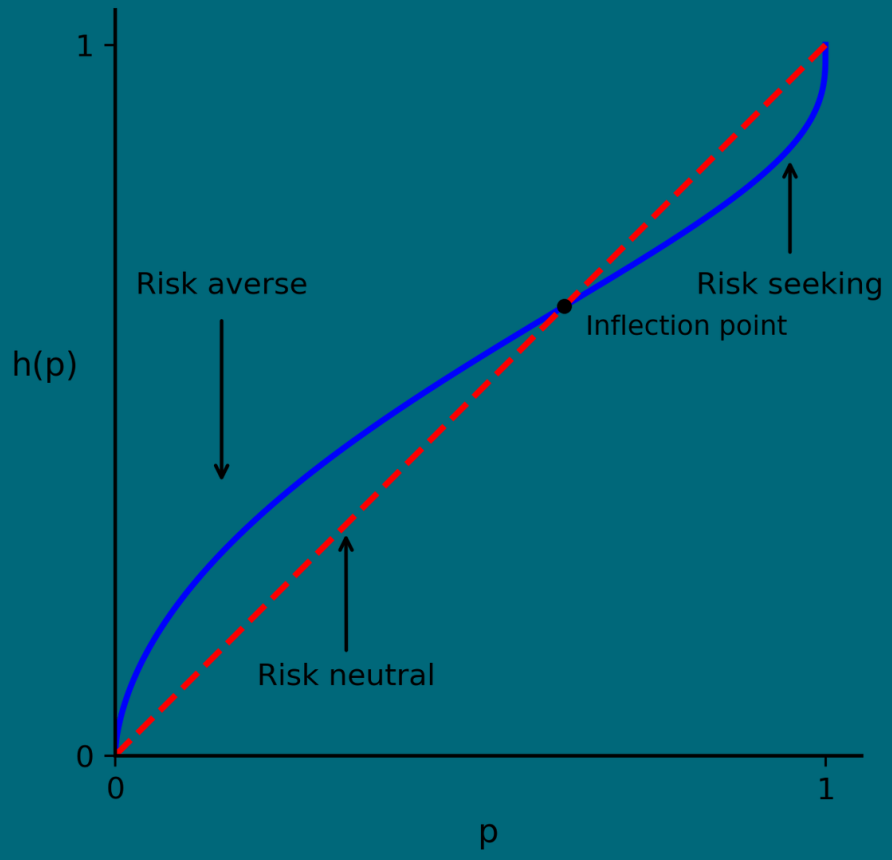
Essays on Distributionally Robust Optimization for Decision-Making under Risk and Ambiguity



$$\int_{-\infty}^{0}\left(h\left(P\left(X>x\right)\right)-1\right)dx+\int_{0}^{\infty}h\left(P\left(X>x\right)\right)dx$$

$$sup_{Q\ll P}E_{Q}\left[X\right]-I_{\phi}\left(Q,P\right)$$

Guanyu Jin

Essays on Distributionally Robust Optimization for Decision-Making under Risk and Ambiguity

Guanyu Jin

Essays on Distributionally Robust Optimization for Decision-Making under Risk and Ambiguity

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op maandag 26 januari 2026, te 13.00 uur

door Guanyu Jin
geboren te ShenZhen

Promotiecommissie

Promotores:	prof. dr. R.J.A. Laeven	Universiteit van Amsterdam
	prof. dr. ir. D. den Hertog	Universiteit van Amsterdam
Overige leden:	prof. dr. S.I. Birbil	Universiteit van Amsterdam
	prof. dr. ir. M.H. Vellekoop	Universiteit van Amsterdam
	dr. S. van Bilsen	Universiteit van Amsterdam
	prof. dr. N.F.F. Schweizer	Tilburg University
	prof. dr. P. Mohajerin Esfahani	TU Delft
	prof. dr. V. Krätschmer	University of Duisburg-Essen

Faculteit Economie en Bedrijfskunde

List of authors

Robust Optimization of Rank-Dependent Models with Uncertain Probabilities

This chapter is co-authored with Roger J. A. Laeven and Dick den Hertog and is based on Jin, Laeven, and den Hertog [2025b].

Constructing Uncertainty Sets for Robust Risk Measures: A Composition of ϕ-Divergences Approach to Combat Tail Uncertainty

This chapter is co-authored with Roger J. A. Laeven, Dick den Hertog and Aharon Ben-Tal and is based on Jin, Laeven, den Hertog, and Ben-Tal [2024].

Sample Average Approximation of Risk Functionals: Non-Asymptotic Error Bounds

This chapter is co-authored with Volker Krätschmer and Roger J. A. Laeven and is based on Jin, Krätschmer, and Laeven [2025a].

A New Distributionally Robust Optimization Model Based on Maxiance Regularization

This chapter is co-authored with Roger J. A. Laeven and Henry Lam and is based on Jin, Laeven, and Lam [2025c].

Preface

Decision-making in life is a perfect example of how optimization problems are sensitive to uncertainty. I still recall the day that I was anxiously waiting for Roger's final decision on my PhD application. Had his email arrived an hour later, I would have taken a completely different career path. It is fortunate that I had the opportunity to pursue a PhD, and therefore I would like to express my appreciation to those who have supported me along the journey.

First and foremost, I would like to express my gratitude to my supervisors, Roger Laeven and Dick den Hertog. Without their guidance, this thesis would have been an impossible task. Roger, thank you for your feedback on my research, academic writing, and career orientation. Dick, thank you for always helping me with difficult optimization problems and for connecting me with other PhD colleagues.

Second, I would like to thank Aharon Ben-Tal, who has made my visit to Technion a pleasant experience. I would also like to thank Henry Lam, who has provided me with valuable insights during my visit to Columbia University.

I wish to express my gratitude to my co-author Volker Krätschmer, with whom I had a very productive collaboration on the third chapter of my PhD thesis. I would also like to thank my co-author and fellow PhD colleague, Justin Starreveld, with whom I have not only collaborated on research, but also enjoyed our visits to the conference EURO 22.

I would like to express my appreciation to all the staff and colleagues at the Amsterdam School of Economics and Business. In particular, I would like to thank my office-mates Evgenii, Justin, Mücahit and Raviar, who made my PhD a joyful journey.

Finally, I would like to express my love and appreciation to my parents, Arno and Huiwen, who have always supported me through all stages of life. I would also like to thank my girlfriend, Kay-leigh Kasan, for her support and encouragement during the difficult parts of my PhD.

Contents

Introduction **9**

1 Robust Optimization of Rank-Dependent Models with Uncertain Probabilities **13**

1.1 Introduction . . . 13
1.1.1 Related Literature . . . 16
1.2 Setup and Notation . . . 17
1.2.1 Rank-Dependent Evaluation . . . 17
1.2.2 ϕ-Divergence Ambiguity Sets . . . 18
1.2.3 Problem Formulations, Terminology and Assumptions . . . 19
1.2.4 Further Notation . . . 21
1.3 Robust Counterpart of Rank-Dependent Models . . . 21
1.3.1 Reformulation into the Robust Counterpart . . . 21
1.3.2 Conic Representability of the Robust Counterpart . . . 23
1.4 Solving the Robust Problem I: A Cutting-Plane Method . . . 26
1.4.1 The Cutting-Plane Algorithm . . . 26
1.4.2 Robust Optimization with a Rank-Dependent Evaluation in the Constraint . . . 28
1.4.3 Choquet Expected Utility . . . 29
1.5 Solving the Robust Problem II: Piecewise-Linear Approximation . . . 30
1.5.1 Piecewise-Linear Distortion Functions . . . 30
1.5.2 Piecewise-Linear Approximation of General Concave Distortion Functions . . . 32
1.5.3 The Piecewise-Linear Approximation Method and its Convergence . . . 32
1.6 Non-Concave Distortion Functions . . . 33
1.6.1 The Nominal Problem . . . 34
1.6.2 Robust Rank-Dependent Models with Inverse S-Shaped Distortion . . . 36
1.7 Numerical Examples . . . 38
1.7.1 Robust Single-Item Newsvendor . . . 39
1.7.2 Robust Multi-Item Newsvendor . . . 41

1.7.3 Robust Portfolio Choice with Concave Distortion Function 43
1.7.4 Robust Portfolio Choice with Inverse S-shaped Distortion Function 44
1.8 Concluding Remarks 47
1.A Tables 1.4 and 1.5 48
1.B Proofs 50
1.C Derivations of the Conjugate Functions and Their Epigraphs 75
1.C.1 Distortion Functions 75
1.C.2 Divergence Functions 80
1.D Robust Rank-Dependent Evaluation Using Penalization 81
1.E A Visualization of the Various Shapes of the Uncertainty Set $\mathcal{U}_{\phi,h}(\mathbf{p})$ . . . 84
1.F The Optimistic Dual Counterpart 84
1.G Optimization of Rank-Dependent Models in the Constraint 87
1.H SOS2-Constraints Formulation 87
1.I The Hit-and-Run Algorithm 88

2 Constructing Uncertainty Sets for Robust Risk Measures: A Composition of ϕ-Divergences Approach to Combat Tail Uncertainty 90
2.1 Introduction 90
2.2 ϕ-Divergences and Risk Measures 93
2.3 Divergence Choices for Robust OCE Risk Measures 95
2.3.1 Necessary and Sufficient Conditions for Finite Robust OCEs 95
2.3.2 Controlling the Tail Properties of Distributions in a Divergence Ball 99
2.3.3 Examples of Tailored Divergences 102
2.4 Other Types of Robust Risk Measures 104
2.5 Elicitation of Divergences from Robust Risk Measures 107
2.6 Computing the Robust OCE Risk Measures 108
2.6.1 Addressing Sampling Errors and Model Misspecification Using Globalized Robust Optimization 110
2.7 Numerical Illustrations 111
2.7.1 Toy Examples 111
2.7.2 Discrete Delta Hedging in the Black-Scholes Model 113
2.7.3 Risk-Averse Newsvendor Optimization 116
2.8 Concluding Remarks 119
2.A Proofs 119
2.B Minimizing Convex Integral 139
2.C Details for Remark 2.3 142
2.D Characterization of Risk Aversion of Robust OCE and its Preservation of Convex Order 143
2.E Additional Details on Tail Analysis 145

2.E.1 Details for Table 2.1 . . . 146
2.E.2 Details for Table 2.2 . . . 149
2.E.3 Details for Table 2.3 . . . 151

3 Sample Average Approximation of Risk Functionals: Non-Asymptotic Error Bounds 154
3.1 Introduction . . . 154
3.1.1 Related literature . . . 157
3.1.2 Outline . . . 159
3.2 Set up . . . 159
3.3 Deviation probabilities . . . 163
3.3.1 The Coherent Case . . . 164
3.3.2 Alternative Bound for Coherent Risk Measures . . . 166
3.3.3 Extension to General Convex Risk Measures . . . 167
3.3.4 Sharpness of Convergence Rates . . . 170
3.4 Distortion Risk Measures with Inverse S-Shaped Distortion Function . . . 171
3.5 Extension to Goal Functions with Unbounded Support . . . 173
3.6 Conclusion and Discussion . . . 176
3.7 Proofs . . . 176
3.7.1 Proof of Theorem 3.10 . . . 179
3.7.2 Proof of Theorem 3.15 . . . 182
3.7.3 Proof of Theorem 3.9 . . . 184
3.7.4 Auxiliary Lemma for Proofs of Theorems 3.10, 3.9 and 3.22 . . . 187
3.7.5 Proofs of Results from Section 3.4 . . . 189
3.7.6 Proof of Theorem 3.20 and Theorem 3.22 . . . 192
3.7.7 Proof of Proposition 3.25 . . . 197
3.8 Additional Details for Distortion Functions . . . 199
3.8.1 Some Details on Table 3.1 . . . 199
3.8.2 Extra Details on Kahneman-Tversky's Distortion Function . . . 200
3.A Kusuoka representation . . . 201
3.B A concentration inequality . . . 205

4 A New Distributionally Robust Optimization Model Based on Maxiance Regularization 206
4.1 Introduction . . . 206
4.2 Preliminaries . . . 211
4.2.1 ϕ-Divergences . . . 211
4.2.2 Maxiance and Distortion Risk Measures . . . 212
4.3 Asymptotic Equivalence of Dual DRO and Mean-Maxiance Optimization . 213
4.4 Optimization of Dual DRO Models . . . 215

4.4.1 Solving Dual DRO 215
4.4.2 A Numerical Investigation of the Advantages of Dual DRO 218
4.5 First-Order Asymptotics of Dual DRO Estimators 222
4.6 Conclusion 226
4.A Proofs 226
4.B Details on Dual DRO Confidence Bounds 249

Bibliography **252**

Summary **263**

Nederlandse Samenvatting (Summary in Dutch) **265**

Introduction

This dissertation is the result of four years of research aimed at developing the necessary mathematical tools to facilitate decision-making under uncertainty. Nearly all decision-making processes are subject to uncertainty. For example, retail business owners must decide about the number of products to be ordered before the realization of actual demand. Financial risk managers are constantly rebalancing their investment portfolios based on estimations of future returns. To facilitate decision-making under uncertainty, we need algorithms that can efficiently extract information from data, make proper risk assessments, and find optimal decisions within limited time frames.

A good algorithm for decision-making in an uncertain environment must be based on a proper theory of risk measurement. Among the various risk theories proposed in the literature, the *expected utility* theory, introduced by Daniel Bernoulli in his work *Exposition of a New Theory on the Measurement of Risk* in 1738, and later axiomatized by von Neumann and Morgenstern [1944], is widely recognized as the benchmark model for rational decision-making. This theory postulates that the risk of a prospect[1] should be measured by its average utility value, where utility may be interpreted as the amount of "satisfaction" that is received from a payoff. Furthermore, utility may be different for each individual, e.g., a gain of 100 euros can be more meaningful to the poor than to the rich.

Despite the popularity of expected utility models in economic theory, the model itself is known to be violated by many empirical observations, such as the famous Allais paradox (see Allais 1953). In particular, expected utility is much criticized among psychologists, who argue that attitude toward risk should not be reflected through the utility values derived from wealth, but rather through the perceptions of probabilities. Namely, risk aversion is a result of overweighting the probability of a bad event, and the converse applies to risk-seeking behavior, where the probability of a good event is overweighted. Therefore, certain non-linearity in probability weighting must be incorporated in a theoretically sound way to produce a new model for decision under risk. This insight was eventually formalized by Quiggin [1982] and Schmeidler [1986, 1989], who introduced the notion of *rank-dependent probabilities* that led to the birth of *rank-dependent utility* models. This

[1] A prospect may refer to a lottery, an investment plan, or anything that involves a future payoff subject to uncertainty.

advancement has also facilitated the further development of *prospect theory* by Kahneman and Tversky [1979], which is one of the most promising state-of-the-art theories on risk measurement.

In all previously mentioned theories, there is the implicit assumption that the probability of each payoff value is known to the decision-maker. This assumption is very idealistic, since in practice the distribution of a future payoff is never fully specified. The lack of prior knowledge on probabilities has a significant effect on decision-making under uncertainty, as demonstrated by Ellsberg 1961 who showed that decision-makers tend to prefer prospects with assessable risks over those that are not. Therefore, besides uncertainty in payoffs, which is commonly referred to as risk, there is also uncertainty in model specification, which in decision theory is known as *ambiguity*. To address both risk and ambiguity, robust decision models have been proposed where risk is measured with respect to a set of distributions. Examples of such robust models include the *maxmin expected utility* by Gilboa and Schmeidler [1989], and the *variational preferences* by Maccheroni et al. [2006]; see also Laeven and Stadje [2024].

Decision-making under uncertainty has also been a central topic in the optimization community. This is motivated by the observation that solutions of optimization problems can be very sensitive to small changes in parameter values that belong to the constraint or objective functions. Notable work that has been done in this direction is the development of *Robust Optimization* by Ben-Tal, Ghaoui, and Nemirovski [2009], who studied optimization problems with worst-case objectives and constraints. Advancements in robust optimization have also attracted researchers from other fields such as statistics and machine learning, who have further developed *Distributionally Robust Optimization* (DRO) techniques with a particular interest to address distributional uncertainty (see e.g., Duchi and Namkoong [2019], Kuhn et al. [2019], Esfahani and Kuhn [2018], van Parys et al. [2021], Lam [2019], Lam and Mottet [2017]).

Ultimately, to provide a methodology for decision under uncertainty that is applicable to real-life problems, one must take into account two major components: a proper theory of risk measurement that is able to capture a decision-maker's attitude toward risk and ambiguity, and efficient optimization techniques that allow us to determine the best solution under the chosen risk model. However, in contrast to expected utility, optimization of more general decision models, such as rank-dependent utility, is still relatively underdeveloped. The same holds for robust decision models, which are inherently more difficult to optimize. Therefore, our first contribution in Chapter 1 is to provide a tractable methodology to solve both distributionally robust and non-robust optimization problems where the risk measures are defined through rank-dependent utility models. The main challenges associated with these types of optimization problems stem from their rank-dependent nature. We show that for many probability weighting functions including concave, convex and inverse S-shaped, both the robust and non-robust optimization problem can be re-

formulated into a rank-independent problem. When the probability weighting function is concave, the problem can be further reformulated into a convex optimization problem with finitely many constraints that admits explicit conic representability for a collection of canonical examples. Although the number of constraints in the reformulated problem scales exponentially with the dimension of the state space, we circumvent this dimensionality curse and provide two types of upper and lower bounds algorithms. They yield tight upper and lower bounds on the exact optimal value and are formally shown to converge asymptotically.

While Chapter 1 provides the necessary computational tools to solve a *given* distributionally robust risk measure optimization problem, it does not answer the key question asked among all practitioners, namely, which robust problem should one solve? Indeed, the specification of a robust optimization problem, which comes down to the construction of a proper uncertainty set, is a different problem of its own. Our second contribution in Chapter 2 aims to answer this question in the context of using robust optimization to address tail uncertainty when measuring risk. We provide a unifying framework to construct uncertainty sets for a broad class of risk measures, given a specified nominal model. The framework is based on a parametrization of robust risk measures using two (or multiple) ϕ-divergence functions (see Subsection 1.2.2), which enables us to provide uncertainty sets that are tailored to both the sensitivity of each risk measure to tail losses and the tail behavior of the nominal distribution. In addition, our formulation allows for a tractable computation of robust risk measures and the elicitation of ϕ-divergences that describe a decision-maker's risk and ambiguity preferences.

As the third contribution of this thesis, presented in Chapter 3, we focus on the statistical properties of Sample Average Approximations (SAA) for risk-averse stochastic optimization problems. SAA is widely used as an estimation method for the optimal value of a stochastic optimization problem by replacing the unknown distribution with its empirical counterpart constructed from data. Since SAA is a statistical estimator, it is essential to examine its convergence to the true optimal value. We study SAA for risk-averse stochastic optimization where the risk of an uncertain goal function is measured by a law-invariant convex risk measure. Non-asymptotic upper estimation of the probability of the absolute deviation error between SAA and the true optimal value is studied and developed under general assumptions on the goal function and risk measure. In particular, the goal function is allowed to be non-linear, unbounded, and does not require any pathwise analytical properties such as continuity or convexity. Conditions on the risk measures are significantly simplified in the coherent case, with prominent examples including distortion risk measures and expectiles. Furthermore, our non-asymptotic bounds can be used to construct confidence regions for the true optimal value, including for stochastic optimization problems with inverse S-shaped distortion risk measures.

Finally, in Chapter 4, we examine decision under uncertainty in machine learning

problems, where the goal is often to find the best learning model that trades off bias and variability. In statistics, this is also known as the bias-variance tradeoff, where variance serves as the standard measure of variability. However, variance does not always adequately represent variability within a population, since it measures deviation only relative to a single reference point (namely, the expected value). This makes variance sensitive to outliers, which can deteriorate the performance of ϕ-divergence DRO models, as these types of DRO models are known for their variance regularization effect. Therefore, we propose a new DRO approach, which we call dual DRO, that is based on maxiance regularization. Maxiance regularization ensures that the dual DRO approach is more robust against outliers than ϕ-divergence DRO, while also providing the bias and variability tradeoff. Moreover, optimization of dual DRO models is just as tractable as ϕ-divergence DRO, and can be used as a tractable lower bound on the true optimal expected value problem with the same $O(1/\sqrt{n})$ asymptotic convergence rate as ϕ-divergence DRO.

Chapter 1

Robust Optimization of Rank-Dependent Models with Uncertain Probabilities

1.1 Introduction

Many stochastic optimization problems in management science, operations research, engineering, economics, and finance arise from decisions involving risk (probabilities given) and ambiguity (probabilities unknown).

A variety of models for decision under risk have been proposed. Among the most popular and empirically viable models is the rank-dependent utility (RDU) model of Quiggin [1982]. In RDU, the utility loss associated to a random variable X under the probabilistic model $\mathbb{P}$ is measured by a rank-dependent evaluation ρ with respect to a non-additive, distorted measure $h \circ \mathbb{P}$:[1]

$$\rho_{u,h,\mathbb{P}}(X) \triangleq \int -u(X)\mathrm{d}(h \circ \mathbb{P}),$$

where $u : \mathbb{R} \to \mathbb{R}$ is a utility function, assumed to be non-decreasing, and $h : [0,1] \to [0,1]$, with $h(0) = 1 - h(1) = 0$ and non-decreasing, is a distortion or probability weighting function. RDU serves as a pivotal building block of prospect theory [Kahneman and Tversky, 1979] and encompasses expected utility [von Neumann and Morgenstern, 1944] when h is linear and the dual theory of Yaari [1987] when u is affine. It accommodates Allais [1953] type phenomena that are incompatible with expected utility. The utility function captures attitude toward wealth and the shape of the distortion function, e.g., concave, convex or (inverse) S-shaped, dictates attitude toward risk. Importantly, under RDU probabilities of outcomes are weighted according to $h(F_X(x)) \triangleq h(\mathbb{P}(X \leq x))$, leading to an evaluation that is non-linear in probabilities, which depend on the ranking of outcomes. These non-linear and rank-dependent features bring major computational

[1] See also Schmeidler [1986, 1989] for related pioneering developments.

challenges.

In the aforementioned theories for decision under risk, $\mathbb{P}$ is assumed to be given. When ambiguity is present, $\mathbb{P}$ is unknown. Ambiguity is often treated via a worst-case approach that is robust against malevolent nature. For example, Gilboa and Schmeidler [1989] introduce maxmin expected utility, a.k.a. multiple priors, under which risks are evaluated according to their worst-case expected utility taken over a set of probabilistic models. Hansen and Sargent [2001, 2008] propose multiplier preferences under which probabilistic models far away from a reference model are penalized according to the Kullback-Leibler divergence (a.k.a. relative entropy). The variational preferences of Maccheroni et al. [2006] admit a general penalty function, thus significantly generalizing the maxmin and multiplier preferences models. Laeven and Stadje [2024] develop a rank-dependent theory for decision under risk and ambiguity that encompasses both the dual and rank-dependent counterparts of Gilboa and Schmeidler [1989] and Maccheroni et al. [2006].

Parallel to these developments, the field of (distributionally) robust optimization has been studying risk and ambiguity from a computational perspective. In robust optimization, ambiguity sets are often constructed exogenously from data, as a confidence region of the true underlying distribution, rather than endogenously based on preferences. A widely used family of statistical estimators for distributional uncertainty is given by ϕ-divergences [Csiszár, 1975, Ben-Tal and Teboulle, 1986, 1987, Pardo, 2006]. Many distributionally robust optimization problems with these types of ambiguity sets can be reformulated into tractable robust counterparts, which can then be solved efficiently using standard optimization algorithms (see e.g., Ben-Tal et al., 2013, Wiesemann et al., 2014, and Esfahani and Kuhn, 2018).

Although a variety of distributionally robust optimization techniques have been developed for optimization under uncertainty, most of the literature is concerned with stochastic models in which the probabilities appear linearly in the optimization problems, such as (possibly generalized) variants of expected value or expected utility maximization. Despite the growing interest in and applications of rank-dependent models across a wide variety of fields (see e.g., Denneberg, 1994, Wakker, 2010, Föllmer and Schied, 2016, and the references therein), optimization of these models—whether ambiguity is present or not—is still relatively underdeveloped. The main difficulties lie in both the non-linearity in probabilities and the rank-dependence: for each value of a decision vector, the rank-dependent evaluation of an uncertain objective or constraint can be permuted, since the ranking of the outcomes depends upon the decision vector.

In this paper, we develop an efficient approach for optimizing rank-dependent models, with and without uncertain probabilities. More precisely, we study the following nominal

and robust minimization problems, in a discrete probability space setting:

$$\inf_{\mathbf{a}\in\mathcal{A}} \rho_{u,h,\mathbf{p}}(f(\mathbf{a},\mathbf{X})), \tag{1.1}$$

$$\inf_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})), \tag{1.2}$$

where $\mathbf{a} \in \mathbb{R}^{n_a}$ is the decision vector, $\mathcal{A}$ is a set of constraints, $\mathbf{X} \in \mathbb{R}^l$ is a random vector, $f : \mathbb{R}^{n_a} \times \mathbb{R}^l \to \mathbb{R}$ is a deterministic function, $\mathcal{D}_\phi(\mathbf{p}, r)$ is a ϕ-divergence ambiguity set (formally defined in (1.8)), and $\rho_{u,h,\mathbf{q}}(.)$ is a rank-dependent evaluation (formally defined in (1.5)), for some probability vectors $\mathbf{p}, \mathbf{q} \in \mathbb{R}^m$.

Our main contributions can be summarized as follows:

- We show that we can reformulate problems (1.1) and (1.2), as well as their epigraph formulations, into rank-independent, tractable (robust) counterparts, for the various empirically relevant shapes of the probability weighting function h, including concave, convex, and inverse S-shaped functions.

- For concave probability weighting functions, we show that the reformulated robust counterpart admits a conic representation, provided h and ϕ are conic representable functions. For a list of canonical examples of h and ϕ, we provide explicit epigraph representations of the reformulated robust counterpart that can directly be implemented into standard conic optimization programs such as CVXPY.

- While the reformulated robust counterparts are more tractable than (1.1) and (1.2), their multiplicity of constraints increases exponentially in the dimension of the underlying discrete probability space. We provide two types of algorithms to circumvent this curse of dimensionality. Each of them, or a combination thereof in the case of convex and inverse S-shaped probability weighting functions, yields tight upper and lower bounds that, as we formally establish, converge to the optimal objective value of the exact problem. Moreover, we show that one of our algorithms can also be applied to a more general type of rank-dependent model, namely Choquet expected utility (CEU, Schmeidler, 1989).[2]

- We provide numerical examples of our approach in applications of robust optimization with rank-dependent models and uncertain probabilities, in the context of inventory planning and portfolio optimization. We efficiently obtain optimal decisions that are robust against uncertainty, and yield accurate optimal objective values. Concrete examples with codes are provided on `https://github.com/GuanJinNL/ROptRDU.github.io.git`.

[2]Under CEU, the non-additive measure need not be obtained by distorting an additive probability measure.

To our best knowledge, our approach is the first that can handle nominal and distributionally robust, generic optimization problems involving rank-dependent evaluations with possibly inverse S-shaped probability weighting functions.

1.1.1 Related Literature

Our work builds on the decision-theoretic literature on evaluating risk and ambiguity. Specifically, we consider in (1.1)–(1.2) the rank-dependent evaluations of Quiggin [1982] and adopt a robust approach as in Gilboa and Schmeidler [1989], Maccheroni et al. [2006], Hansen and Sargent [2001, 2008] and Laeven and Stadje [2024], which can be viewed as (generalized) decision-theoretic foundations of the classical decision rule of Wald [1950]; see also Huber [1981]. We contribute to this literature by developing corresponding optimization techniques.

An initial connection between robust risk measures and robust optimization has been studied in an early paper by Bertsimas and Brown [2009]. They show that a decision-maker's preference, as represented by a coherent risk measure, can provide a device for constructing an uncertainty set for robust optimization purposes. The authors were able to obtain tractability results for distortion risk measures under a specific parameterization and the (strict) assumption that the underlying discrete probability space is uniform. Our paper relaxes these assumptions, while also providing a blueprint for constructing uncertainty sets tailored to general risk and ambiguity preferences. Postek et al. [2016] studied distributionally robust optimization (DRO) problems for a broad collection of uncertainty sets and risk measures. However, for classes of risk measures that are non-linear in probabilities, such as distortion risk measures, the tractability remained unknown (see their Table I); they are covered as special cases in this paper. Although different choices of uncertainty sets are possible, our paper focuses on uncertainty sets defined by ϕ-divergences, which constitute a family of divergences including the Kullback-Leibler divergence, Burg entropy and Hellinger distance. The study of ϕ-divergences in DRO is motivated by several earlier studies; see e.g., Ben-Tal and Teboulle [1986, 1987] and Ben-Tal et al. [2013].

Recently, Cai et al. [2023] and Pesenti et al. [2020] studied DRO problems involving distortion risk measures with general non-concave distortion functions. Their interesting results show that the DRO problem associated with a non-concave distortion function is equivalent to that with its concave envelope approximation provided that the underlying uncertainty set obeys certain moment conditions. In this paper, we show that this equivalence is typically not satisfied in our general setting, thus requiring novel techniques. Optimization of non-expected utility models with possibly non-concave distortion functions has also been studied in innovative work of He and Zhou [2011], in particular in the context of portfolio choice, introducing the so-called quantile method; see also Carlier

and Dana [2006]. The effectiveness of this method relies on the ability to parametrize the distributions of all random objective functions $\{f(\mathbf{a}, \mathbf{X}), \mathbf{a} \in \mathcal{A}\}$ by a set of quantile functions that satisfy finitely many constraints. For a general convex function f and feasibility set $\mathcal{A}$ that are considered in (1.1), there is, however, no systematic approach to obtain such a quantile formulation. Our work provides a method to optimize (both robust and nominal) rank-dependent models for a broad class of decision problems and can directly be implemented in standard optimization software. Finally, Delage et al. [2022] and Wang and Xu [2023] analyze 'preference robust' optimization, which considers uncertainty in the distortion function rather than in the underlying probabilistic model.

The remainder of the paper is organized as follows. Section 1.2 introduces the setting and notation. Sections 1.3 to 1.5 study the optimization problems (1.1) and (1.2) for concave distortion functions. The techniques we develop are extended in Section 1.6 to non-concave, inverse S-shaped distortion functions. Section 1.7 presents our numerical experiments. Conclusions are in Section 1.8. In Appendix 1.A–1.I, we provide all the proofs, additional examples, and technical details.

1.2 Setup and Notation

1.2.1 Rank-Dependent Evaluation

Let $(\Omega, \mathcal{F})$ be a measurable space. We define the rank-dependent evaluation $\rho_{u,h,\mathbb{Q}}$ of the utility loss associated with a random variable $X : \Omega \to \mathbb{R}$ under a given probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ as the integral with respect to the non-additive measure $h \circ \mathbb{Q}$ [Denneberg, 1994], or equivalently, as a Choquet integral [Choquet, 1954]:

$$\rho_{u,h,\mathbb{Q}}(X) \triangleq \int -u(X)\,\mathrm{d}(h \circ \mathbb{Q}) \tag{1.3}$$

$$= \int_0^\infty h\left(\mathbb{Q}[-u(X) > t]\right)\mathrm{d}t + \int_{-\infty}^0 \left(h\left(\mathbb{Q}[-u(X) > t]\right) - 1\right)\mathrm{d}t. \tag{1.4}$$

Here, u is a non-decreasing utility function defined on a suitable domain containing the support of X. Furthermore, the function $h : [0,1] \to [0,1]$ is a distortion, or probability weighting, function that is non-decreasing and satisfies $h(0) = 0$ and $h(1) = 1$. We note that $\rho_{u,h,\mathbb{Q}}$ is also known as a *distortion risk measure* when u is the identity function. In this paper, we consider distortion functions that may be concave, convex or inverse S-shaped.

Definition 1.1. We say that h is inverse S-shaped if, for some $p^0 \in (0,1)$, we have that h is concave for $p \leq p^0$ and h is convex for $p \geq p^0$.

If Ω is discrete with $|\Omega| = m$, then the integral (1.4) reduces to a rank-dependent sum. Let $x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(m)}$ denote the ranked realizations of X, with $\{(i)\}_{i=1}^m$ denoting

the indices of the ranked realizations. The monotonicity of u preserves the ranking of $(x_i)_{i=1}^m$. Therefore, we have

$$\rho_{u,h,\mathbf{q}}(X) = \sum_{i=1}^{m} - \left(h\left(\sum_{j=i}^{m} q_{(j)} \right) - h\left(\sum_{j=i+1}^{m} q_{(j)} \right) \right) u\left(x_{(i)} \right), \tag{1.5}$$

with $\mathbf{q} \in [0,1]^m$ the probability vector associated to Ω and $\sum_{j=m+1}^{m} q_{(j)} \triangleq 0$, by convention.

A well-known example of a rank-dependent evaluation is the Conditional-Value-at-Risk (a.k.a. Expected Shortfall; Föllmer and Schied, 2016, Section 4.6), $\mathrm{CVaR}_{1-\alpha}(X)$, which is defined as

$$\mathrm{CVaR}_{1-\alpha}(X) \triangleq \frac{1}{1-\alpha} \int_0^{1-\alpha} \mathrm{VaR}_\gamma(X)\, \mathrm{d}\gamma, \quad 0 \leq \alpha < 1, \tag{1.6}$$

where $\mathrm{VaR}_\gamma(X) \triangleq \inf\{x : \mathbb{Q}(-X \leq x) \geq 1-\gamma\}$, $0 < \gamma < 1$, is the Value-at-Risk. For a certain level $\alpha \in [0,1)$, $\mathrm{CVaR}_{1-\alpha}(X)$ can be interpreted as the (sign-changed) average of the left $(1-\alpha)\cdot 100\%$ tail of the risk X. It is easily verified that $\mathrm{CVaR}_{1-\alpha}$ is a rank-dependent evaluation with linear utility function and distortion function $h(p) = \min\left\{\frac{p}{1-\alpha}, 1\right\}$. Other canonical examples of distortion functions are given in Table 1.4 of Appendix 1.A. The distortion function captures attitude toward risk whereas the utility function describes attitude toward wealth (see e.g., Quiggin, 1982, Yaari, 1987, Chew et al., 1987, and Eeckhoudt and Laeven, 2021).

1.2.2 ϕ-Divergence Ambiguity Sets

We construct ambiguity (or uncertainty) sets using ϕ-divergences. For discrete outcome spaces Ω with $|\Omega| = m$, the ϕ-divergence $I_\phi(\mathbf{q}, \mathbf{p})$ between two probability vectors $\mathbf{q}, \mathbf{p} \in [0,1]^m$ is defined as

$$I_\phi(\mathbf{q}, \mathbf{p}) \triangleq \sum_{i=1}^{m} p_i \phi\left(\frac{q_i}{p_i} \right). \tag{1.7}$$

Here, $\phi : [0,\infty) \to \mathbb{R}$ is a convex function that satisfies the following conventions: $\phi(1) = 0$, $0\phi(0/0) \triangleq 0$ and $0\phi(x/0) \triangleq x \lim_{t\to\infty} \phi(t)/t$ (see Pardo, 2006, Definition 1.1). For a nominal probability vector $\mathbf{p}$ and $r > 0$, the ϕ-divergence ambiguity set is defined as

$$\mathcal{D}_\phi(\mathbf{p}, r) \triangleq \left\{ \mathbf{q} \in \mathbb{R}^m \;\middle|\; \mathbf{q} \geq \mathbf{0},\ \mathbf{q}^T\mathbf{1} = 1, I_\phi(\mathbf{q}, \mathbf{p}) \leq r \right\}. \tag{1.8}$$

Here, $\mathbf{0}, \mathbf{1} \in \mathbb{R}^m$ are the vectors with all entries equal to 0 and 1, respectively. As outlined in Ben-Tal et al. [2013] (Section 3.2), one may construct $\mathcal{D}_\phi(\mathbf{p}, r)$ as a statistical confidence set by replacing $\mathbf{p}$ with an empirical estimator $\hat{\mathbf{p}}_n$. Indeed, as shown in Pardo [2006] (Corollary 3.1), under the null-hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0$, and provided ϕ is twice

continuously differentiable in a neighborhood of 1 with $\phi''(1) > 1$, the following object converges to a chi-squared distribution:[3]

$$\frac{2n}{\phi''(1)} I_\phi(\mathbf{p}_0, \hat{\mathbf{p}}_n) \rightsquigarrow \chi^2_{m-1,1-\alpha}. \tag{1.9}$$

Here, n is the sample size used for constructing the empirical estimator, and $\chi^2_{m-1,1-\alpha}$ is the $(1-\alpha)$-quantile of the chi-square distribution with $m-1$ degrees of freedom.[4] Using (1.9), one can construct an asymptotic confidence set $\mathcal{D}_\phi(\hat{\mathbf{p}}_n, r)$ by choosing

$$r = \frac{\phi''(1)}{2n} \chi^2_{m-1,1-\alpha}. \tag{1.10}$$

1.2.3 Problem Formulations, Terminology and Assumptions

We study the following nominal and robust minimization problems:

$$\inf_{\mathbf{a}\in\mathcal{A}} \rho_{u,h,\mathbf{p}}(f(\mathbf{a}, \mathbf{X})), \tag{P-Nom}$$

$$\inf_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}, \mathbf{X})), \tag{P}$$

where $\mathbf{a} \in \mathbb{R}^{n_a}$ is the decision vector contained in a compact set of constraints $\mathcal{A}$ consisting of convex inequalities, $\mathbf{X} \in \mathbb{R}^l$ is a random vector, $f : \mathbb{R}^{n_a} \times \mathbb{R}^l \to \mathbb{R}$ is a jointly convex function in $(\mathbf{a}, \mathbf{x})$, $\mathcal{D}_\phi(\mathbf{p}, r)$ is a ϕ-divergence ambiguity set defined in (1.8) with respect to a nominal probability vector $\mathbf{p}$, and $\rho_{u,h,\mathbf{q}}(.)$ is the rank-dependent evaluation defined in (1.5), with respect to some probability vector $\mathbf{q} \in \mathbb{R}^m$.

Henceforth, we often consider the nominal and robust rank-dependent evaluations in a constraint form induced by the following epigraph formulation:

$$\rho_{u,h,\mathbf{p}}(f(\mathbf{a}, \mathbf{X})) \leq c, \tag{P-Nom-EG}$$

$$\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}, \mathbf{X})) \leq c. \tag{P-EG}$$

That is, we emphasize that we are able to deal with a robust rank-dependent evaluation both in the objective and in the constraint, where in the latter case we consider the

[3] From a statistical perspective, it may be more natural to consider $I_\phi(\hat{\mathbf{p}}_n, \mathbf{p}_0)$, where the true model $\mathbf{p}_0$ appears as the reference model. However, in the robust optimization literature, $I_\phi(\mathbf{p}_0, \hat{\mathbf{p}}_n)$ appears more commonly. According to Theorem 3.1 and Corollary 3.1 of Pardo [2006], both objects have the same limiting distribution under the null.

[4] That is, $\mathbb{P}(Z \leq \chi^2_{m-1,1-\alpha}) = 1 - \alpha$, for $Z \sim \chi^2_{m-1}$.

following type of problem:

$$\begin{aligned} &\min_{\mathbf{a}\in\mathcal{A}} \; g(\mathbf{a}) \\ &\text{s.t.} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \le c, \end{aligned} \tag{P-constraint}$$

where g is convex in $\mathbf{a}$, and $c \in \mathbb{R}$ is a fixed parameter (in contrast to being an epigraph variable). The nominal version of (P-constraint) is defined analogously, with the only difference that the robust constraint is replaced by the nominal constraint (P-Nom-EG).

Additionally, we would like to mention the alternative, but highly related, approach of defining a robust rank-dependent evaluation, which appears in the decision theory literature (see e.g., Laeven and Stadje, 2024):

$$\tilde{\rho}_{\mathrm{rob}}(X) \triangleq \sup_{\mathbf{q}\ge\mathbf{0},\mathbf{q}^T\mathbf{1}=1} \rho_{u,h,\mathbf{q}}(X) - \theta I_\phi(\mathbf{q},\mathbf{p}), \quad \theta > 0. \tag{1.11}$$

In Appendix 1.D, we provide additional details on how (1.11) can be reformulated similar to (P-EG), and that the minimization problem (P) is equivalent to minimizing (1.11) for a specific θ.

Let $[m] \triangleq \{1,\dots,m\}$ for an integer m. The following assumptions are made throughout this paper:

Assumption 1.2. *The optimization problem* (P) *is finite:* $-\infty <$(P)$ < \infty$.

Assumption 1.3. *The nominal probability vector* $\mathbf{p}$ *satisfies* $p_i > 0$ *for all* $i \in [m]$.

Assumption 1.4. *The functions* $\phi(t), -h(p), -u(f(\mathbf{a},\mathbf{x}))$ *are lower-semicontinuous on their respective domains.*

Assumption 1.5. (P-constraint) *is finite and contains a Slater point, i.e., there exists* $\mathbf{a}_0 \in \mathrm{int}(\mathcal{A})$ *such that* $\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}_0,\mathbf{X})) < c$. *Furthermore, we assume* (P-constraint) $> \inf_{\mathbf{a}\in\mathcal{A}} g(\mathbf{a})$.

Assumption 1.2 can be satisfied if e.g., the image set $\{u(f(\mathbf{a},\mathbf{x})) \mid \mathbf{a}\in\mathcal{A}, \mathbf{x}\in\mathrm{supp}(\mathbf{X})\}$ is contained in a bounded interval. Assumption 1.3 constitutes a weak redundancy condition. Assumption 1.4 is to ensure that the optimization problem (P) has an optimal solution in $\mathcal{A}$, and Assumption 1.5 is to ensure that (P-constraint) satisfies the strong duality theorem.

In this paper, the term *robust solution* refers to an optimal solution of (P) or (P-constraint). Similarly, a *nominal solution* refers to an optimal solution of (P-Nom) or the nominal version of (P-constraint).

1.2.4 Further Notation

For a convex function $g : \mathbb{R}^d \to \mathbb{R}$, we denote by g^* its convex *conjugate* $g^*(\mathbf{y}) \triangleq \sup_{\mathbf{z}\in\mathrm{dom}(g)}\{\mathbf{y}^T\mathbf{z} - g(\mathbf{z})\}$, where $\mathrm{dom}(g) \triangleq \{\mathbf{z} \mid g(\mathbf{z}) < +\infty\}$ is the effective domain of g. Note that g^* is always convex, since it is the pointwise supremum of a linear function in $\mathbf{y}$. Furthermore, g^* is non-decreasing if $\mathrm{dom}(g) \subset [0,\infty)$. For $\lambda > 0$, the *perspective* of g is the function $\tilde{g} : (\mathbf{z},\lambda) \mapsto \lambda g\left(\frac{\mathbf{z}}{\lambda}\right)$. By convention, $\tilde{g}(\mathbf{0},0) = 0$ and $\tilde{g}(\mathbf{z},0) = \infty$ for $x \neq 0$. Note that $\tilde{g}$ is convex if g is convex. The *epigraph* of a function g is the set $\mathrm{Epi}(g) \triangleq \{(\mathbf{z},t) : g(\mathbf{z}) \leq t\}$. An epigraph may have a conic representation expressed in a conic inequality $\succeq_{\mathbf{K}}$, where $\mathbf{K}$ is a proper cone (i.e., pointed, closed, convex and with a non-empty interior) and $\mathbf{K}^*$ denotes its dual cone, defined as

$$\mathbf{K}^* \triangleq \{\boldsymbol{\lambda} \in \mathbb{R}^d : \boldsymbol{\lambda}^T\mathbf{z} \geq 0, \forall \mathbf{z} \in \mathbf{K}\};$$

see Chares [2009] and Ben-Tal and Nemirovski [2019] for further details.

1.3 Robust Counterpart of Rank-Dependent Models

In this section, we show how the optimization problems (P-Nom)–(P) can be reformulated into rank-independent problems with finitely many constraints. Our idea hinges on the insight that a rank-dependent evaluation with a concave distortion function admits a dual representation that is itself a robust optimization problem with linear probabilities and convex uncertainty set. Although the dual representation holds only for concave distortion functions, we show in Section 1.6 how the same idea can be extended to encompass convex and inverse S-shaped distortion functions. This makes it possible to conduct robust optimization for a broad class of rank-dependent models.

1.3.1 Reformulation into the Robust Counterpart

We start by reformulating the constraints (P-EG)–(P-Nom-EG) with concave distortion functions, since they form the basis of the more complicated convex and inverse S-shaped cases described in Section 1.6. In Sections 1.3–1.5, without further mentioning, we assume the utility function to be concave; Section 1.6 allows u to be non-concave. The reformulation relies on utilizing the following dual representation, which involves a *composite uncertainty set*.

Theorem 1.6. *Let $h : [0,1] \to [0,1]$ be a concave distortion function. Then, for all $(\mathbf{a},c) \in \mathbb{R}^{n_a+1}$, we have that* (P-EG) *is satisfied if and only if*

$$\sup_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_{\phi,h}(\mathbf{p})} \sum_{i=1}^{m} -\bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)) \leq c, \tag{1.12}$$

where $\mathcal{U}_{\phi,h}(\mathbf{p})$ *is a convex composite uncertainty set given by*

$$\mathcal{U}_{\phi,h}(\mathbf{p}) \triangleq \left\{ (\mathbf{q}, \bar{\mathbf{q}}) \in \mathbb{R}^{2m} \;\middle|\; \begin{array}{l} \sum_{i=1}^m q_i = \sum_{i=1}^m \bar{q}_i = 1 \\ \sum_{i=1}^m p_i \phi\left(\frac{q_i}{p_i}\right) \leq r \\ \sum_{i\in J} \bar{q}_i \leq h\left(\sum_{i\in J} q_i\right), \ \forall J \subset [m] \\ q_i, \bar{q}_i \geq 0, \ \forall i \in [m] \end{array} \right\}. \tag{1.13}$$

As a special case of interest, (P-Nom-EG) *is satisfied if and only if*

$$\sup_{\bar{\mathbf{q}} \in M_h(\mathbf{p})} \sum_{i=1}^m -\bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)) \leq c, \tag{1.14}$$

where $M_h(\mathbf{p})$ *is the set induced by* h*:*

$$M_h(\mathbf{p}) \triangleq \left\{ \bar{\mathbf{q}} \in \mathbb{R}^m \;\middle|\; \begin{array}{l} \sum_{i=1}^m \bar{q}_i = 1 \\ \sum_{i\in J} \bar{q}_i \leq h\left(\sum_{i\in J} p_i\right), \ \forall J \subset [m] \\ \bar{q}_i \geq 0, \ \forall i \in [m] \end{array} \right\}. \tag{1.15}$$

As a consequence, Theorem 1.6 implies that the robust problem (P) is equivalent to the following rank-independent problem that is more suitable for robust optimization techniques:

$$\min_{\mathbf{a}\in\mathcal{A}} \sup_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_{\phi,h}(\mathbf{p})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)). \tag{P-ref}$$

Similarly, the nominal problem (P-Nom) can be reformulated to

$$\min_{\mathbf{a}\in\mathcal{A}} \sup_{\bar{\mathbf{q}}\in M_h(\mathbf{p})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)). \tag{P-Nom-ref}$$

Moreover, the equivalence between (P-EG) and (1.12) features an interpretation somewhat similar to that in Bertsimas and Brown [2009], which connects the shape of an uncertainty set to decision theory. Indeed, this equivalence can be viewed as a device for constructing a myriad of preference-based uncertainty sets $\mathcal{U}_{\phi,h}(\mathbf{p})$ for robust optimization. In Figure 1.5 of Appendix 1.E, we provide an illustrative visualization of the widely varying shapes of uncertainty sets $\mathcal{U}_{\phi,h}(\mathbf{p})$ that can be generated by selecting specific examples of the deterministic, univariate functions h and ϕ.

The following theorem establishes how the semi-infinite constraints (1.12) and (1.14) can be further reformulated into finitely many constraints, yielding a (more) tractable robust counterpart problem.

Theorem 1.7. *Let* $h : [0,1] \to [0,1]$ *be a concave distortion function. Then, we have that, for all* $(\mathbf{a}, c) \in \mathbb{R}^{n_a+1}$*, the inequality* (1.12) *holds if and only if there exist* α, β, γ*,*

$(\nu_j)_{j=1}^{2^m-2}, (\lambda_j)_{j=1}^{2^m-2} \in \mathbb{R}$, *such that*

$$\begin{cases} \alpha + \beta + \gamma r + \sum_{i=1}^{m} p_i \gamma \phi^* \left(\frac{-\alpha + \sum_{j:i\in I_j} \nu_j}{\gamma} \right) + \sum_{j=1}^{2^m-2} \lambda_j (-h)^* \left(\frac{-\nu_j}{\lambda_j} \right) \leq c, \\ -u(f(\mathbf{a}, \mathbf{x}_i)) - \beta - \sum_{j:i\in I_j} \lambda_j \leq 0, \qquad \forall i \in [m] \\ \lambda_j, \gamma \geq 0, \qquad \forall j \in [2^m - 2], \end{cases} \tag{1.16}$$

where $I_1, \ldots, I_{2^m-2}$ *are all subsets of* $[m]$*, except the empty set and the set* $[m]$ *itself.*

In the nominal case, we have that the inequality (1.14) *holds if and only if there exist* $\beta \in \mathbb{R}$, $(\lambda_j)_j \geq 0$ *such that*

$$\begin{cases} \beta + \sum_{j=1}^{2^m-2} \lambda_j h \left(\sum_{k\in I_j} p_k \right) \leq c, \\ -u(f(\mathbf{a}, \mathbf{x}_i)) - \beta - \sum_{j:i\in I_j} \lambda_j \leq 0, \ \forall i \in [m] \\ \beta \in \mathbb{R}, \lambda_j \geq 0, \ \forall j \in [2^m - 2]. \end{cases} \tag{1.17}$$

Remark 1.8. For technical reasons, the conjugate $(-h)^*$ in Theorem 1.7 is taken with respect to the domain $[0, \infty)$: $(-h)^*(y) = \sup_{t\geq 0}\{yt + h(t)\}$, by defining $h(x) = 1$ for all $x > 1$.

Remark 1.9. In Table 1.4, we provide $(-h)^*$ explicitly for a collection of canonical examples. (P-ref) may also be further reformulated using the "optimistic dual counterpart" [Gorissen et al., 2014]. This approach is particularly useful if $(-h)^*$ cannot be computed analytically, as is the case e.g., for $h(p) = \Phi(\Phi^{-1}(p) + \nu)$, $\nu > 0$, with Φ the standard normal cdf; see Wang [2000] and Goovaerts and Laeven [2008]. Further details are provided in Appendix 1.F.

1.3.2 Conic Representability of the Robust Counterpart

In this subsection, we explore the conic representability of the robust counterpart reformulated in Theorem 1.7. Following Ben-Tal and Nemirovski [2019], a set $\mathcal{S} \subset \mathbb{R}^n$ is conic representable by a cone $\mathbf{K}$ if and only if

$$\mathbf{x} \in \mathcal{S} \Leftrightarrow \exists\ \mathbf{w}, \mathbf{A}, \mathbf{b} : \mathbf{A} \begin{pmatrix} \mathbf{x} \\ \mathbf{w} \end{pmatrix} - \mathbf{b} \succeq_{\mathbf{K}} \mathbf{0}.$$

A function is said to be conic representable if its epigraph is. The reformulated robust counterpart (1.16) in Theorem 1.7 contains constraints that are expressed in the perspective of the univariate conjugate functions $(-h)^*$ and ϕ^*. We focus on the conic representability of these constraints:

$$\lambda(-h)^* \left(\frac{-\nu}{\lambda} \right) \leq z, \text{ and } \gamma \phi^* \left(\frac{s}{\gamma} \right) \leq t.$$

In practice, the derivation of the conjugate functions might be difficult. However, the epigraph representation of a conjugate function does not always require an explicit form of the conjugate function itself. The following two lemmas provide a generic approach for determining the conic representation of the epigraph of the perspective function through that of the function itself. Interestingly, this procedure does not require any derivation of the conjugate function. In Tables 1.4 and 1.5, we provide the explicit conic representation for many canonical examples of the conjugate functions $(-h)^*$ and ϕ^*, where the representations are composed of a combination of standard cones such as the quadratic, power, and exponential. These explicit conic representations are useful for the implementation of these constraints in standard optimization software such as CVXPY. Details on the derivations of the epigraph representations can be found in Appendix 1.C.

The following two lemmas originate from Ben-Tal and Nemirovski [2019] (Propositions 2.3.2 and 2.3.4), where they were stated only for the quadratic cones. However, they can be extended to general cones.

Lemma 1.10. *If f is conic representable with a cone $\mathbf{K}$, i.e., there exist $(\mathbf{A}, \mathbf{v}, \mathbf{B}, \mathbf{b})$ such that*

$$\mathrm{Epi}(f) = \{(\mathbf{x}, t) : \exists\ \mathbf{w} : \mathbf{A}\mathbf{x} + t\mathbf{v} + \mathbf{B}\mathbf{w} + \mathbf{b} \succeq_{\mathbf{K}} \mathbf{0}\},$$

and $\mathbf{A}\mathbf{x} + t\mathbf{v} + \mathbf{B}\mathbf{w} + \mathbf{b} \succ_{\mathbf{K}} \mathbf{0}$ for some $(\mathbf{x}, t, \mathbf{w})$, then f^* is conic representable with the dual cone $\mathbf{K}^*$:

$$\mathrm{Epi}(f^*) = \{(\mathbf{y}, s) : \exists\ \boldsymbol{\xi} \in \mathbf{K}^* : \mathbf{A}^T\boldsymbol{\xi} = -\mathbf{y}, \mathbf{B}^T\boldsymbol{\xi} = \mathbf{0}, \mathbf{v}^T\boldsymbol{\xi} = 1, s \geq \mathbf{b}^T\boldsymbol{\xi}\}.$$

In particular, f and f^* are representable by the same cone if $\mathbf{K}$ is self-adjoint.

Lemma 1.11. *If f is conic representable with a cone $\mathbf{K}$, then so is its perspective.*

We provide an example illustrating how the ideas in the proofs of Lemmas 1.10 and 1.11 can be systematically applied to determine the conic representation of the epigraph of the perspective transformation of a distortion function.

Example 1.12. Consider $h(p) = p^r(1 - \log(p^r)),\ 0 < r < 1$. Instead of deriving a closed-form expression of $(-h)^*$, which can be challenging in specific cases, we determine a conic representation of the epigraph $\lambda(-h)^*\left(\frac{-\nu}{\lambda}\right) \leq t$ by first determining a conic representation of $\mathrm{Epi}(-h)$. We have

$$(p, t) \in \mathrm{Epi}(-h) \Leftrightarrow \exists\ w \geq 0 : -w + w\log(w) \leq t, w \leq p^r, w \leq 1,$$

due to the fact that $x \mapsto -x + x\log(x)$ is decreasing on $[0, 1]$. Following the proof of

Lemma 1.10, we have that $(y, s) \in \text{Epi}((-h)^*)$ if and only if

$$-s \leq \min_{p,w\geq 0, t\in\mathbb{R}} \{-yp + t \mid -w + w\log(w) \leq t, w \leq p^r, w \leq 1\}.$$

Since this is a bounded convex problem satisfying Slater's condition, the duality theorem implies that the minimization problem is equal to

$$\begin{aligned}
&\max_{\xi_1,\xi_2,\xi_3\geq 0} -\xi_3 + \inf_{p,w\geq 0, t\in\mathbb{R}} \{-yp - \xi_2 p^r + (\xi_2 + \xi_3 - \xi_1)w + \xi_1 w\log(w) + (1-\xi_1)t\} \\
&= \max_{\xi_2,\xi_3\geq 0} -\xi_3 + \inf_{p,w\geq 0} \{-yp - \xi_2 p^r + (\xi_2 + \xi_3 - 1)w + w\log(w)\} \\
&= \max_{\xi_2,\xi_3\geq 0} \{-\xi_3 - e^{-\xi_2-\xi_3} + \left(r^{\frac{1}{1-r}} - r^{\frac{r}{1-r}}\right)\xi_2^{\frac{1}{1-r}}|y|^{1-\frac{1}{1-r}} \mid y \leq 0\},
\end{aligned}$$

where in the last equality we computed $\inf_{p\geq 0}\{-yp - \xi_2 p^r\}$ explicitly.[5] This gives that $(y, s) \in \text{Epi}((-h)^*)$ if and only if there exist $\xi_2, \xi_3, \xi_4 \geq 0$ such that

$$\xi_3 + e^{-\xi_2-\xi_3} + \left(r^{\frac{r}{1-r}} - r^{\frac{1}{1-r}}\right)\xi_4 \leq s, \quad \xi_2 \leq |y|^r \xi_4^{1-r}, \quad y \leq 0.$$

Lemma 1.11 then yields that $\lambda(-h)^*\left(\frac{-\nu}{\lambda}\right) \leq z$ if and only if there exist $\xi_2, \xi_3, \xi_4 \geq 0$ such that

$$\xi_3 + \lambda e^{-(\xi_2+\xi_3)/\lambda} + \left(r^{\frac{r}{1-r}} - r^{\frac{1}{1-r}}\right)\xi_4 \leq z, \quad \xi_2 \leq |\nu|^r \xi_4^{1-r}, \quad \nu \geq 0,$$

which is a combination of the power cone and the exponential cone.

In some cases, one can also calculate a conjugate function by writing it as an inf-convolution. For example, if f is a sum of individual f_i's: $f(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x})$, then the conjugate of a sum is the inf-convolution of the sum of the conjugates (see Theorem 2.1, Bertsimas and den Hertog, 2022):

$$\left(\sum_{i=1}^n f_i\right)^*(\mathbf{s}) = \inf_{\{\mathbf{v}^i\}_{i\in[n]}} \left\{\sum_{i=1}^n f_i^*(\mathbf{v}^i) \;\middle|\; \sum_{i=1}^n \mathbf{v}^i = \mathbf{s}\right\},$$

where the infimum can eventually be omitted when considering the epigraph formulation.

[5] $\inf_{p\geq 0} -yp - \xi_2 p^r = \begin{cases} \left(r^{\frac{1}{1-r}} - r^{\frac{r}{1-r}}\right)\xi_2^{\frac{1}{1-r}}|y|^{1-\frac{1}{1-r}} & y < 0, \quad \xi_2 > 0 \\ 0 & y \leq 0, \quad \xi_2 = 0 \\ -\infty & y \geq 0, \quad \xi_2 > 0. \end{cases}$

1.4 Solving the Robust Problem I: A Cutting-Plane Method

As shown in Theorem 1.7, the number of reformulated constraints grows exponentially, as 2^m, where m is the dimension of the probability vector. If m is of small or moderate size, then the reformulations in (1.16) and (1.17) can be applied to solve the minimization problems (P) and (P-Nom) exactly. However, for larger m, this becomes computationally intractable. In this section, we show how the reformulation of (P) to (P-ref) (and similarly for (P-Nom) to (P-Nom-ref)) by Theorem 1.6 enables us to devise a suitable cutting-plane method, which circumvents this curse of dimensionality.

1.4.1 The Cutting-Plane Algorithm

The standard cutting-plane algorithm is one of the most applied algorithms to solve robust optimization problems and has shown great efficiency in many applications (e.g., Mutapcic and Boyd, 2009, Bertsimas et al., 2016). The general idea is simple: Approximate the uncertainty set $\mathcal{U}$ with a suitably chosen subset $\mathcal{U}_j \subset \mathcal{U}$ and solve the corresponding robust problem, which is often a simpler problem, similar to the nominal problem. If the solution is feasible according to the robust evaluation with respect to the original set $\mathcal{U}$, then the process is terminated. Otherwise, the worst-case parameter in $\mathcal{U}$ associated with the current solution is added to $\mathcal{U}_j$, and the process is repeated.

In our case, we apply the cutting-plane method to the reformulated problem (P-ref), where the uncertainty set is given by the *composite uncertainty set* $\mathcal{U}_{\phi,h}(\mathbf{p})$. After obtaining a candidate solution using a suitable subset $\mathcal{U}_j$, we verify its feasibility with respect to the original uncertainty set $\mathcal{U}_{\phi,h}(\mathbf{p})$. This involves solving the following optimization problem, for a given solution $\mathbf{a}_*$:

$$\sup_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_{\phi,h}(\mathbf{p})} \sum_{i=1}^{m} -\bar{q}_i u(f(\mathbf{a}_*, \mathbf{x}_i)).$$

At first sight, this seems problematic due to the 2^m number of constraints in $\mathcal{U}_{\phi,h}(\mathbf{p})$. Fortunately, this can be avoided, since the equivalent robust rank-dependent evaluation is the optimization problem:

$$\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}_*, \mathbf{X})). \tag{1.18}$$

Problem (1.18) can be solved efficiently. Indeed, for a given solution $\mathbf{a}_*$, we can assess the ranking of the outcomes: $u(f(\mathbf{a}_j, \mathbf{x}_{(1)})) \geq \ldots \geq u(f(\mathbf{a}_j, \mathbf{x}_{(m)}))$. Then, by rewriting the alternating sum in (1.5), problem (1.18) becomes equal to the following convex optimization

problem:

$$\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \sum_{k=1}^{m} h\left(\sum_{j=k}^{m} q_{(j)}\right)\left(u(f(\mathbf{a}_*,\mathbf{x}_{(j-1)})) - u(f(\mathbf{a}_*,\mathbf{x}_{(j)}))\right), \tag{1.19}$$

where $u(f(\mathbf{a}_*,\mathbf{x}_{(0)})) := 0$. The optimal solution of (1.19) is a probability vector $\mathbf{q}^* \in \mathcal{D}_\phi(\mathbf{p},r)$. This probability vector further yields a $\bar{\mathbf{q}}^*$ such that $\bar{q}^*_{(i)} = h\left(\sum_{k=i}^{m} q^*_{(k)}\right) - h\left(\sum_{k=i+1}^{m} q^*_{(k)}\right)$ and $\bar{q}^*_{(m)} = h(q^*_{(m)})$, which is precisely the probability vector that constitutes the rank-dependent sum (1.5). The following lemma justifies that we may add the probability vectors $(\mathbf{q}^*, \bar{\mathbf{q}}^*)$ as the worst-case probabilities at each iteration of the cutting-plane procedure.

Lemma 1.13. *We have that* $(\mathbf{q}^*, \bar{\mathbf{q}}^*) \in \mathcal{U}_{\phi,h}(\mathbf{p})$.

We can now describe the cutting-plane procedure in full detail, which we do more precisely in Algorithm 1. We note that the cutting-plane algorithm always computes

Algorithm 1 Cutting-Plane Method

1: Start with $\mathcal{U}_1 = \{(\mathbf{p},\mathbf{p})\}$. Fix a tolerance parameter $\epsilon_{\text{tol}} > 0$.
2: At the j-th iteration, solve the following problem with the uncertainty set $\mathcal{U}_j$:

$$\min_{\mathbf{a}\in\mathcal{A}} \sup_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_j} -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)). \tag{1.20}$$

3: Let $(\mathbf{a}_j, c_j)$ be the optimal solution and objective value of (1.20). Determine a ranking of the realizations:

$$-u(f(\mathbf{a}_j,\mathbf{x}_{(1)})) \leq \ldots \leq -u(f(\mathbf{a}_j,\mathbf{x}_{(m)})).$$

Then, solve the optimization problem (1.19), which gives an optimal objective value v_j and a solution $(\mathbf{q}^*_j, \bar{\mathbf{q}}^*_j)$.
4: If $v_j - c_j \leq \epsilon_{\text{tol}}$, then the solution is accepted and the process is terminated.
5: If not, set $\mathcal{U}_{j+1} = \mathcal{U}_j \cup \{(\mathbf{q}^*_j, \bar{\mathbf{q}}^*_j)\}$ and repeat steps 2–5.

a *lower* bound c_j on the optimal objective value of (P-ref), since solving (1.20) at each iteration j with respect to a subset $\mathcal{U}_j \subset \mathcal{U}_{\phi,h}(\mathbf{p})$ is a less conservative problem. Moreover, the lower bound is improved iteratively, since $\mathcal{U}_j \subset \mathcal{U}_{j+1}$. Furthermore, by solving (1.19), which is a robust rank-dependent evaluation of a particular feasible solution, we obtain an *upper* bound v_j on (P-ref) at each iteration. Hence, the final step of the cutting-plane algorithm yields an upper and a lower bound on the exact optimal objective value (P-ref), with a gap value $v_j - c_j \leq \epsilon_{\text{tol}} > 0$ that can be chosen arbitrarily small.

The natural, formal question that arises is whether the cutting-plane algorithm actually terminates after finitely many iterations, thus yielding convergent upper and lower bounds as $\epsilon_{\text{tol}} \to 0$. The following theorem, which has its roots in Mutapcic and Boyd

[2009], states that, for any $\epsilon_{\text{tol}} > 0$, the termination is indeed guaranteed. We note that Mutapcic and Boyd [2009] imposes Lipschitz continuity conditions, which we avoid in our setting.

Theorem 1.14. *Let h be a concave distortion function. Suppose $\sup_{\mathbf{a}\in\mathcal{A},i\in[m]} u(f(\mathbf{a},\mathbf{x}_i)) < \infty$. Then, for all $\epsilon_{\text{tol}} > 0$, Algorithm 1 terminates after finitely many iterations.*

Remark 1.15. Algorithm 1 also applies to the nominal case (P-Nom), where $r \equiv 0$ in the set $\mathcal{U}_{\phi,h}(\mathbf{p})$. Furthermore, Theorem 1.14 still holds in this case, since its proof does not depend on the choice of r.

1.4.2 Robust Optimization with a Rank-Dependent Evaluation in the Constraint

In the previous subsection, we have shown that our cutting-plane algorithm yields convergent upper and lower bounds to problems (P-Nom)–(P), where the rank-dependent evaluation appears in the objective function. In this subsection, we discuss how the cutting-plane algorithm can be adapted to the robust problem (P-constraint), where the RDU model appears in the constraint, to provide a convergent lower bound. The adaptation of Algorithm 1 to problem (P-constraint) is presented in full detail in Algorithm 4 in Appendix 1.G.[6] The following theorem establishes the convergence of the cutting-plane algorithm.

Theorem 1.16. *Let h be concave. Suppose $\sup_{\mathbf{a}\in\mathcal{A},i\in[m]} u(f(\mathbf{a},\mathbf{x}_i)) < \infty$. Then, for all $\epsilon_{\text{tol}} > 0$, Algorithm 4 terminates after finitely many iterations. Moreover, if the function $\mathbf{a} \mapsto u(f(\mathbf{a},\mathbf{x}))$ is concave for all $\mathbf{x} \in \mathbb{R}^l$, then the optimal objective value of the final solution obtained from Algorithm 4 converges to that of* (P-constraint)*, as $\epsilon_{\text{tol}} \to 0$.*

Thus, using the cutting-plane algorithm, we can obtain lower bounds that converge to the exact optimal objective value of (P-constraint), as $\epsilon_{\text{tol}} \to 0$. Naturally, one would also like to obtain an upper bound on (P-constraint), as well as a feasible solution of (P-constraint). We note that this is not guaranteed by the cutting-plane algorithm as described in Algorithm 4, since it only provides a solution that is feasible within a tolerance $\epsilon_{\text{tol}} > 0$. With this objective in mind, we explore the rank-dependent nature of $\rho_{u,h,\mathbf{q}}(.)$ and propose a method to obtain an upper bound and a feasible solution of (P-constraint), which requires solving an optimization problem with only $3m + 3$ number of constraints. We first state the following definition.

Definition 1.17. Given an $\mathbf{a} \in \mathcal{A}$, we let $\mathcal{I}(\mathbf{a})$ denote the set of all permutations $(i_1,\ldots,i_m)$ of the index vector $(1,\ldots,m)$ such that the ranking $-u(f(\mathbf{a},\mathbf{x}_{i_1})) \leq \ldots \leq -u(f(\mathbf{a},\mathbf{x}_{i_m}))$ holds.

[6] In a nutshell, we replace (1.20) in Algorithm 1 by (P-constraint) adapted to the set $\mathcal{U}_j$ and replace c_j in step 4 by the constraint parameter c.

The idea is that for any given solution $\mathbf{a}_0 \in \mathcal{A}$, we can determine a ranking of the realizations $\{-u(f(\mathbf{a}_0, \mathbf{x}_i))\}_{i=1}^m$ and obtain a vector of indices $(i_1, \ldots, i_m) \in \mathcal{I}(\mathbf{a}_0)$. A natural upper bound on (P-constraint) can then be computed by solving the following optimization problem induced by the ranking $(i_1, \ldots, i_m)$ of $\mathbf{a}_0$:

$$U^*(\mathbf{a}_0) \triangleq \min_{\mathbf{a}\in\mathcal{A}} \left\{ g(\mathbf{a}) \;\middle|\; \sup_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_{\phi,h}^{(i_1,\ldots,i_m)}(\mathbf{p})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)) \le c \right\}, \tag{1.21}$$

where the rank-dependent uncertainty set is

$$\mathcal{U}_{\phi,h}^{(i_1,\ldots,i_m)}(\mathbf{p}) \triangleq \left\{ (\mathbf{q}, \bar{\mathbf{q}}) \in \mathbb{R}^{2m} \;\middle|\; \begin{array}{l} \sum_{i=1}^m q_i = \sum_{i=1}^m \bar{q}_i = 1 \\ \sum_{i=1}^m p_i \phi\left(\frac{q_i}{p_i}\right) \le r \\ \sum_{j=k}^m \bar{q}_{i_j} \le h\left(\sum_{j=k}^m q_{i_j}\right), \ \forall k \in [m] \\ q_i, \bar{q}_i \ge 0, \ \forall i \in [m] \end{array} \right\}.$$

Indeed, since $\mathcal{U}_{\phi,h}(\mathbf{p}) \subset \mathcal{U}_{\phi,h}^{(i_1,\ldots,i_m)}(\mathbf{p})$, solving (1.21) (which can be done using Theorem 1.7) yields an upper bound and a feasible solution of (P-constraint). In particular, if $\mathbf{a}_0$ is a good approximation of the optimal solution of (P-constraint), such that the ranking coincides, then $U^*(\mathbf{a}_0)$ is exactly equal to the optimal objective value of (P-constraint). This observation is made precise in the following lemma and is pivotal for developing a convergent upper bound for (P-constraint).

Lemma 1.18. *Let* $\mathbf{a}^*$ *be a minimizer of* (P-constraint)*. Then, for any* $\mathbf{a}_0 \in \mathcal{A}$ *such that* $\mathcal{I}(\mathbf{a}_0) \subset \mathcal{I}(\mathbf{a}^*)$*, we have that the value* $U^*(\mathbf{a}_0)$ *as defined in* (1.21) *is equal to the optimal objective value of* (P-constraint)*.*

Lemma 1.18 suggests that if one approximates the optimal solution of (P-constraint) with a sequence of cutting-plane solutions as $\epsilon_{\text{tol}} \to 0$, then under certain continuity conditions, the upper bounds computed in (1.21) also converge to the exact value. This is established in the following theorem.

Theorem 1.19. *Let* h *be concave. Assume* $\sup_{\mathbf{a}\in\mathcal{A}, i\in[m]} u(f(\mathbf{a}, \mathbf{x}_i)) < \infty$*. If the functions* $g(\mathbf{a})$ *and* $u(f(\mathbf{a}, \mathbf{x}_i))$ *are continuous on* $\mathcal{A}$*, for all* $i \in [m]$*, and* $\mathbf{a}_n$ *is a sequence of solutions obtained from Algorithm 4 with* $\epsilon_{\text{tol},n} \to 0$*, then* $(U^*(\mathbf{a}_n))_{n\ge1}$ *is a sequence of upper bounds that converges to the optimal objective value of* (P-constraint)*.*

1.4.3 Choquet Expected Utility

In this subsection, we briefly discuss how our cutting-plane method can also be extended to solve optimization problems for a more general rank-dependent model, namely the Choquet expected utility model [Schmeidler, 1986, 1989]. Let $c : \mathcal{F} \to [0, 1]$ be a monotone

set function[7] such that $c(\emptyset) = 0,\ c(\Omega) = 1$. The Choquet expected utility model evaluates a random variable X by the Choquet integral:

$$\int -X \,\mathrm{d}c \triangleq \int_0^\infty c(-X > t)\,\mathrm{d}t + \int_{-\infty}^0 (c(-X > t) - 1)\,\mathrm{d}t. \tag{1.22}$$

The rank-dependent evaluation $\rho_{u,h,\mathbb{Q}}$ as defined in (1.4) is a special case of the Choquet expected utility, where $c(A) = h(\mathbb{Q}(A))$. If c is a submodular set function,[8] then the Choquet expected utility can also be written as a worst-case expectation (see Denneberg, 1994), similar to Theorem 1.6:

$$\int -X \,\mathrm{d}c = \sup_{\bar{\mathbf{q}} \in \mathcal{U}_c} -\sum_{i=1}^m \bar{q}_i x_i, \quad \mathcal{U}_c \triangleq \left\{ \bar{\mathbf{q}} \in \mathbb{R}^m \;\middle|\; \begin{array}{l} \sum_{i=1}^m \bar{q}_i = 1 \\ \sum_{i \in J} \bar{q}_i \le c(J),\ \forall J \subset [m] \\ \bar{q}_i \ge 0,\ \forall i \in [m] \end{array} \right\}. \tag{1.23}$$

The uncertainty set $\mathcal{U}_c$ has essentially the same structure as the set $\mathcal{U}_{\phi,h}(\mathbf{p})$ defined in (1.13) and consists of only linear constraints. Therefore, the same procedure of the cutting-plane algorithm can be extended naturally to solve the more general rank-dependent minimization problem $\min_{\mathbf{a} \in \mathcal{A}} \int -f(\mathbf{a}, \mathbf{X})\,\mathrm{d}c$. By contrast, the piecewise-linear approximation methods, which we will discuss in Section 1.5, constitute a more restrictive, but as we will explicate in some cases more efficient, approach.

1.5 Solving the Robust Problem II: Piecewise-Linear Approximation

In this section, we develop a different method for computing lower and upper bounds on the optimal objective value of (P-ref), which relies on a suitable approximation of the distortion functions, and is leveraged in Section 1.6.

1.5.1 Piecewise-Linear Distortion Functions

In this subsection, we show that if the distortion function h is concave and piecewise-linear, then the exponential complexity of the constraints in $\mathcal{U}_{\phi,h}(\mathbf{p})$ can be reduced to the order of $m \cdot K$, where m is the dimension of the probability vector, and K is the number of linear pieces of h.

More precisely, we consider a function $h = \min_{j \in [K]} h_j$, with affine functions $h_j(p) = l_j \cdot p + b_j$, such that the slopes $l_1 > \cdots > l_K$ are decreasing and the intercepts $b_1 < \cdots < b_K$ are increasing. The affine functions are assumed to be defined on a set of support points

[7] $c(A) \le c(B)$, if $A \subset B$.
[8] $c(A \cap B) + c(A \cup B) \le c(A) + c(B),\ \forall A, B \in \mathcal{F}$.

$0 = x_0 < x_1 < \cdots < x_K = 1$, such that $h(p) = h_j(p)$, if $p \in [x_{j-1}, x_j]$, for all $j \in [K]$. Furthermore, we impose $b_1 \equiv 0$ and $l_K + b_K \equiv 1$, so that $h(0) = 0$ and $h(1) = 1$. We refer to this type of function as a *K-piecewise-linear distortion function*. Clearly, this function is non-decreasing and concave.

We prove the following lemma.

Lemma 1.20. *Let h be a K-piecewise-linear distortion function. Then, $(\mathbf{q}, \bar{\mathbf{q}}) \in \mathcal{U}_{\phi,h}(\mathbf{p})$ if and only if there exists $\mathbf{t} \triangleq (\{t_{i,j}\}_{i=1}^m)_{j=1}^K \in \mathbb{R}^{mK}$ such that the variables $(\mathbf{q}, \bar{\mathbf{q}}, \mathbf{t})$ satisfy the constraints*

$$\begin{cases} q_i, \bar{q}_i, t_{i,j} \geq 0, \ \forall i \in [m], \ \forall j \in [K] \\ \bar{q}_i \leq l_j \cdot q_i + t_{i,j}, \ \forall i \in [m], \ \forall j \in [K] \\ \sum_{i=1}^m t_{i,j} \leq b_j, \ \forall j \in [K] \\ \sum_{i=1}^m q_i = \sum_{i=1}^m \bar{q}_i = 1 \\ \sum_{i=1}^m p_i \phi\left(\frac{q_i}{p_i}\right) \leq r. \end{cases} \tag{1.24}$$

Hence, for K-piecewise-linear distortion functions, we obtain the following reformulation of the robust counterpart (1.12).

Theorem 1.21. *Let h be a K-piecewise-linear distortion function. Then, we have that $(\mathbf{a}, c)$ satisfies the constraint (1.12) if and only if there exist variables $\alpha, \beta, \gamma, \{\lambda_{ij}\}_{i\in[m],j\in[K]}$, $\{\nu_j\}_{j=1}^K \in \mathbb{R}$, such that*

$$\begin{cases} \alpha + \beta + \gamma r + \sum_{j=1}^K \nu_j b_j + \sum_{i=1}^m p_i \gamma \phi^*\left(\frac{-\alpha + \sum_{j=1}^K \lambda_{ij} l_j}{\gamma}\right) \leq c \\ -u(f(\mathbf{a}, \mathbf{x}_i)) - \beta - \sum_{j=1}^K \lambda_{ij} \leq 0, \ \forall i \in [m] \\ \lambda_{ij} \leq \nu_j, \ \forall i \in [m], \ \forall j \in [K] \\ \gamma, \lambda_{ij}, \nu_j \geq 0, \ \forall i \in [m], \ \forall j \in [K]. \end{cases} \tag{1.25}$$

In the nominal case, we have that (1.14) holds if and only if there exist variables $\beta, \lambda_{ij}, \nu_j \in \mathbb{R}$ such that (1.25) holds with $\alpha = \gamma \equiv 0$, whence the first line becomes $\beta + \sum_{j=1}^K \nu_j b_j + \sum_{i=1}^m \sum_{j=1}^K \lambda_{ij} l_j p_i \leq c$.

Thus, for K-piecewise-linear distortion functions, the exponential complexity in the uncertainty set $\mathcal{U}_{\phi,h}(\mathbf{p})$ can be reduced. This result also provides a tractable approximation for general non-piecewise-linear concave distortion functions, as we outline in the following subsection, exploiting that each concave function can be approximated uniformly by a piecewise-linear concave function.

1.5.2 Piecewise-Linear Approximation of General Concave Distortion Functions

For a general concave h, we can reduce the complexity in $\mathcal{U}_{\phi,h}(\mathbf{p})$ by approximating h with a piecewise-linear function h_ϵ, such that $\max_{x\in[0,1]} |h(x) - h_\epsilon(x)| \leq \epsilon$ for any given $\epsilon > 0$. An upper approximation $h_\epsilon \geq h$ and a lower approximation $h_\epsilon \leq h$ yield an upper and lower bound on the optimal objective value of (P-ref), respectively. Uniform approximation of general concave functions with piecewise-linear functions has been studied in the literature (see, e.g., Imamoto and Tang, 2008, Cox, 1971).

More specifically, using the concavity of h, one can approximate h from below by choosing a set of support points $\{x_i\}_{i=0}^K$ and considering the concave piecewise-linear function that connects the values $\{h(x_i)\}_i$. Given an $\epsilon > 0$, the required number of support points K can be minimized. We briefly describe our approach of determining the minimal K support points $\{x_i\}_{i=0}^K$ such that h is lower approximated by a piecewise-linear function with a prescribed error ϵ. Set $x_0 = 0$. At the i-th iteration, we choose the next support point x_{i+1} such that the maximal error is equal to ϵ:

$$e_i(x_{i+1}) \triangleq \sup_{x\in[x_i,x_{i+1}]} \left\{ h(x) - \frac{h(x_{i+1}) - h(x_i)}{x_{i+1} - x_i}(x - x_i) - h(x_i) \right\} = \epsilon. \tag{1.26}$$

If $e_i(1) \leq \epsilon$, then we simply choose $x_{i+1} = 1$. Otherwise, we choose x_{i+1} such that (1.26) holds. By construction, the piecewise-linear approximation h_ϵ induced by this set of support points has a maximum approximation error of ϵ and satisfies $h_\epsilon \leq h$, due to concavity. The existence of such x_{i+1} at each iteration is verified in the following lemma. Moreover, it implies that we can solve (1.26) for x_{i+1} using the bisection method.

Lemma 1.22. *Let h be a continuous, increasing, strictly concave function on $[0, 1]$. Then, $e_i(x_{i+1})$ is an increasing and continuous function in x_{i+1}. In particular, for any $\epsilon > 0$ and iteration i, if $e_i(1) > \epsilon$, then there exists a $x_{i+1} \in (x_i, 1)$ such that $e_i(x_{i+1}) = \epsilon$.*

1.5.3 The Piecewise-Linear Approximation Method and its Convergence

Given a piecewise-linear lower approximation $h_\epsilon \leq h$ of h, we can approximate the uncertainty set $\mathcal{U}_{\phi,h}(\mathbf{p})$ with $\mathcal{U}_{\phi,h_\epsilon}$ and apply Theorem 1.21. This yields a lower bound on (P-ref). Similarly, the concave distortion function $\tilde{h}_\epsilon$, which we define as

$$\tilde{h}_\epsilon(p) \triangleq \begin{cases} 0, & p = 0 \\ \min\{h_\epsilon(p) + \epsilon, 1\}, & 0 < p \leq 1, \end{cases} \tag{1.27}$$

yields an upper approximation of h with uniform error ϵ. Hence, we can also approximate $\mathcal{U}_{\phi,h}(\mathbf{p})$ by $\mathcal{U}_{\phi,\tilde{h}_\epsilon}(\mathbf{p})$ and apply Theorem 1.21 (note that the constraints (1.25) for $\tilde{h}_\epsilon$ are the same as those for h_ϵ, but with b_j replaced by $b_j + \epsilon$). This yields an upper bound for (P-ref) since $\mathcal{U}_{\phi,h}(\mathbf{p}) \subset \mathcal{U}_{\phi,\tilde{h}_\epsilon}(\mathbf{p})$. Therefore, we obtain an upper and lower bound using the piecewise-linear approximation method that we summarize in Algorithm 2. We also show that both bounds converge to the optimal objective value of the exact problem (P-ref), if the approximation error ϵ approaches zero. Thus, Algorithm 2 terminates for any parameter $\delta > 0$. This also applies, *mutatis mutandis*, to problem (P-constraint).

Algorithm 2 Piecewise-Linear Approximation Method

1: Set an $\epsilon > 0$. Approximate h from below by a K-piecewise-linear distortion function h_ϵ such that: (1) The number of pieces, K, is minimized; (2) $h_\epsilon(x) \leq h(x)$, $\forall x \in [0,1]$; (3) $\sup_{x\in[0,1]} h(x) - h_\epsilon(x) \leq \epsilon$.
2: Solve problem (P-ref) with $\mathcal{U}_{\phi,h}(\mathbf{p})$ replaced by $\mathcal{U}_{\phi,h_\epsilon}(\mathbf{p})$, using Theorem 1.21. This gives a lower bound L^* and an optimal solution $\mathbf{a}^*$. Do the same with $\tilde{h}_\epsilon$ to obtain an upper bound U^*.
3: If $U^* - L^* < \delta$ for some prescribed $\delta > 0$, then we take $\mathbf{a}^*$ as the final solution.
4: Otherwise, set $\epsilon \to \epsilon/2$ and perform all previous steps to obtain new bounds and solutions.

Theorem 1.23. *Let h be concave. Suppose $\sup_{\mathbf{a}\in\mathcal{A},i\in[m]} u(f(\mathbf{a},\mathbf{x}_i)) < \infty$. Then, Algorithm 2 terminates after finitely many iterations, for any $\epsilon, \delta > 0$. If the function $\mathbf{a} \mapsto u(f(\mathbf{a},\mathbf{x}))$ is concave for all $\mathbf{x} \in \mathbb{R}^l$, then this also holds when Algorithm 2 is applied to problem* (P-constraint).

1.6 Non-Concave Distortion Functions

In the previous sections, we have discussed how to optimize a rank-dependent model when the distortion function is concave. We now extend these ideas to convex and inverse S-shaped distortion functions, which are often found in empirical work (see e.g., Wakker, 2010, Prelec, 1998). Optimization problems with non-concave distortion functions are challenging due to their non-convex nature. Moreover, rank-dependent models with non-concave distortion functions lack a dual representation, which was a pivotal tool that enabled us to reformulate the rank-dependent problem. In the recent literature (e.g., Cai et al., 2023 and Pesenti et al., 2020), uncertainty sets with a specific structure have been identified for which optimizing the robust distortion risk measure with non-concave distortion function is equivalent to optimizing the same model with h replaced by its concave envelope $\hat{h}$.[9] However, this equivalence is typically not satisfied in our setting, as we state in the following proposition.

[9]The concave envelope of a function h is the smallest concave function that dominates h.

Proposition 1.24. *Let $\mathcal{U}$ be a compact set of probability vectors. Enumerate the realizations of X as $x_1 > \ldots > x_m$ and suppose that $\mathcal{U}$ does not contain the probability vector for which $q_1 = 1$, which is concentrated on x_1. Then, there exists a continuous distortion function h such that*

$$\sup_{\mathbf{q}\in\mathcal{U}} \rho_{h,\mathbf{q}}(X) < \sup_{\mathbf{q}\in\mathcal{U}} \rho_{\hat{h},\mathbf{q}}(X).$$

Therefore, it is necessary to study how to reformulate a robust or nominal rank-dependent evaluation (constraint), in the case of a convex or inverse S-shaped distortion function. We will first study how to reformulate a nominal constraint of the form (P-Nom-EG).

1.6.1 The Nominal Problem

Although rank-dependent models do not admit a dual representation for general non-concave distortion functions, we can still express them as inner robust optimization problems if the distortion function is convex or inverse S-shaped. The idea is to treat the concave and convex parts of the inverse S-shaped function separately, where the convex part is transformed into a concave part via the dual function $\bar{h}(p) \triangleq 1 - h(1-p)$. This result is stated in the following theorem. Henceforth, the utility function is allowed to be non-concave.

Theorem 1.25. *Let $h : [0,1] \to [0,1]$ be an inverse S-shaped distortion function with $p^0 \in [0,1]$ as in Definition 1.1. Define the dual function $\bar{h}(p) \triangleq 1 - h(1-p)$. Then, we have that, for all $(\mathbf{a}, c) \in \mathbb{R}^{n_a+1}$,* (P-Nom-EG) *is satisfied if and only if there exist $z \in \mathbb{R}, \bar{\mathbf{q}} \in N_{\bar{h}}^{cv}(\mathbf{p})$ such that*

$$\begin{cases} z + \displaystyle\sum_{i=1}^{m} -\bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)) \leq c \\ \displaystyle\sup_{\mathbf{q}\in M_h^{ca}(\mathbf{p})} \sum_{i=1}^{m} -q_i u(f(\mathbf{a}, \mathbf{x}_i)) \leq z, \end{cases} \tag{1.28}$$

where

$$M_h^{ca}(\mathbf{p}) \triangleq \left\{ \mathbf{q} \in \mathbb{R}^m \;\middle|\; \begin{array}{l} q_i \geq 0 \\ \sum_{i\in J} q_i \leq h\left(\sum_{i\in J} p_i\right), \ \forall J \subset [m] : \sum_{i\in J} p_i \leq p^0 \\ \sum_{i=1}^m q_i = h(p^0) \end{array} \right\}, \tag{1.29}$$

and

$$N_{\bar{h}}^{cv}(\mathbf{p}) \triangleq \left\{ \bar{\mathbf{q}} \in \mathbb{R}^m \;\middle|\; \begin{array}{l} \bar{q}_i \geq 0 \\ \sum_{i\in J} \bar{q}_i \leq \bar{h}\left(\sum_{i\in J} p_j\right), \ \forall J \subset [m] : \sum_{i\in J} p_i \leq 1 - p^0 \\ \sum_{i=1}^m \bar{q}_i = \bar{h}(1-p^0) \end{array} \right\}. \tag{1.30}$$

Remark 1.26. In particular, if h is convex, we have that

$$\rho_{u,h,\mathbf{p}}(f(\mathbf{a},\mathbf{X})) = -\sup_{\bar{\mathbf{q}}\in N_h^{\mathrm{cv}}(\mathbf{p})} \sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)), \tag{1.31}$$

with $p^0 \equiv 0$.

Theorem 1.25 shows that (P-Nom-EG) can be reformulated into rank-independent constraints (1.28), where the supremum constraint can be reformulated further as in Theorem 1.7. However, both (1.29) and (1.30) still feature an exponential number of constraints. As outlined in the previous section, we can circumvent this using piecewise-linear approximation. Let h be a piecewise-linear, inverse S-shaped distortion function with its concave and convex parts specified by the concave functions $h_l, \bar{h}_l$ (note that $\bar{h}_l$ is the dual of the convex part of h), respectively, on the domains $[0, p^0]$ and $[0, 1-p^0]$. Due to concavity and monotonicity, they can be expressed as minima of affine functions:

$$h_l(p) = \min\{l_1^{(1)}p + b_1^{(1)}, \ldots, l_{K_1}^{(1)}p + b_{K_1}^{(1)}\}, \quad \bar{h}_l(p) = \min\{l_1^{(2)}p + b_1^{(2)}, \ldots, l_{K_2}^{(2)}p + b_{K_2}^{(2)}\}, \tag{1.32}$$

defined on the support points

$$0 = s_0^{(1)} \le s_1^{(1)}, \ldots, s_{K_1-1}^{(1)} \le s_{K_1}^{(1)} = p^0, \quad 0 = s_0^{(2)} \le s_1^{(2)}, \ldots, s_{K_2-1}^{(2)} \le s_{K_2}^{(2)} = 1-p^0,$$

such that, for all $k \in \{1, \ldots, K_1\}, v \in \{1, \ldots, K_2\}$,

$$h_l(p) = l_k^{(1)}p + b_k^{(1)}, \forall p \in [s_{k-1}^{(1)}, s_k^{(1)}],\ \bar{h}_l(p) = l_v^{(2)}p + b_v^{(2)}, \forall p \in [s_{v-1}^{(2)}, s_v^{(2)}]. \tag{1.33}$$

The following theorem establishes a reformulation of the constraints in (1.28), when h is piecewise-linear.

Theorem 1.27. *Let h be a piecewise-linear, inverse S-shaped distortion function with its concave and convex part specified by $h_l, \bar{h}_l$ as in* (1.32)*. Then, for any $(\mathbf{a}, c) \in \mathbb{R}^{n_a+1}$,* (P-Nom-EG) *is satisfied if and only if there exist variables $\lambda_{ik}, \nu_k, t_{ik}, \bar{q}_i \ge 0$, $\beta \in \mathbb{R}$ such that*

$$\begin{cases} \beta \cdot h(p^0) + \sum_{k=1}^{K_1} \nu_k b_k^{(1)} + \sum_{i=1}^{m}\sum_{k=1}^{K_1} \lambda_{ik} l_k^{(1)} p_i - \sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)) \le c \\ -u(f(\mathbf{a},\mathbf{x}_i)) - \beta - \sum_{k=1}^{K_1} \lambda_{ik} \le 0,\ \forall i \in [m] \\ \lambda_{ik} \le \nu_k,\ \forall i \in [m],\ \forall k \in [K_1] \\ \bar{q}_i \le l_k^{(2)} p_i + t_{ik},\ \forall i \in [m],\ \forall k \in [K_2] \\ \sum_{i=1}^{m} t_{ik} \le b_k^{(2)},\ \forall k \in [K_2] \\ \sum_{i=1}^{m} \bar{q}_i = \bar{h}(1-p^0). \end{cases} \tag{1.34}$$

Importantly, this yields a problem with only $O(m \cdot K_1 \vee K_2)$ constraints. We note that, not surprisingly, (1.34) contains a non-convex constraint due to the product term $\bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i))$. Hence, the computation of (P-Nom-EG) with the reformulated constraints in (1.34) requires a mixed-integer nonlinear programming (MINLP) solver, such as BARON that uses the branch and reduce search method to obtain a global optimum (see e.g., Ryoo and Sahinidis, 1996). In particular, if the set $\mathcal{A}$ is polyhedral (i.e., $\mathcal{A} = \{\mathbf{a} \in \mathbb{R}^{n_a} : \mathbf{D}\mathbf{a} \leq \mathbf{d}\}$), $f(\mathbf{a}, \mathbf{x}) = \mathbf{a}^T\mathbf{x}$ and u is piecewise-linear, then the nomimal problem (P-Nom-EG) with constraints (1.34) can also be solved efficiently by Gurobi (Gurobi Optimization, LLC, 2023), using bilinear and special ordered sets of type 2 (SOS2) constraints; see the precise formulation in (1.73) in Appendix 1.H, further elaboration on the SOS2 constraints in (1.36), and a numerical study on the performance of Gurobi in this setting in Section 1.7.4.

1.6.2 Robust Rank-Dependent Models with Inverse S-Shaped Distortion

In this subsection, we study the robust constraint (P-EG) when h is inverse-S-shaped. We note that then (P-EG) cannot be reformulated using a composite uncertainty set as in (1.12), since the convex part of the distortion function gives rise to a sup-inf term when combining the set $\mathcal{D}_\phi(\mathbf{p}, r)$ with (1.30). However, we can still circumvent this using a suitable cutting-plane approach, where we iteratively solve a nominal problem using (1.34), and compute the robust rank-dependent evaluation for each nominal solution.

Indeed, for any given feasible solution $\mathbf{a}_* \in \mathcal{A}$, we can assess the ranking of the outcomes $u(f(\mathbf{a}^T, \mathbf{x}_{(1)})) \geq \ldots \geq u(f(\mathbf{a}^T, \mathbf{x}_{(m)}))$. Then, we calculate the robust rank-dependent evaluation:

$$\sup_{\mathbf{q} \in \mathcal{D}_\phi(\mathbf{p}, r)} \sum_{i=1}^{m} h\left(\sum_{k=i}^{m} q_{(k)}\right) \left(u(f(\mathbf{a}_*), \mathbf{x}_{(i-1)})) - u(f(\mathbf{a}_*, \mathbf{x}_{(i)}))\right). \tag{1.35}$$

Problem (1.35) is, quite naturally, non-convex and can again be computed using a solver such as BARON. If h is piecewise-linear and the divergence function ϕ is linear or quadratic, then one can also solve (1.35) using SOS2 constraints in Gurobi. Specifically, given a set of support points $(t_j, h(t_j))_{j=0}^K$, where $0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = 1$, the

SOS2 constraints can be formulated as follows:

$$\begin{aligned}
\max_{\substack{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)\\ \boldsymbol{\lambda}\in\mathbb{R}^{K\times m}\\ \mathbf{z}\in\mathbb{R}^m}} \quad & -u(f(\mathbf{a}_*,\mathbf{x}_{(1)})) + \sum_{i=2}^{m} z_i(u(f(\mathbf{a}_*,\mathbf{x}_{(i-1)})) - u(f(\mathbf{a}_*,\mathbf{x}_{(i)})))\\
\text{subject to } & \sum_{k=i}^{m} q_{(k)} = \sum_{j=1}^{K} \lambda_{ij} t_j,\ \forall i \in \{2,\dots,m\}\\
& z_i = \sum_{j=1}^{K} \lambda_{ij} h(t_j),\ \forall i \in \{2,\dots,m\}\\
& \sum_{j=1}^{K} \lambda_{ij} = 1,\ \forall i \in [m]\\
& \lambda_{ij} \geq 0, \text{SOS2},\ \forall i \in [m],\ \forall j \in [K].
\end{aligned} \tag{1.36}$$

The SOS2 constraints control, for each $i \in [m]$, the variables $\{\lambda_{ij}\}_{j=1}^{K}$ in a way such that only two adjacent $\lambda_{ij}, \lambda_{i,j+1}$ can be non-zero. Since the support points $\{t_j\}_{j=0}^{K}$ are ordered, the two non-zero adjacent variables $\lambda_{ij}, \lambda_{i,j+1}$ ensure that we are only optimizing within each interval $[t_j, t_{j+1}]$.

Thus, we can now introduce Algorithm 3, the cutting-plane approach that enables us to solve the robust problem (P) for piecewise-linear, inverse S-shaped distortion functions.

Algorithm 3 Cutting-Plane Method with Inverse S-Shaped Distortion Function

1: Start with $\mathcal{U}_1 = \{\mathbf{p}\}$. Fix a tolerance parameter $\epsilon_{\text{tol}} > 0$.
2: At the j-th iteration, solve the following problem with the uncertainty set $\mathcal{U}_j$, to obtain a solution $\mathbf{a}_j$ and optimal objective value c_j:

$$\min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{U}_j} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})), \tag{1.37}$$

by applying the reformulation in Theorem 1.27 to each element of $\mathcal{U}_j$.
3: Determine the robust rank-dependent evaluation v_j of $\mathbf{a}_j$ by solving (1.36), which gives an optimal solution $\mathbf{q}_j^*$.
4: If $v_j - c_j \leq \epsilon_{\text{tol}}$, then the solution $\mathbf{a}_j$ is accepted and the process terminated.
5: If not, set $\mathcal{U}_{j+1} = \mathcal{U}_j \cup \{\mathbf{q}_j^*\}$, and repeat steps 2–5.

The following theorem establishes that this cutting-plane method terminates after finitely many iterations.

Theorem 1.28. *Let h be an inverse S-shaped, piecewise-linear distortion function. Suppose that $\sup_{\mathbf{a}\in\mathcal{A},i\in[m]} |u(f(\mathbf{a},\mathbf{x}_i))| < \infty$. Then, for all $\epsilon_{\text{tol}} > 0$, Algorithm 3 terminates after finitely many iterations.*

Next, we examine the convergence of the piecewise-linear approximation to problems (P-Nom) and (P). In conjunction with Theorem 1.28, it constitutes, in a sense, the

master theorem of the paper.

Theorem 1.29. *Let h be an inverse S-shaped distortion function and let h_ϵ be such that $\sup_{p\in[0,1]} |h(p)-h_\epsilon(p)| \leq \epsilon$. Suppose that $\sup_{\mathbf{a}\in\mathcal{A},i\in[m]} |u(f(\mathbf{a},\mathbf{x}_i))| < \infty$. Then, the optimal objective value of* (P-Nom) *for h_ϵ converges to that of h, as $\epsilon \to 0$.*

Similarly, for the robust problem (P)*, Algorithm 3, if applied to h_ϵ, yields an optimal objective value that converges to the exact optimal objective value of* (P) *for h, as both $\epsilon_{\text{tol}}, \epsilon \to 0$.*

We note that when h is an inverse-S-shaped distortion function, the piecewise-linear approximation specified by $\{h_l, \bar{h}_l\}$ in (1.32) provides neither a lower bound nor an upper bound on $h(p)$, for all $p \in [0,1]$. Indeed, h_l is a lower approximation of h on $[0, p^0]$, whereas $1-\bar{h}_l(1-p)$ is an upper approximation of h on $[p_0, 1]$. Nonetheless, if one aims to bound h on the entire $[0,1]$, then we can apply the same device as in (1.27): either translate h_l with an ϵ (the maximal error) to obtain an upper piecewise-linear approximation of h on $[0, p^0]$, or do the same with the dual function $\bar{h}_l$ for a lower approximation on $[p^0, 1]$. Theorem 1.29 then implies that the corresponding lower and upper bounds on the optimal objective value of (P-Nom) (or (P)), computed from these piecewise-linear approximations of h, will converge to the exact optimal value, as $\epsilon \to 0$.

Finally, we note that Algorithm 3 can also be adapted (similar to Algorithm 4 in the Appendix) to solve problem (P-constraint), where the rank-dependent evaluation with inverse-S-shaped distortion function is in the constraint. Hence, we also present the following theorems with similar statements as in Theorems 1.28 and 1.29.

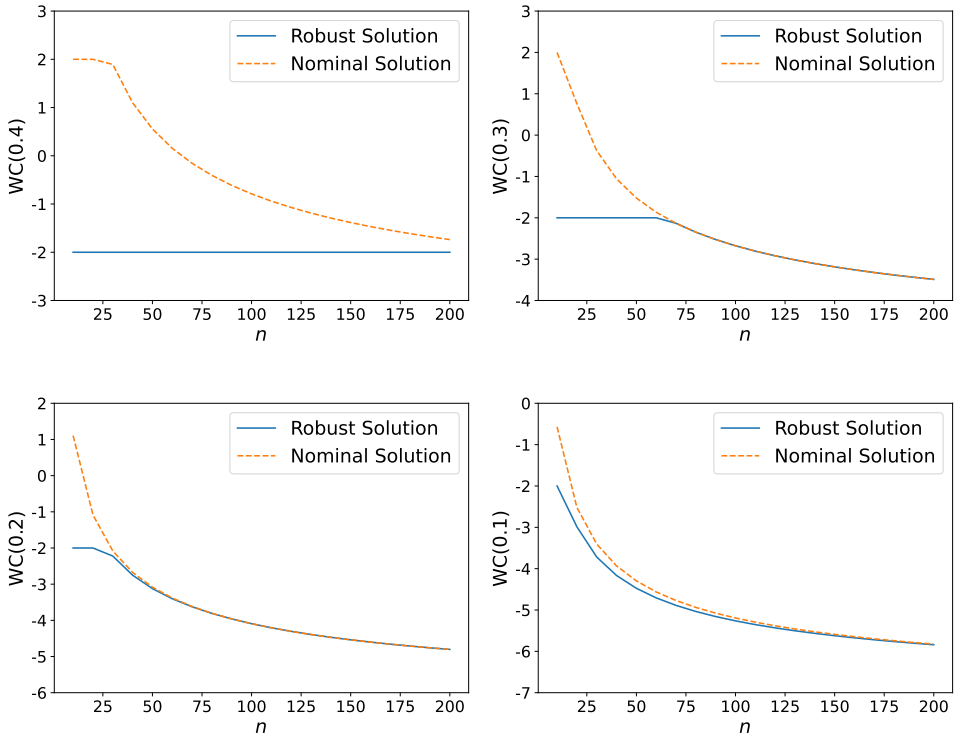
Theorem 1.30. *Let h be an inverse S-shaped, piecewise-linear distortion function. Suppose that $u(f(\mathbf{a},\mathbf{x}_i))$ is continuous in $\mathbf{a} \in \mathcal{A}$ for all $i \in [m]$. Then, for any $\epsilon_{\text{tol}} > 0$, Algorithm 3 terminates after finitely many iterations when applied to* (P-constraint)*. Furthermore, if h is also continuous, then the final solution $\mathbf{a}_{\epsilon_{\text{tol}}}$ of Algorithm 3 gives a lower bound $g(\mathbf{a}_{\epsilon_{\text{tol}}})$ that converges to the optimal objective value of* (P-constraint) *as $\epsilon_{\text{tol}} \to 0$.*

Theorem 1.31. *Let h be an inverse S-shaped distortion function. Suppose that $u(f(\mathbf{a},\mathbf{x}_i))$ is continuous in $\mathbf{a} \in \mathcal{A}$ for all $i \in [m]$. If g, appearing in the objective, and h are continuous, then for any distortion function h_ϵ such that $h_\epsilon(p) \leq h(p)$ for all $p \in [0,1]$ and $\sup_{p\in[0,1]} |h(p) - h_\epsilon(p)| \leq \epsilon$, the optimal objective value of* (P-constraint) *for h_ϵ converges to that of h, as $\epsilon \to 0$.*

1.7 Numerical Examples

In this section, we illustrate the various methods developed in this paper by applying them to two canonical examples of optimization problems: the *newsvendor* problem and the *portfolio choice* problem. In both examples, we compare the solution of the robust problem

Figure 1.1: Single-item newsvendor problem. This figure displays the worst-case evaluation $\mathrm{WC}(\alpha_0) \triangleq \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r(n))} \mathrm{CVaR}_{1-\alpha_0}(.)$ under the robust and nominal solutions, for a range of values of $r(n) = \chi^2_{2,0.95}/(2n)$ and $\alpha_0 = 0.4, 0.3, 0.2, 0.1$.



(P) to the solution of the nominal problem (P-Nom). We use the phrase "robust/nominal solution" to refer to the solution of the robust/nominal problem.

1.7.1 Robust Single-Item Newsvendor

In the single-item newsvendor problem, a seller is uncertain about the demand for a certain product, and has to decide in advance how many units of the product have to be ordered and stocked in the inventory. Let $d_i \geq 0$ denote the realization of the demand in state i. Furthermore, let: c be the ordering cost of one unit, $v > c$ be the selling price, $s < c$ be the salvage value per unsold item returned to the factory, and l be the loss per unit of unmet demand, which may include both the cost of a lost sale and a penalty for the lost customer goodwill. Finally, let y be the number of items ordered, i.e., the decision variable. The profit function $\pi(d_i, y)$ is defined as

$$\begin{aligned}\pi(d_i, y) &\triangleq v\min\{d_i, y\} + s(y - d_i)_+ - l(d_i - y)_+ - cy \\ &= (s - v)(y - d_i)_+ - l(d_i - y)_+ + (v - c)y. \end{aligned} \tag{1.38}$$

Note that $\pi(d_i, y)$ is concave in y since it is a sum of concave piecewise-linear functions in y for all i. Assume that the demand $d_i \in \{4, 8, 10\}$, which corresponds to low, medium,

and high demand, with nominal probabilities $\mathbf{p} = \{0.375, 0.375, 0.25\}$. Furthermore, we assume that the number of items ordered will not exceed the maximum demand, i.e., $y \leq 10$. Finally, we set the parameter values $c = 4, v = 6, s = 2, l = 4$.

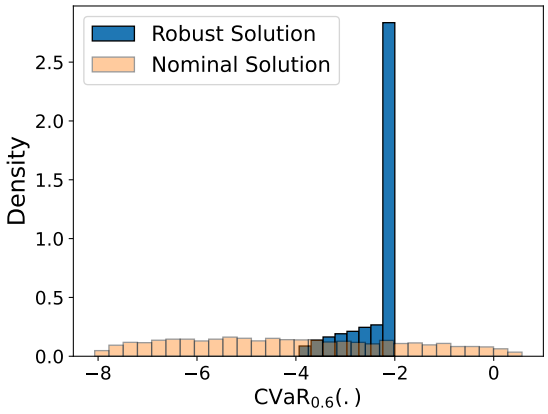
We solve both the robust problem (P) and the nominal problem (P-Nom), where the decision variable y is subject to the constraint $0 \leq y \leq 10$, $\mathbf{d}$ is the uncertain parameter, and $\pi(\mathbf{d}, y)$ is the random objective function. The nominal and robust newsvendor problem is studied in Eeckhoudt et al. [1995] and Ben-Tal et al. [2013], respectively, assuming expected utility preferences. We choose $\text{CVaR}_{1-\alpha_0}$ (see (1.6)) to be our rank-dependent evaluation $\rho_{u,h,\mathbf{q}}$. This corresponds to a piecewise-linear distortion function $h(p) = \min\{p/(1-\alpha_0), 1\}$ and a linear utility function $u(x) = x$. For the robust problem (P), we choose the KL-divergence function $\phi(t) = t\log(t) - t + 1$, and set the radius of the divergence set $\mathcal{D}_\phi(\mathbf{p}, r)$ to be $r = \chi^2_{2,0.95}/(2n)$ as in (1.10), where n is the sample size that we assume that $\mathbf{p}$ is estimated from. A larger value of the objective function indicates a larger (i.e., worse) evaluation of the utility loss; a value of $\alpha_0 = 0.4$ means that the worst 60% of the utility loss distribution is considered.

As the cardinality of realizations of the uncertain parameter d is merely 3, we can readily invoke Theorem 1.7 to solve both the reformulated nominal and robust problems (P-ref) and (P-Nom-ref) exactly, without reducing the number of constraints. We then perform the following experiment: For each α_0, we first obtain a nominal solution to (P-Nom). Then, for each radius $r(n) = \chi^2_{2,0.95}/(2n)$, we obtain a robust solution to (P). To compare the robust solution with the nominal solution, we calculate their worst-case evaluation $\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r(n))} \text{CVaR}_{1-\alpha_0}(.)$ under the radius $r(n)$. This is repeated for a range of $n = 10, 20, \ldots, 200$, and $\alpha_0 = 0.4, 0.3, 0.2, 0.1$. As n increases, the ϕ-divergence radius r decreases, and thus there is less ambiguity.

The results are displayed in Figure 1.1. Quite naturally, we observe that the differences between the worst-case evaluations of both solutions decrease as n increases. We also observe a decrease in the value of the worst-case evaluation as α_0 decreases, reflecting that the $\text{CVaR}_{1-\alpha_0}$ risk measure itself is monotonically decreasing in α_0. Furthermore, we see that for lower values of α_0, the differences between the worst-case evaluations of both solutions are already small for small sample size n. Moreover, we observe a flat, constant worst-case evaluation of the robust solution at level -2. This is the largest (i.e., worst) value that the robust rank-dependent evaluation attains for a robust solution $y^* = 7$. This conservativeness occurs when n is relatively small.

To further illustrate the differences between the robust and nominal solutions, we use the Hit-and-Run algorithm (see Appendix 1.I for further details) to sample 5,000 probability vectors $\mathbf{q}$ from the KL-divergence uncertainty set for $n = 50$. For each sampled vector $\mathbf{q}$, we calculate the $\text{CVaR}_{1-\alpha_0}$ evaluation, with $\alpha_0 = 0.4$. As shown in Figure 1.2, the evaluation of the nominal solution exhibits a large variance. It also exceeds the largest (i.e., worst) evaluation of the robust solution. On the other hand, the robust solution

Figure 1.2: Single-item newsvendor problem (*continued*): This figure displays, for each sampled probability vector $\mathbf{q}$ from the KL-divergence uncertainty set for $n = 50$, the corresponding $\text{CVaR}_{1-\alpha_0}(.)$ evaluation with $\alpha_0 = 0.4$ of the robust and nominal solutions.



shows less variance and is concentrated at the value -2, which as mentioned above, is the most conservative evaluation that the robust optimal solution attains.

1.7.2 Robust Multi-Item Newsvendor

We next examine the multi-item newsvendor problem, where each item in a set of items features its own uncertain demand. Let $d_i^{(j)}$ be the i-th realization of the j-th item's demand, $i = 1, 2, 3$. We take $j = 1, 2, 3$ and consider the sum of the individual profit functions

$$\pi_{\text{tot}}(\mathbf{d}, \mathbf{y}) = \pi_1(d^{(1)}, y_1) + \pi_2(d^{(2)}, y_2) + \pi_3(d^{(3)}, y_3), \tag{1.39}$$

where

$$\pi_j(d^{(j)}, y_j) = v_j \min\{d^{(j)}, y_j\} + s_j(y_j - d^{(j)})_+ - l_j(d^{(j)} - y_j)_+ - c_j y_j, \tag{1.40}$$

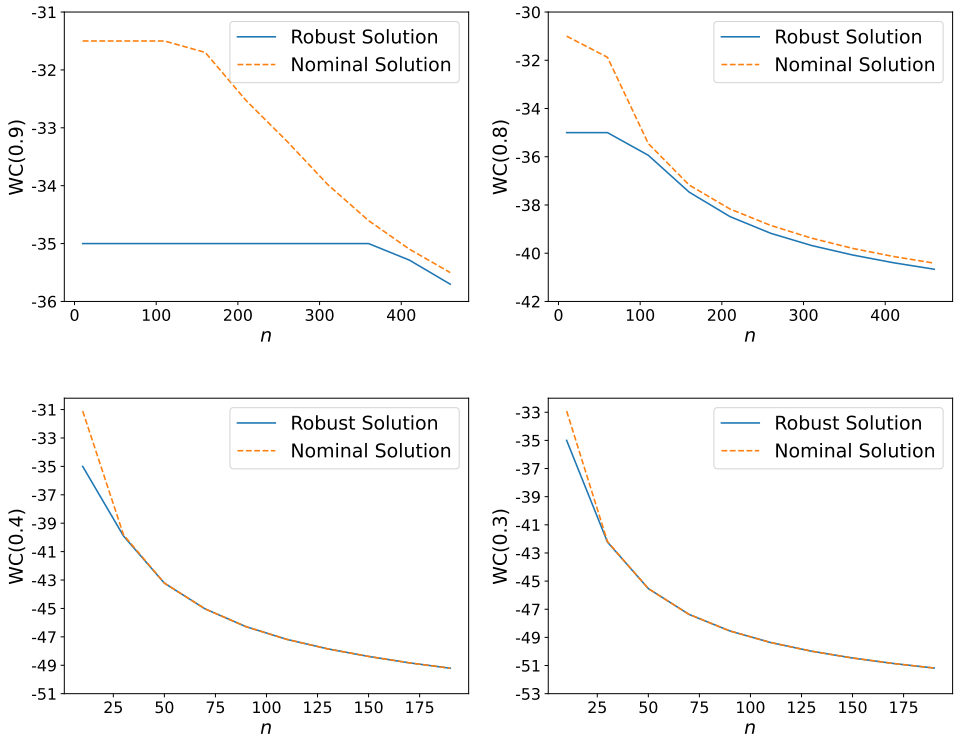
with v_j, s_j, l_j, c_j the parameters corresponding to item j. We assume that the demand takes on the same possible values for all items, i.e., $d_i^{(j)} \in \{4, 8, 10\}$ for all j. Since each realization of $d^{(j)}$ contributes to a possible realization of $\pi_{\text{tot}}(\mathbf{y}, \mathbf{d})$, there are in total $m = 3^3 = 27$ possible realizations. We solve again problems (P) and (P-Nom) with the same preference specification as in the single-item problem. Since the $\text{CVaR}_{1-\alpha_0}$ risk measure has a piecewise-linear distortion function, we can apply Theorem 1.21 to solve (P-ref) and (P-Nom-ref) in this higher-dimensional setting. We take the nominal probability $\mathbf{p} \in \mathbb{R}^{27}$ to be the probability of each combination of realizations of $(d^{(1)}, d^{(2)}, d^{(3)})$, which is the product of the probabilities $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \mathbf{p}^{(3)}$ of each individual realization. The parameters we use are given in Table 1.1.

Similar to the single-item problem, we investigate the differences between the worst-

Table 1.1: Multi-item newsvendor problem. This table displays the parameters used in the multi-item newsvendor problem.

Item(j)	c	v	s	l	$p_1^{(j)}$	$p_2^{(j)}$	$p_3^{(j)}$
1	4	6	2	4	0.375	0.375	0.25
2	5	8	2.5	3	0.25	0.25	0.5
3	4	5	1.5	4	0.127	0.786	0.087

Figure 1.3: Multi-item newsvendor problem. This figure displays the worst-case evaluation $\mathrm{WC}(\alpha_0) \triangleq \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r(n))} \mathrm{CVaR}_{1-\alpha_0}(.)$ under the robust and nominal solutions, for a range of values of $r(n) = \chi^2_{2,0.95}/(2n)$ and $\alpha_0 = 0.9, 0.8, 0.4, 0.3$.



case evaluations of the robust and nominal solutions, for a range of n and α_0. We choose $\alpha_0 = 0.4, 0.3$ to compare the results with the single-item problem, and $\alpha_0 = 0.9, 0.8$ to explore higher values of α_0. From Figure 1.3, we observe that in the multi-item problem, the worst-case evaluation is much lower (i.e., better) than in the single-item problem, which speaks to the diversification benefits of a multi-item inventory. The overall pattern is similar to the single-item case: for relatively low α_0, the difference in worst-case evaluation between the robust solution and the nominal solution is smaller than for larger α_0. The same holds as the sample size n increases.

1.7.3 Robust Portfolio Choice with Concave Distortion Function

In this subsection, we investigate the performance of the cutting-plane algorithm and the piecewise-linear approximation method described in Algorithms 1 and 2, by studying robust portfolio optimization problems. We consider the returns of six portfolios formed on size and book-to-market ratio (2×3) obtained from Kenneth French's data library.[10] We use monthly returns from January 1984 to January 2014. This gives us a total of $m = 360$ return realizations for each of the six portfolios. We denote by $\mathbf{r} \in \mathbb{R}^6$ the uncertain return of the six portfolios, with empirical realizations $\mathbf{r}_1, \ldots, \mathbf{r}_{360}$. We solve the following nominal and robust portfolio optimization problems:

$$\min_{\mathbf{a}\in\mathcal{A}} \rho_{u,h,\mathbf{p}}(1 + \mathbf{a}^T\mathbf{r}) \tag{1.41}$$

$$\min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(1 + \mathbf{a}^T\mathbf{r}), \tag{1.42}$$

where $1 + \mathbf{a}^T\mathbf{r}$ is the total wealth after one period with initial wealth normalized to unity, which is evaluated under the rank-dependent utility model. The decision variable $\mathbf{a} \in \mathbb{R}^6$ is subject to the constraints $\mathcal{A} = \{\mathbf{a} \in \mathbb{R}^6 \mid \sum_{j=1}^6 a_j = 1, a_j \geq 0\}$. The nominal probability $\mathbf{p}$ is set to be the empirical distribution with $p_i = \frac{1}{360}$. We use the modified chi-squared divergence function $\phi(x) = (x-1)^2$, with radius $r = \frac{1}{n}\chi^2_{359,0.95}$. We choose the concave distortion function $h(p) = 1 - (1-p)^2$, motivated by two decision-theoretic papers: Eeckhoudt et al. [2020] and Eeckhoudt and Laeven [2021]. Furthermore, we choose the exponential utility function $u(x) = 1 - e^{-x/\lambda}$ with $\lambda = 10$.

We solve problems (1.41)–(1.42) with the cutting-plane method described in Algorithm 1, and compare its performance to the piecewise-linear method described in Algorithm 2, formally relying on Theorems 1.14 and 1.23. As shown in Table 1.2, both algorithms yield a very small gap between the upper and lower bounds they provide. The piecewise-linear approximation method is typically more efficient, but this is not surprising, since the cutting-plane method is a more general method and utilizes less structure of the problems at hand.

[10] `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

Table 1.2: Robust portfolio choice. This table displays the results of Algorithms 1 and 2 when applied to the robust problem (1.42) and the nominal problem (1.41). The respective tolerance parameters are set to $\epsilon_{\text{tol}} = 0.0001$ and $\epsilon = 0.001$.

Panel A: The cutting-plane method

Problem	Lower Bound	Upper Bound	# Cuts	Run Time
Robust	-0.08953	-0.08948	5	54 sec
Nominal	-0.09403	-0.09397	5	22 sec

Panel B: Piecewise-linear approximation

Problem	Lower Bound	Upper Bound	Run Time
Robust	-0.08951	-0.08948	14 sec
Nominal	-0.09401	-0.09398	7 sec

Additionally, for the cutting-plane algorithm, we calculate the worst-case rank-dependent evaluation of the nominal solution, which is equal to -0.08938. We also calculate the evaluation of the robust solution under the nominal distribution $\mathbf{p}$, which is equal to -0.09395. As we can see from the table, the worst-case evaluation of the nominal solution is not much larger than that of the robust solution (cf. the first rows of the Panels A and B), suggesting that the nominal solution is already "near-optimal" for the robust problem (1.42). Similarly, the robust solution is also "near-optimal" for the nominal problem (1.41). It appears that when the dimension m of the state space is large, the robust and nominal problems do not yield highly different solutions in this example.

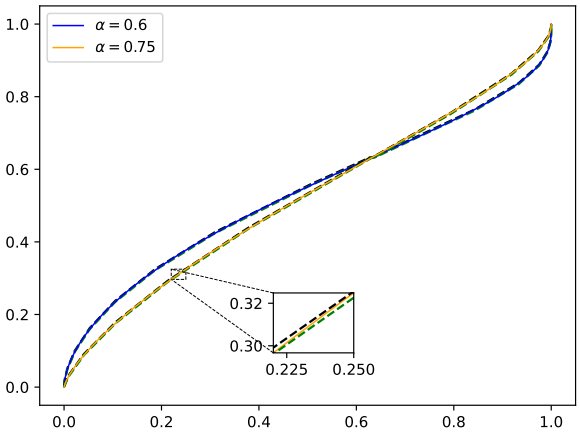
1.7.4 Robust Portfolio Choice with Inverse S-shaped Distortion Function

In this subsection, we investigate portfolio optimization problems (1.41)–(1.42) when h is an inverse S-shaped distortion function. In particular, we examine the Prelec [1998] distortion function:

$$h_\alpha(p) \triangleq 1 - \exp(-(-\log(1-p))^\alpha), \quad 0 < \alpha < 1. \tag{1.43}$$

We focus primarily on $\alpha = 0.6,\ 0.75$, which are broadly consistent with common empirical findings (Wakker, 2010, p. 260). We consider a linear utility function, $u(x) = x$, to isolate the effect of inverse S-shaped probability weighting. To obtain an upper and lower bound on problems (1.41)–(1.42), we approximate Prelec's function from above and below, using piecewise-linear functions. The approximation procedure is carried out by applying the method described in Section 1.5.2 separately to the concave and convex parts of Prelec's function, where we minimize the number of linear pieces under a pre-specified

Figure 1.4: Prelec's distortion. This figure displays Prelec's distortion function (1.43) and its upper and lower piecewise-linear approximations (dashed) for $\alpha = 0.6, 0.75$. The approximation error is set to $\epsilon = 0.($



approximation error.[11] The shape of Prelec's function and its respective upper and lower piecewise-linear approximations are displayed in Figure 1.4.

We generate 100 return data for 5 assets using the same numerical scheme as in Esfahani and Kuhn [2018], where the returns for each asset $j \in \{1, \ldots, 5\}$ are composed of two factors: $r_j = \psi + \gamma_j$, with a systematic risk factor $\psi \sim N(0, 0.02)$ and an idiosyncratic risk factor $\gamma_j \sim N(0.03j, 0.25j)$, for the j-th asset. Here, $N(\mu, \sigma)$ denotes the Gaussian distribution. By construction, the asset with a higher index j has a higher expected return and standard deviation.

We obtain lower and upper bounds on the nominal problem (1.41) by implementing the constraints (1.34) of Theorem 1.27 for the lower and upper piecewise-linear approximation of h, respectively, in the Gurobi solver (Version 11.0.3). Next, we invoke Theorems 1.28–1.29. Hence, a lower bound for the robust problem (1.42) is obtained by calculating the lower bound implied by the cutting-plane method described in Algorithm 3 (where we set $\epsilon_{\text{tol}} = 0.001$), while using the lower approximation of h. Then, with the solution obtained from the cutting-plane method, we determine an upper bound on (1.42) by calculating its worst-case evaluation using the SOS2-constraints as in (1.36), where we use the upper approximation of h. For the robust problem, we consider the total variation divergence $\phi(t) = |t - 1|$. We choose the radius for the robust problem to be $r = \frac{1}{m}\chi^2_{m-1,0.95}$, where $m = 100$.

The results are displayed in Table 1.3. We observe very tight upper and lower bounds. Furthermore, Gurobi solves both the nominal and robust problems (1.41)–(1.42) efficiently. As we can observe from Panel A, the evaluation of the nominal solution decreases with α. This is not surprising. Indeed, as we can see in Figure 1.4, Prelec's function has

[11] The approximation error is chosen to be $\epsilon = 0.003$ leading to 19, 13, and 6 linear pieces for $\alpha = 0.6, 0.75, 0.95$.

Table 1.3: Robust portfolio choice with inverse S-shaped distortion function. Upper and lower bounds (UB and LB) obtained for the nominal problem (1.41) and the robust problem (1.42) when h is Prelec's distortion function with $\alpha \in \{0.6, 0.75, 0.95\}$ (see Figure 1.4) and the divergence function is given by $\phi(t) = |t - 1|$. The cardinality of return realizations is $m = 100$.

Panel A: Solutions Nominal Problem

α	LB	UB	Run Time (LB)	Run Time (UB)
0.6	-1.144	-1.142	1.34 sec	1.21 sec
0.75	-1.153	-1.152	0.44 sec	0.90 sec
0.95	-1.162	-1.160	0.43 sec	0.30 sec

Panel B: Solutions Robust Problem

α	LB	UB	# Cuts	Run Time (LB)	Run Time (UB)
0.6	-1.041	-1.040	9	314 sec	2.77 sec
0.75	-1.039	-1.038	9	142 sec	1.45 sec
0.95	-1.036	-1.035	9	98 sec	0.37 sec

the property that if $\alpha_0 < \alpha_1$, then $h_{\alpha_0}(p) > h_{\alpha_1}(p)$ for the concave segment $p \in [0, 1-1/e]$, which covers more than half of the interval $[0, 1]$. Interestingly, Panel B of Table 1.3 reveals that the robust evaluation of the robust solution can exhibit an opposite, mildly increasing relationship with respect to α. This is because in the robust case, if the radius r is sufficiently large, then the worst-case probability $q^*_{(m)}$ of the worst-case realization can be such that $q^*_{(m)} \geq 1 - 1/e$ (recall that (m) denotes the index of the highest ranked realization, as in (1.35)). When this happens, all distorted decumulative probabilities in the objective function of (1.35) are increasing in α, since then all decumulative probabilities are on the interval $[1 - 1/e, 1]$. If the radius r would be set sufficiently small, then the same decreasing pattern with respect to α as in Panel A would also emerge in the robust case.

1.8 Concluding Remarks

In this paper, we have shown that nominal and robust optimization problems involving rank-dependent models can be reformulated into rank-independent, tractable optimization problems. When the distortion function is concave, we have demonstrated that this reformulation admits a conic representation, which we have explicitly derived for canonical distortion and divergence functions. Whereas the multiplicity of constraints in the reformulation increases exponentially with the dimension of the underlying probability space, we have developed two types of algorithms, and combinations thereof, to circumvent this curse of dimensionality. We have formally established that the upper and lower bounds the algorithms generate converge to the optimal objective value. Finally, we have illustrated the good performance of our methods in two examples involving concave as well as inverse S-shaped distortion functions, efficiently yielding nominal and robust solutions that generate very tight upper and lower bounds on the optimal objective values.

As a direction for future research, one can investigate whether the approach developed in this paper can be extended to encompass robust optimization problems with general law-invariant convex functionals (Föllmer and Schied [2016], S. 4.5), which also admit a dual representation.

Appendix

1.A Tables 1.4 and 1.5

Distortion family	*Function* $h(p)$, $p \in [0,1]$	*Conjugate* $(-h)^*(y)$, $y \le 0$	*Epigraph of perspective* $\lambda(-h)^*\left(\frac{-\nu}{\lambda}\right) \le z,\ \lambda > 0$	*Conic representation*
Expectation	p	$0,\ y \le -1$	$0 \le z,\ \nu \ge \lambda$	CQ
$\mathrm{CVaR}_{1-\alpha}$	$\min\{\frac{p}{1-\alpha}, 1\},\ \alpha \in [0,1)$	$\max\{(1-\alpha)y + 1, 0\}$	$\max\{-(1-\alpha)\nu + \lambda, 0\} \le z,\ \nu \ge 0$	CQ
Proportional Hazard	$\begin{cases} p^r \\ r \in (0,1) \end{cases}$	$r^{\frac{1}{1-r}}(1+\frac{1}{r})\lvert y\rvert^{\frac{r}{r-1}}$	$\begin{cases} \lambda(r^{\frac{r}{1-r}} - r^{\frac{1}{1-r}})^{1-r} \le z^{1-r}\nu^r \\ \nu \ge 0 \end{cases}$	PC
Absolute Deviation	$\begin{cases} (1+r)p & p < 1/2 \\ (1-r)p + r & p \ge 1/2 \\ r \in (0,1) \end{cases}$	$\max\{(y+1+r)/2, 0\},\ y \le -(1-r)$	$\begin{cases} \max\{(-\nu + \lambda + \lambda r)/2, 0\} \le z \\ -\nu \le -(1-r)\lambda \end{cases}$	CQ
Gini Principles	$\begin{cases} (1+r)p - rp^2 \\ r \in (0,1) \end{cases}$	$\begin{cases} \frac{1}{4r}\max\{y+1+r, 0\}^2 & y+1 \le r \\ y+1 & y+1 > r \end{cases}$	$\begin{cases} z = z_1 + z_2 \\ -\nu + \lambda(1-r) = \xi_1 + \xi_2 \\ (z_1 + \lambda) \ge \sqrt{\frac{w^2}{r} + (z_1 - \lambda)^2},\ z_2 \ge \xi_2 \\ \xi_1 + 2r \le w \\ \xi_1 \le 0, \xi_2, w, \nu \ge 0 \end{cases}$	CQ
Dual Moments	$\begin{cases} 1 - (1-p)^n \\ n > 1 \end{cases}$	$\begin{cases} y + c(n)\cdot\min\{\lvert y\rvert, n\}^{\frac{n}{n-1}} + 1 \\ c(n) = \left(n^{-\frac{1}{n-1}} - n^{-\frac{n}{n-1}}\right) \end{cases}$	$\begin{cases} \lambda - \xi_3 + (n^{-\frac{1}{n-1}} - n^{-\frac{n}{n-1}})\xi_4 \le z \\ \xi_2 \le \xi_4^{\frac{n-1}{n}} \cdot \lambda^{1-\frac{n-1}{n}} \\ -\nu + \xi_3 \le 0,\ \xi_2 \ge \xi_3, \\ \xi_2, \xi_3, \xi_4 \ge 0. \end{cases}$	PC
MAXMINVAR	$\begin{cases} (1-(1-p)^n)^{1/n} \\ n > 1 \end{cases}$	$\begin{cases} y\left(1 - s^{\frac{1}{n}}\right) + (1-s)^{\frac{1}{n}} \\ s = \frac{\lvert y\rvert^{\frac{n}{n-1}}}{1+\lvert y\rvert^{\frac{n}{n-1}}} \end{cases}$	$\begin{cases} -\nu + \xi_3 \le 0, \xi_2 \ge \lambda \\ -\xi_3 + \lVert(\xi_2, \xi_3)\rVert_{\frac{n}{n-1}} \le z \\ \xi_2, \xi_3 \ge 0 \end{cases}$	PC
Lookback Transform	$\begin{cases} p^r(1 - \log(p^r)) \\ r \in (0,1) \end{cases}$	Closed form unknown	$\begin{cases} \xi_3 + \lambda e^{-(\xi_2+\xi_3)/\lambda} + (r^{\frac{r}{1-r}} - r^{\frac{1}{1-r}})\xi_4 \le z \\ \xi_2 \le \nu^r \xi_4^{1-r} \\ \xi_2, \xi_3, \xi_4, \nu \ge 0 \end{cases}$	EXP × PC

Table 1.4: Canonical distortion functions with explicit expressions of their conjugates $(-h)^*$ and conic representations of the epigraphs of their perspectives. The cones are abbreviated as: CQ: quadratic cone, PC: power cone, EXP: exponential cone. Note that $(-h)^*(y) = +\infty$ for $y > 0$; see Remark 1.8. For Expectation and $\mathrm{CVaR}_{1-\alpha}$, see Föllmer and Schied [2016], Section 4.6; for Proportional Hazard, see Wang [1995]; for Absolute Deviation and Gini Principles, see Denneberg [1990b]; for Dual Moments, see Muliere and Scarsini [1989] and Eeckhoudt et al. [2020]; for MAXMINVAR, see Cherny and Madan [2009]; for Lookback Transform, see Denneberg [1990a].

Divergence family	*Function* $\phi(x),\ x \geq 0$	*Conjugate* $\phi^*(y),\ y \in \mathbb{R}$	*Epigraph of perspective* $\gamma\phi^*\left(\frac{s}{\gamma}\right) \leq t,\ \gamma > 0$	*Conic representation*
Kullback-Leibler	$x \log x - x + 1$	$e^y - 1$	$\begin{cases} \gamma \log(\frac{\gamma}{w}) + s \leq 0 \\ w - \gamma \leq t. \end{cases}$	EXP
Burg entropy	$-\log x + x - 1$	$-\log(1-y),\ y < 1$	$\begin{cases} \gamma \log(\frac{\gamma}{v}) \leq t \\ v = \gamma - s,\ v > 0 \end{cases}$	EXP
χ^2-distance	$\frac{1}{x}(x-1)^2$	$2 - 2\sqrt{1-y},\ y < 1$	$\begin{cases} 2\gamma - 2w \leq t \\ \sqrt{w^2 + \frac{1}{4}(\gamma - v)^2} \leq \frac{1}{2}(\gamma + v) \\ w \geq 0,\ v = \gamma - s,\ v > 0 \end{cases}$	CQ
Variation distance	$\lvert x - 1 \rvert$	$\max\{s, -1\},\ s \leq 1$	$\max\{s, -\gamma\} \leq t,\ s \leq \gamma$	CQ
Modified χ^2-distance	$(x-1)^2$	$\max\{0, y/2 + 1\}^2 - 1$	$\begin{cases} \sqrt{w^2 + \frac{t^2}{4}} \leq \frac{t+2\gamma}{2} \\ 0 \leq w, s/2 + \gamma \leq w \end{cases}$	CQ
Hellinger distance	$(\sqrt{x} - 1)^2$	$\frac{y}{1-y},\ y < 1$	$\begin{cases} -\gamma + v \leq t \\ \sqrt{\gamma^2 + \frac{1}{4}(v - w)^2} \leq \frac{1}{2}(v + w) \\ w = \gamma - s,\ w > 0. \end{cases}$	CQ
χ-divergence of order $\theta > 1$	$\lvert x - 1 \rvert^\theta$	$y + (\theta - 1)(\frac{\lvert y \rvert}{\theta})^{\frac{\theta}{\theta - 1}}$	$\begin{cases} s + (\theta - 1)\theta^{\frac{\theta}{1-\theta}} w \leq t, 0 \leq w \\ \lvert s \rvert \leq w^{\frac{\theta}{\theta-1}} \cdot \gamma^{1 - \frac{\theta}{\theta-1}} \end{cases}$	PC
Cressie and Read	$\frac{1-\theta+\theta x - x^\theta}{\theta(1-\theta)},\ 0 < \theta < 1$	$\begin{cases} \frac{1}{\theta}(1 - y(1-\theta))^{\frac{\theta}{\theta-1}} - \frac{1}{\theta} \\ y > \frac{1}{1-\theta} \end{cases}$	$\begin{cases} \lvert w \rvert \leq (t\theta + \gamma)^{\frac{\theta-1}{\theta}} \cdot \gamma^{1 - \frac{\theta-1}{\theta}} \\ w = \gamma - s(1-\theta),\ s < \frac{\gamma}{1-\theta} \end{cases}$	PC

Table 1.5: Canonical ϕ-divergence functions taken from Table 2 of Ben-Tal et al. [2013], but with the explicit epigraphs of their perspective functions (provided in the fourth column) and corresponding conic representations (provided in the fifth column). The cones are abbreviated as: CQ: quadratic cone, PC: power cone, EXP: exponential cone.

1.B Proofs

Proof of Theorem 1.6. Let $\mathbf{q}$ be an arbitrary probability vector and denote by $\mathbb{Q}$ the corresponding probability measure on the sigma-algebra $2^{|\Omega|}$, i.e., the set of all subsets of Ω. Recall Eqn. (1.3). By Denneberg [1994], pp. 16–17, the set function induced by the composition $h \circ \mathbb{Q}$ is monotone and submodular. It then follows from Proposition 10.3 of Denneberg [1994] that, for any $X : \Omega \to \mathbb{R}$,

$$\rho_{u,h,\mathbf{q}}(X) = \sup_{\bar{\mathbf{q}} \in M_h(\mathbf{q})} \mathbb{E}_{\bar{\mathbf{q}}}[-u(X)],$$

with $M_h(\mathbf{q})$ defined in (1.15).

Hence, we have that

$$\begin{aligned}\sup_{\mathbf{q} \in \mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) &= \sup_{\mathbf{q} \in \mathcal{D}_\phi(\mathbf{p},r)} \sup_{\bar{\mathbf{q}} \in M_h(\mathbf{q})} \mathbb{E}_{\bar{\mathbf{q}}}[-u(f(\mathbf{a},\mathbf{X}))] \\ &= \sup_{(\mathbf{q},\bar{\mathbf{q}}) \in \mathcal{U}_{\phi,h}(\mathbf{p})} -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)),\end{aligned}$$

where $\mathcal{U}_{\phi,h}(\mathbf{p})$ is as defined in (1.13). This proves the stated result. □

Proof of Theorem 1.7. Suppose that a pair $(\mathbf{a}, c) \in \mathbb{R}^{n_a+1}$ satisfies the inequality

$$\sup_{(\mathbf{q},\bar{\mathbf{q}}) \in \mathcal{U}_{\phi,h}(\mathbf{p})} -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)) \leq c. \tag{1.44}$$

Then, the left-hand side of (1.44) constitutes a maximization problem upper bounded by $c < \infty$. Moreover, the set $\mathcal{U}_{\phi,h}(\mathbf{p})$ contains $(\mathbf{p},\mathbf{p})$ as a Slater point, since $\phi(1) = 0 < r$ and $h(\sum_{k \in I_j} p_k) > \sum_{k \in I_j} p_k$ for all subsets $I_j \subset [m]$ that are not $\emptyset$ and $[m]$.[12] Therefore,

[12]This is because $h(x) > x$ for all $x \in (0,1)$ due to concavity (excluding the trivial case where $h(x) = x$ for all $x \in [0,1]$).

strong duality holds, and we consider the Lagrangian function

$$
\begin{aligned}
&L(\mathbf{a}, \mathbf{q}, \bar{\mathbf{q}}, \alpha, \beta, (\lambda_j)_j, \gamma) \\
&= -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)) - \alpha \left(\sum_{i=1}^{m} q_i - 1 \right) - \beta \left(\sum_{i=1}^{m} \bar{q}_i - 1 \right) - \sum_{j=1}^{2^m-2} \lambda_j \left(\sum_{k \in I_j} \bar{q}_k - h \left(\sum_{k \in I_j} q_k \right) \right) \\
&\quad - \gamma \left(\sum_{i=1}^{m} p_i \phi \left(\frac{q_i}{p_i} \right) - r \right) \\
&= \alpha + \beta + \gamma r + \sum_{i=1}^{m} -(u(f(\mathbf{a}, \mathbf{x}_i)) + \beta) \bar{q}_i - \sum_{j=1}^{2^m-2} \lambda_j \sum_{k \in I_j} \bar{q}_k + \sum_{i=1}^{m} -\alpha q_i - \gamma \left(\sum_{i=1}^{m} p_i \phi \left(\frac{q_i}{p_i} \right) \right) \\
&\quad - \sum_{j=1}^{2^m-2} -\lambda_j h \left(\sum_{k \in I_j} q_k \right),
\end{aligned}
$$

for $\alpha, \beta \in \mathbb{R}$ and $\lambda_j, \gamma \geq 0$. We analyze $\sup_{\mathbf{q}, \bar{\mathbf{q}} \geq 0} L(\mathbf{a}, \mathbf{q}, \bar{\mathbf{q}}, \alpha, \beta, (\lambda_j)_j, \gamma)$, which excluding the constant $\alpha + \beta + \gamma r$ is equal to:

$$
\begin{aligned}
&\sup_{\mathbf{q}, \bar{\mathbf{q}} \geq 0} \sum_{i=1}^{m} -(u(f(\mathbf{a}, \mathbf{x}_i)) + \beta) \bar{q}_i - \sum_{j=1}^{2^m-2} \lambda_j \sum_{k \in I_j} \bar{q}_k + \sum_{i=1}^{m} -\alpha q_i - \gamma p_i \phi \left(\frac{q_i}{p_i} \right) \\
&\qquad - \sum_{j=1}^{2^m-2} \lambda_j (-h) \left(\sum_{k \in I_j} q_k \right) \\
&= \sup_{\bar{\mathbf{q}} \geq 0} \sum_{i=1}^{m} -(u(f(\mathbf{a}, \mathbf{x}_i)) + \beta) \bar{q}_i - \sum_{j=1}^{2^m-2} \lambda_j \sum_{k \in I_j} \bar{q}_k + \sup_{\mathbf{q} \geq 0} \sum_{i=1}^{m} -\alpha q_i - \gamma p_i \phi \left(\frac{q_i}{p_i} \right) \\
&\qquad - \sum_{j=1}^{2^m-2} \lambda_j (-h) \left(\sum_{k \in I_j} q_k \right).
\end{aligned}
$$

We examine both supremum terms separately. The first supremum gives:

$$
\begin{aligned}
\sup_{\bar{\mathbf{q}} \geq 0} \sum_{i=1}^{m} -(u(f(\mathbf{a}, \mathbf{x}_i)) + \beta) \bar{q}_i - \sum_{j=1}^{2^m-2} \lambda_j \sum_{k \in I_j} \bar{q}_k &= \sup_{\bar{\mathbf{q}} \geq 0} \sum_{i=1}^{m} - \left(u(f(\mathbf{a}, \mathbf{x}_i)) + \beta + \sum_{j : i \in I_j} \lambda_j \right) \bar{q}_i \\
&= \begin{cases} 0 & \text{if } u(f(\mathbf{a}, \mathbf{x}_i)) + \beta + \sum_{j : i \in I_j} \lambda_j \geq 0, \ \forall i \\ \infty & \text{else.} \end{cases}
\end{aligned}
$$

The second supremum gives:

$$
\begin{aligned}
&\sup_{\mathbf{q}\geq 0} \sum_{i=1}^{m} -\alpha q_i - \gamma p_i \phi\left(\frac{q_i}{p_i}\right) - \sum_{j=1}^{2^m-2} \lambda_j(-h)\left(\sum_{k\in I_j} q_k\right) \\
&= \sup_{\substack{\mathbf{q}\geq \mathbf{0} \\ w_1,\ldots,w_{2^m-2}\geq 0}} \left\{ \sum_{i=1}^{m} -\alpha q_i - \gamma p_i \phi\left(\frac{q_i}{p_i}\right) - \sum_{j=1}^{2^m-2} \lambda_j(-h)(w_j) \;\middle|\; w_j = \sum_{k\in I_j} q_k \right\} \\
&= \inf_{\nu_1,\ldots,\nu_{2^m-2}} \sup_{\substack{\mathbf{q}\geq \mathbf{0} \\ w_1,\ldots,w_{2^m-2}\geq 0}} \sum_{i=1}^{m} -\alpha q_i - \gamma p_i \phi\left(\frac{q_i}{p_i}\right) - \sum_{j=1}^{2^m-2} \lambda_j(-h)(w_j) - \sum_{j=1}^{2^m-2} \nu_j \left(w_j - \sum_{k\in I_j} q_k\right) \\
&= \inf_{\nu_1,\ldots,\nu_{2^m-2}} \sup_{\mathbf{q}\geq 0} \sum_{i=1}^{m} \left(-\alpha + \sum_{j:i\in I_j} \nu_j\right) q_i - \gamma p_i \phi\left(\frac{q_i}{p_i}\right) + \sup_{w_1,\ldots,w_{2^m-2}\geq 0} \sum_{j=1}^{2^m-2} -\nu_j w_j - \lambda_j(-h)(w_j) \\
&= \inf_{\nu_1,\ldots,\nu_{2^m-2}} \sum_{i=1}^{m} p_i \left(\sup_{t\geq 0} \left(-\alpha + \sum_{j:i\in I_j} \nu_j\right) t - \gamma\phi(t)\right) + \sum_{j=1}^{2^m-2} \lambda_j(-h)^*\left(\frac{-\nu_j}{\lambda_j}\right) \\
&= \inf_{\nu_1,\ldots,\nu_{2^m-2}} \sum_{i=1}^{m} p_i \gamma \phi^*\left(\frac{-\alpha + \sum_{j:i\in I_j} \nu_j}{\gamma}\right) + \sum_{j=1}^{2^m-2} \lambda_j(-h)^*\left(\frac{-\nu_j}{\lambda_j}\right),
\end{aligned}
$$

with ϕ^* and $(-h)^*$ the conjugates of ϕ and $-h$.

Therefore, strong duality implies that a pair $(\mathbf{a}, c)$ satisfies (1.44) if and only if

$$
\begin{aligned}
&\inf_{\substack{\alpha,\beta,\nu_j\in\mathbb{R} \\ \lambda_j,\gamma\geq 0}} \alpha + \beta + \gamma r + \sum_{i=1}^{m} p_i \gamma \phi^*\left(\frac{-\alpha + \sum_{j:i\in I_j} \nu_j}{\gamma}\right) + \sum_{j=1}^{2^m-2} \lambda_j(-h)^*\left(\frac{-\nu_j}{\lambda_j}\right) \leq c \\
&\text{subject to } -u(f(\mathbf{a},\mathbf{x}_i)) - \beta - \sum_{j:i\in I_j} \lambda_j \leq 0,\ \forall i \in [m].
\end{aligned}
$$

Since the infimum is attained due to the boundedness of the primal problem (1.44) and the strong duality theorem, we may remove the infimum sign and obtain that the above holds if and only if there exist $\lambda_j, \gamma \geq 0, \alpha, \beta, \nu_j \in \mathbb{R}$ such that

$$
\begin{cases}
\alpha + \beta + \gamma r + \sum_{i=1}^{m} p_i \gamma \phi^*\left(\frac{-\alpha + \sum_{j:i\in I_j} \nu_j}{\gamma}\right) + \sum_{j=1}^{2^m-2} \lambda_j(-h)^*\left(\frac{-\nu_j}{\lambda_j}\right) \leq c \\
-u(f(\mathbf{a},\mathbf{x}_i)) - \beta - \sum_{j:i\in I_j} \lambda_j \leq 0,\ \forall i \in [m] \\
\lambda_j, \gamma \geq 0 \\
\alpha, \beta, \nu_j \in \mathbb{R},\ \forall j \in [2^m - 2].
\end{cases}
$$

For the nominal problem, the Lagrangian function is given by

$$\begin{aligned} L(\mathbf{a}, \bar{\mathbf{q}}, \beta, (\lambda_j)_j) &= -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)) - \beta \left(\sum_{i=1}^{m} \bar{q}_i - 1 \right) - \sum_{j=1}^{2^m-2} \lambda_j \left(\sum_{k \in I_j} \bar{q}_k - h\left(\sum_{k \in I_j} p_k \right) \right) \\ &= \beta + \sum_{j=1}^{2^m-2} \lambda_j h\left(\sum_{k \in I_j} p_k \right) + \sum_{j=1}^{m} \bar{q}_i \left(-u(f(\mathbf{a}, \mathbf{x}_i)) - \beta - \sum_{j: i \in I_j} \lambda_j \right). \end{aligned}$$

Hence, we have

$$\begin{aligned} &\sup_{\bar{\mathbf{q}} \in M_h(\mathbf{p})} \sum_{i=1}^{m} -\bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)) \\ &= \inf_{\beta \in \mathbb{R}, \lambda_j \geq 0} \sup_{\bar{\mathbf{q}} \geq \mathbf{0}} L(\mathbf{a}, \bar{\mathbf{q}}, \beta, (\lambda_j)_j) \\ &= \inf_{\beta \in \mathbb{R}, \lambda_j \geq 0} \left\{ \beta + \sum_{j=1}^{2^m-2} \lambda_j h\left(\sum_{k \in I_j} p_k \right) \;\middle|\; -u(f(\mathbf{a}, \mathbf{x}_i)) - \beta - \sum_{j: i \in I_j} \lambda_j \leq 0, \ \forall i \in [m] \right\}, \end{aligned}$$

where the above strong duality holds since we are solving a linear programming problem. Therefore, we have that $\sup_{\bar{\mathbf{q}} \in M_h(\mathbf{p})} \sum_{i=1}^{m} -\bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)) \leq c$ if and only if there exist $\beta \in \mathbb{R}, (\lambda_j)_j \geq 0$ such that

$$\begin{cases} \beta + \sum_{j=1}^{2^m-2} \lambda_j h\left(\sum_{k \in I_j} p_k \right) \leq c \\ -u(f(\mathbf{a}, \mathbf{x}_i)) - \beta - \sum_{j: i \in I_j} \lambda_j \leq 0, \ \forall i \in [m] \\ \beta \in \mathbb{R}, \lambda_j \geq 0, \ \forall j \in [2^m - 2]. \end{cases}$$

□

Proof of Lemma 1.10. The proof follows Ben-Tal and Nemirovski [2019], except that we replace the quadratic cone with a general cone. We have

$$\mathrm{Epi}(f^*) = \{(\mathbf{y}, s) : \mathbf{y}^T \mathbf{x} - f(\mathbf{x}) \leq s, \forall \mathbf{x}\} = \{(\mathbf{y}, s) : \mathbf{y}^T \mathbf{x} - t \leq s, \forall (\mathbf{x}, t) \in \mathrm{Epi}(f)\}.$$

Therefore, $(\mathbf{y}, s) \in \mathrm{Epi}(f^*)$ if and only if

$$\min_{\mathbf{x}, t, \mathbf{w}} \{-\mathbf{y}^T \mathbf{x} + t : \mathbf{A}\mathbf{x} + t\mathbf{v} + \mathbf{B}\mathbf{w} + \mathbf{b} \succeq_{\mathbf{K}} \mathbf{0}\} \geq -s.$$

This minimization problem is strictly feasible and bounded from below. Hence, the conic duality theorem (see Ben-Tal and Nemirovski, 2019) implies that it is equal to

$$\max_{\boldsymbol{\xi}} \{-\mathbf{b}^T \boldsymbol{\xi} : \mathbf{A}^T \boldsymbol{\xi} = -\mathbf{y}, \mathbf{B}^T \boldsymbol{\xi} = \mathbf{0}, \mathbf{v}^T \boldsymbol{\xi} = 1, \boldsymbol{\xi} \in \mathbf{K}^*\}.$$

Therefore,

$$\text{Epi}(f^*) = \{(\mathbf{y}, s) : \exists\ \boldsymbol{\xi} \in \mathbf{K}^* : \mathbf{A}^T\boldsymbol{\xi} = -\mathbf{y}, \mathbf{B}^T\boldsymbol{\xi} = \mathbf{0}, \mathbf{v}^T\boldsymbol{\xi} = 1, s \geq \mathbf{b}^T\boldsymbol{\xi}\}.$$

□

Proof of Lemma 1.11. Following Ben-Tal and Nemirovski [2019], we have

$$\begin{aligned}
\text{Epi}(\tilde{f}) = \left\{(\mathbf{x}, \lambda, t) : \lambda f\left(\frac{\mathbf{x}}{\lambda}\right) \leq t\right\} &= \left\{(\mathbf{x}, \lambda, t) : \left(\frac{\mathbf{x}}{\lambda}, \frac{t}{\lambda}\right) \in \text{Epi}(f)\right\} \\
&= \{(\mathbf{x}, \lambda, t) : \exists\ \mathbf{w} \in \mathbb{R}^k : \mathbf{A}(\mathbf{x}/\lambda, \mathbf{w}, t/\lambda)^T - b \succeq_{\mathbf{K}} \mathbf{0}\} \\
&= \{(\mathbf{x}, \lambda, t) : \exists\ \tilde{\mathbf{w}} \in \mathbb{R}^k : \mathbf{A}(\mathbf{x}/\lambda, \tilde{\mathbf{w}}/\lambda, t/\lambda)^T - b \succeq_{\mathbf{K}} \mathbf{0}\} \\
&= \{(\mathbf{x}, \lambda, t) : \exists\ \tilde{\mathbf{w}} \in \mathbb{R}^k : \mathbf{A}(\mathbf{x}, \tilde{\mathbf{w}}, t)^T - \lambda b \succeq_{\mathbf{K}} \mathbf{0}\} \\
&= \{(\mathbf{x}, \lambda, t) : \exists\ \tilde{\mathbf{w}} \in \mathbb{R}^k : [\mathbf{A}, -b](\mathbf{x}, \tilde{\mathbf{w}}, t, \lambda)^T \succeq_{\mathbf{K}} \mathbf{0}\}.
\end{aligned}$$

□

Proof of Lemma 1.13. By definition, $\mathbf{q}^* \in \mathcal{D}_\phi(\mathbf{p}, r)$. Hence, we only need to show that $\bar{\mathbf{q}}^* \in M_h(\mathbf{q}^*)$. Using Lemma 4.98 of Föllmer and Schied [2016], we have that $\rho_{u,h,\mathbf{q}^*}(X) \geq \mathbb{E}_{\bar{\mathbf{q}}^*}[-X]$ for all random variables X (where $\rho_{u,h,\mathbf{q}^*}(X)$ is as defined in (1.4) with measure $\mathbb{Q} = \mathbf{q}^*$). In particular, this holds for all $X = -\mathbb{1}_A$, for any measurable set $A \subset \Omega$. Hence, $(\mathbf{q}^*, \bar{\mathbf{q}}^*) \in \mathcal{U}_{\phi,h}(\mathbf{p})$. □

Proof of Theorem 1.14. First, we have, for any $\mathbf{a}^1, \mathbf{a}^2 \in \mathcal{A}$,

$$\begin{aligned}
\left|\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}^2, \mathbf{x}_i)) - \sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}^1, \mathbf{x}_i))\right| &\leq \|\bar{\mathbf{q}}\|_2 \|(u(f(\mathbf{a}^2, \mathbf{x}_i)) - u(f(\mathbf{a}^1, \mathbf{x}_i)))_{i=1}^m\|_2 \\
&\leq \|\bar{\mathbf{q}}\|_1 \|(u(f(\mathbf{a}^2, \mathbf{x}_i)) - u(f(\mathbf{a}^1, \mathbf{x}_i)))_{i=1}^m\|_2 \\
&= \|(u(f(\mathbf{a}^2, \mathbf{x}_i)) - u(f(\mathbf{a}^1, \mathbf{x}_i)))_{i=1}^m\|_2,
\end{aligned} \tag{1.45}$$

where the first inequality follows from Cauchy-Schwarz; and the second inequality follows from $\|\bar{\mathbf{q}}\|_2 \leq \|\bar{\mathbf{q}}\|_1$ for a probability vector $\bar{\mathbf{q}}$, since $\bar{q}_k^2 \leq \bar{q}_k$ for $|\bar{q}_k| \leq 1$.

Suppose now that the cutting-plane method has not terminated at the t-th iteration, i.e., the optimal solution and objective value $(\mathbf{a}_t, c_t)$ violate the ϵ_{tol}-feasibility condition at step 4 of Algorithm 1. Let $(\mathbf{q}_t^*, \bar{\mathbf{q}}_t^*)$ be the new worst-case scenarios that are added to $\mathcal{U}_t$ at step 5 of Algorithm 1. Then, by definition, we have

$$-\sum_{i=1}^m \bar{q}_{t,i}^* u(f(\mathbf{a}_t, \mathbf{x}_i)) - c_t > \epsilon_{\text{tol}}. \tag{1.46}$$

For any $s > t$, we also have that at the s-th iteration:

$$-\sum_{i=1}^{m} \bar{q}^*_{t,i} u(f(\mathbf{a}_s, \mathbf{x}_i)) - c_s \leq 0, \tag{1.47}$$

since this is part of the constraint at the s-th iteration. Let $\tilde{c}$ be the optimal objective value of (P-ref). By Assumption 1.2, we have that $-\infty < \tilde{c} < \infty$. Furthermore, we have $\tilde{c} \geq c_i$ for any iteration i since the cutting-plane algorithm always yields a lower bound on (P-ref). Hence, we may assume that for t and s sufficiently large, the lower bound improvement is upper bounded, i.e., $c_s - c_t \leq \frac{1}{2}\epsilon_{\text{tol}}$, since otherwise the cutting-plane will yield a lower bound that exceeds $\tilde{c}$, after finitely many iterations. Therefore, it follows from (1.46)–(1.47) that

$$\left| \sum_{i=1}^{m} \bar{q}^*_{t,i} u(f(\mathbf{a}_t, \mathbf{x}_i)) - \sum_{i=1}^{m} \bar{q}^*_{t,i} u(f(\mathbf{a}_s, \mathbf{x}_i)) \right| > \epsilon_{\text{tol}} - (c_s - c_t) \geq \frac{1}{2}\epsilon_{\text{tol}},$$

which implies that

$$\|(u(f(\mathbf{a}_s, \mathbf{x}_i)) - u(f(\mathbf{a}_t, \mathbf{x}_i)))_{i=1}^m\|_2 > \frac{1}{2}\epsilon_{\text{tol}}. \tag{1.48}$$

This shows that the minimum distance between any two outcomes of the cutting-plane method, when evaluated in utilities, is at least $\frac{1}{2}\epsilon_{\text{tol}}$.

The idea is now to show that if the cutting-plane method does not terminate, then there is a sequence of infinitely many cutting-plane solutions $\{\mathbf{a}_j\}_{j=1}^{\infty}$ for which the corresponding vector $(u(f(\mathbf{a}_j, \mathbf{x}_i)))_{i=1}^m$ remains in a bounded set. Since we know from above that for each of these solutions their utility values are a distance $\frac{1}{2}\epsilon_{\text{tol}}$ away from each other, we conclude that the cutting-plane method must terminate since a bounded set can not contain infinitely many disjoint balls with radius $\frac{1}{2}\epsilon_{\text{tol}}$, as argued in Mutapcic and Boyd [2009].

Therefore, we define

$$\mathcal{T} \triangleq \left\{ (u(f(\mathbf{a}, \mathbf{x}_i)))_{i=1}^m \;\middle|\; \mathbf{a} \in \mathcal{A},\ -\sum_{i=1}^{m} p_i u(f(\mathbf{a}, \mathbf{x}_i)) \leq \tilde{c} \right\}.$$

Then, for all iterations t, we have $(u(f(\mathbf{a}_t, \mathbf{x}_i)))_{i=1}^m \in \mathcal{T}$, since $\mathbf{p} \in \mathcal{U}_t$ for all $t \geq 1$, hence we have the inequality $-\sum_{i=1}^m p_i u(f(\mathbf{a}_t, \mathbf{x}_i)) \leq c_t \leq \tilde{c}$. We show that $\mathcal{T}$ is bounded in the Euclidean 2-norm $\|.\|_2$. Indeed, by assumption, $M \triangleq \sup_{\mathbf{a}\in\mathcal{A}, i\in[m]} u(f(\mathbf{a}, \mathbf{x}_i)) < \infty$. Hence, for all vectors $(u(f(\mathbf{a}, \mathbf{x}_i)))_{i=1}^m \in \mathcal{T}$, its individual entry is always bounded from above by M. It remains to show that all its entries are also bounded from below. Let $p_{\min} = \min_{i=1}^m p_i > 0$ (by Assumption 1.3) and $i_{\min}(\mathbf{a}) = \text{argmin}_i u(f(\mathbf{a}, \mathbf{x}_i)$. Then, we

have for all $(u(f(\mathbf{a}, \mathbf{x}_i)))_{i=1}^m \in \mathcal{T}$,

$$-p_{i_{\min(\mathbf{a})}} u(f(\mathbf{a}, \mathbf{x}_{i_{\min(\mathbf{a})}})) - \sum_{i \neq i_{\min(\mathbf{a})}} p_i u(f(\mathbf{a}, \mathbf{x}_i)) \leq \tilde{c},$$

which implies

$$\begin{aligned} -u(f(\mathbf{a}, \mathbf{x}_{i_{\min(\mathbf{a})}})) \leq & \frac{\tilde{c} + M}{p_{i_{\min(\mathbf{a})}}} \leq \frac{\tilde{c} + M}{p_{\min}} \\ & \Leftrightarrow \\ u(f(\mathbf{a}, \mathbf{x}_{i_{\min(\mathbf{a})}})) \geq & -\frac{\tilde{c} + M}{p_{\min}} > -\infty. \end{aligned}$$

Hence, $\mathcal{T}$ is bounded in the Euclidean 2-norm $\|.\|_2$. □

Proof of Theorem 1.16. Before we proceed with the proof, we recall the notion of a KKT-vector. For a generic convex optimization problem

$$\min_{x \in \mathcal{X}} \{f_0(x) \mid f_i(x) \leq 0, \forall i \in [L]\},$$

with $\mathcal{X}$ a suitable convex subset and $f_0, \ldots, f_L$ convex functions, the KKT-vector corresponding to the constraints $f_i(x) \leq 0, i \in [L]$ is a vector $\lambda \in \mathbb{R}^L$ such that

$$\inf_{x \in \mathcal{X}} \left\{ f_0(x) + \sum_{i=1}^{L} \lambda_i f_i(x) \right\} = \min_{x \in \mathcal{X}} \{f_0(x) \mid f_i(x) \leq 0, \forall i \in [L]\}.$$

The existence of the KKT-vector is guaranteed when Slater's condition and the boundedness of the optimization problem are satisfied (see Theorem 28.2, Rockafellar, 1970).

The proof now consists of two parts. In the first part, we show that Algorithm 4 terminates after finitely many iterations, for any $\epsilon_{\text{tol}} > 0$. The second part shows the convergence as $\epsilon_{\text{tol}} \to 0$.

First part: We reexamine the proof of Theorem 1.14. First, we have, for any $\mathbf{a}^1, \mathbf{a}^2 \in \mathcal{A}$, the following inequality:

$$\left| \sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}^2, \mathbf{x}_i)) - \sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}^1, \mathbf{x}_i)) \right| \leq \|(u(f(\mathbf{a}^2, \mathbf{x}_i)) - u(f(\mathbf{a}^1, \mathbf{x}_i)))_{i=1}^m\|_2.$$

Let t be an iteration where Algorithm 4 has not yet terminated and let $(\mathbf{q}_t^*, \bar{\mathbf{q}}_t^*)$ be the new worst-case scenarios that are added to $\mathcal{U}_t$. Then, we have $-\sum_{i=1}^m \bar{q}_{t,i}^* u(f(\mathbf{a}_t, \mathbf{x}_i)) - c > \epsilon_{\text{tol}}$. For any $s > t$, we also have that at the s-th iteration $-\sum_{i=1}^m \bar{q}_{t,i}^* u(f(\mathbf{a}_s, \mathbf{x}_i)) - c \leq 0$. Hence,

$$\left| \sum_{i=1}^m \bar{q}_{t,i}^* u(f(\mathbf{a}_t, \mathbf{x}_i)) - \sum_{i=1}^m \bar{q}_{t,i}^* u(f(\mathbf{a}_s, \mathbf{x}_i)) \right| > \epsilon_{\text{tol}},$$

which implies that

$$\|(u(f(\mathbf{a}_s, \mathbf{x}_i)) - u(f(\mathbf{a}_t, \mathbf{x}_i)))_{i=1}^m\|_2 > \epsilon_{\text{tol}}.$$

We define the set

$$\mathcal{T} \triangleq \left\{ (u(f(\mathbf{a}, \mathbf{x}_i)))_{i=1}^m \;\middle|\; \mathbf{a} \in \mathcal{A},\ -\sum_{i=1}^m p_i u(f(\mathbf{a}, \mathbf{x}_i)) \leq c \right\}.$$

Then, for all iterations t, we have $(u(f(\mathbf{a}_t, \mathbf{x}_i)))_{i=1}^m \in \mathcal{T}$, since $\mathbf{p} \in \mathcal{U}_t$ for all $t \geq 1$. It follows from the proof of Theorem 1.14 that $\mathcal{T}$ must be bounded in the Euclidean 2-norm $\|.\|_2$. Hence, Algorithm 4 must terminate after finitely many steps.

Second part: Let $g(\mathbf{a}_{\epsilon_{\text{tol}}})$ be the objective value of the solution $\mathbf{a}_{\epsilon_{\text{tol}}}$ obtained at the final iteration of Algorithm 4, for a tolerance parameter $\epsilon_{\text{tol}} > 0$. Let $P(0)$ be the optimal objective value of (P-constraint). Define, for any $\epsilon > 0$,

$$P(\epsilon) \triangleq \min_{\mathbf{a} \in \mathcal{A}} \left\{ g(\mathbf{a}) \;\middle|\; \sup_{(\mathbf{q}, \bar{\mathbf{q}}) \in \mathcal{U}_{\phi,h}(\mathbf{p})} -\sum_{k=1}^N \bar{q}_k u(x_k(\mathbf{a})) \leq c + \epsilon \right\}.$$

By construction, we have $P(\epsilon_{\text{tol}}) \leq g(\mathbf{a}_{\epsilon_{\text{tol}}})$ since $\mathbf{a}_{\epsilon_{\text{tol}}}$ is feasible for the problem of $P(\epsilon_{\text{tol}})$. Furthermore, the cutting-plane algorithm yields a lower bound on $P(0)$. Hence, $g(\mathbf{a}_{\epsilon_{\text{tol}}}) \leq P(0)$ for all $\epsilon_{\text{tol}} > 0$. Therefore,

$$0 \leq P(0) - g(\mathbf{a}_{\epsilon_{\text{tol}}}) \leq P(0) - P(\epsilon_{\text{tol}}).$$

Thus, it remains to bound $P(0) - P(\epsilon_{\text{tol}})$ from above. This can be done by utilizing the strong duality theorem, which is guaranteed by Assumption 1.5. Therefore, we have, for any $\epsilon > 0$,

$$\begin{aligned} P(\epsilon) &= \sup_{\lambda \geq 0} \inf_{\mathbf{a} \in \mathcal{A}} g(\mathbf{a}) + \lambda \left(\sup_{(\mathbf{q}, \bar{\mathbf{q}}) \in \mathcal{U}_{\phi,h}(\mathbf{p})} -\sum_{k=1}^N \bar{q}_k u(x_k(\mathbf{a})) - c - \epsilon \right) \\ &\geq -\lambda^* \epsilon + \inf_{\mathbf{a} \in \mathcal{A}} g(\mathbf{a}) - \lambda^* \left(\sup_{(\mathbf{q}, \bar{\mathbf{q}}) \in \mathcal{U}_{\phi,h}(\mathbf{p})} -\sum_{k=1}^N \bar{q}_k u(x_k(\mathbf{a})) - c \right) \\ &= -\lambda^* \epsilon + P(0), \end{aligned}$$

where λ^* is the KKT-vector of the problem $P(0)$ corresponding to the supremum constraint, which is a strictly positive constant (by Assumption 1.5) independent of ϵ. Hence, we have

$$P(0) - P(\epsilon_{\text{tol}}) \leq \lambda^* \epsilon_{\text{tol}}.$$

Therefore, the convergence follows as $\epsilon_{\text{tol}} \to 0$. □

Proof of Lemma 1.18. We first show that, for any $\mathbf{a} \in \mathcal{A}$, we have that

$$\sup_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_{\phi,h}(\mathbf{p})} \sum_{i=1}^{m} -\bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)) = \sup_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_{\phi,h}^{(i_1,\ldots,i_m)}(\mathbf{p})} -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)), \tag{1.49}$$

for any ranking $(i_1,\ldots,i_m) \in \mathcal{I}(\mathbf{a})$ as in Definition 1.17. We fix $\mathbf{a}$, denote $Y \triangleq f(\mathbf{a},\mathbf{X})$ and consider the ranking $u(y_{(1)}) \geq \ldots \geq u(y_{(m)})$. By the definition of $\rho_{u,h,\mathbf{q}}(Y)$ in (1.4), we have that in our discrete setting $\rho_{u,h,\mathbf{q}}(Y)$ is equal to the rank-dependent sum

$$\rho_{u,h,\mathbf{q}}(Y) = -\sum_{i=1}^{m} h\left(\sum_{j=i}^{m} q_{(j)}\right)(u(y_{(i)}) - u(y_{(i-1)})),$$

where $-u(y_{(0)}) \triangleq 0$. We now claim that we have an equality between the rank-dependent sum above and the following optimization problem:

$$\begin{aligned}
\max_{\bar{\mathbf{q}}\in\mathbb{R}^m} \quad & \sum_{i=1}^{m} -\left(\sum_{j=i}^{m} \bar{q}_{(j)}\right)(u(y_{(i)}) - u(y_{(i-1)})) \\
\text{subject to} \quad & \sum_{j=i}^{m} \bar{q}_{(j)} \leq h\left(\sum_{j=i}^{m} q_{(j)}\right), \quad \forall i \in [m] \\
& \sum_{i=1}^{m} \bar{q}_i = 1 \\
& \bar{q}_i \geq 0, \quad \forall i \in [m].
\end{aligned}$$

Indeed, we can define $\bar{q}^*_{(i)} \triangleq h\left(\sum_{j=i}^{m} q_{(j)}\right) - h\left(\sum_{j=i+1}^{m} q_{(j)}\right)$, where $\bar{q}^*_{(m)} \triangleq h(q_{(m)})$. Then, $\sum_{j=i}^{m} \bar{q}^*_{(i)} = h\left(\sum_{j=i}^{m} q_{(j)}\right)$, and $\bar{\mathbf{q}}^*$ is a probability vector since $h(1) = 1$, $h(0) = 0$, and h is non-decreasing. Hence, $\bar{\mathbf{q}}^*$ is feasible for the above optimization problem. Furthermore, since $-(u(y_{(i)}) - u(y_{(i-1)})) \geq 0$ for all $i \geq 2$, the maximum is attained at a vector $\bar{\mathbf{q}}$ such that the constraint $\sum_{j=i}^{m} \bar{q}_{(i)} \leq h\left(\sum_{j=i}^{m} q_{(j)}\right)$ is an equality for all i, which uniquely defines $\bar{\mathbf{q}}^*$. An expansion of the alternating sum $\sum_{i=1}^{m} -\left(\sum_{j=i}^{m} q_{(j)}\right)(u(y_{(i)}) - u(y_{(i-1)}))$ shows that it is also equal to the sum $\sum_{i=1}^{m} -q_{(i)}u(y_{(i)})$. Therefore, (1.49) holds for all $\mathbf{a} \in \mathcal{A}$ and any ranking $(i_1,\ldots,i_m) \in \mathcal{I}(\mathbf{a})$. Let V^* be the optimal objective value of (P-constraint). Then, for any $\mathbf{a}_0$ such that $\mathcal{I}(\mathbf{a}_0) \subset \mathcal{I}(\mathbf{a}^*)$, we have that $U^*(\mathbf{a}_0) \leq V^*$, since $\mathbf{a}^*$ is feasible for the problem (1.21) by (1.49). On the other hand, we also have the upper bound relation $V^* \leq U^*(\mathbf{a}_0)$, since $\mathcal{U}_{\phi,h}(\mathbf{p}) \subset \mathcal{U}_{\phi,h}^{(i_1,\ldots,i_m)}(\mathbf{p})$ for any index vector $(i_1,\ldots,i_m)$. Hence, $U^*(\mathbf{a}_0) = V^*$. □

Proof of Theorem 1.19. Let $(\mathbf{q}^*_n, \bar{\mathbf{q}}^*_n) \in \operatorname{argmax}_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_{\phi,h}(\mathbf{p})} -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a}_n,\mathbf{x}_i))$. De-

note $g(\mathbf{a}, \mathbf{q}, \bar{\mathbf{q}}) = -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i))$. Since the set $\mathcal{A} \times \mathcal{U}_{\phi,h}(\mathbf{p})$ is compact,[13] we may assume that there exists a limit $(\mathbf{a}_n, \mathbf{q}_n^*, \bar{\mathbf{q}}_n^*) \to (\mathbf{a}_L, \mathbf{q}_L, \bar{\mathbf{q}}_L) \in \mathcal{A} \times \mathcal{U}_{\phi,h}(\mathbf{p})$, as $n \to \infty$. We will show that $\mathbf{a}_L$ must be an optimal solution for (P-constraint). Indeed, since $(\mathbf{q}_n^*, \bar{\mathbf{q}}_n^*)$ are maximizers, we have that

$$c + \epsilon_{\text{tol},n} \geq -\sum_{i=1}^{m} \bar{q}_{n,i}^* u(f(\mathbf{a}_n, \mathbf{x}_i)) \geq \max_{(\mathbf{q},\bar{\mathbf{q}}) \in \mathcal{U}_{\phi,h}(\mathbf{p})} -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a}_n, \mathbf{x}_i)).$$

Taking the limit as $n \to \infty$ yields

$$c \geq -\sum_{i=1}^{m} \bar{q}_{L,i}^* u(f(\mathbf{a}_L, \mathbf{x}_i)) \geq \max_{(\mathbf{q},\bar{\mathbf{q}}) \in \mathcal{U}_{\phi,h}(\mathbf{p})} -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a}_L, \mathbf{x}_i)).$$

Hence, by Theorem 1.6, this implies that $\mathbf{a}_L$ is feasible for (P-constraint). Since $\mathbf{a}_n$ is a sequence of solutions such that $g(\mathbf{a}_n)$ converges to the optimal objective value of (P-constraint) by Theorem 1.16, it follows from the continuity of g that $g(\mathbf{a}_L)$ equals to the optimal objective value of (P-constraint). Hence, $\mathbf{a}_L$ is an optimal solution of (P-constraint). Continuity of the functions $-u(f(\mathbf{a}, \mathbf{x}_i))$ in $\mathbf{a}$, for all $i \in [m]$, implies that there exists some $N > 0$, such that for all $n \geq N$, we have the inclusion of the ranking set: $\mathcal{I}(\mathbf{a}_n) \subset \mathcal{I}(\mathbf{a}_L)$. Therefore, Lemma 1.18 implies that $U^*(\mathbf{a}_n)$ converges to the optimal objective value of (P-constraint). □

Proof of Lemma 1.20. If $(\mathbf{q}, \bar{\mathbf{q}}, \mathbf{t})$ satisfies the constraints in (1.24), then $\bar{q}_i \leq l_j \cdot q_i + t_{i,j}$ for all $i \in [m]$ and $j \in [K]$. Then, by the non-negativity of the variables $t_{i,j}$, we also have that, for any subset $I \subset [m]$ and any $j \in [K]$,

$$\sum_{i \in I} \bar{q}_i \leq l_j \sum_{i \in I} q_i + \sum_{i \in I} t_{i,j} \leq l_j \sum_{i \in I} q_i + \sum_{i=1}^{m} t_{i,j} \leq l_j \sum_{i \in I} q_i + b_j.$$

Hence, $\sum_{i \in I} \bar{q}_i \leq \min_{j \in [K]} h_j(\sum_{i \in I} q_i)$ and thus $(\mathbf{q}, \bar{\mathbf{q}}) \in \mathcal{U}_{\phi,h}(\mathbf{p})$.

Conversely, let $(\mathbf{q}, \bar{\mathbf{q}}) \in \mathcal{U}_{\phi,h}(\mathbf{p})$. Then, $\bar{q}_i \leq l_j \cdot q_i + t_{i,j}$ for all $i \in [m]$ and $j \in [K]$ with $t_{i,j} \triangleq \max\{\bar{q}_i - l_j q_i, 0\}$. Moreover, we have that, for all $j \in [K]$,

$$\sum_{i=1}^{m} t_{i,j} = \sum_{i=1}^{m} \max\{\bar{q}_i - l_j q_i, 0\} = \sum_{i \in I_+} \bar{q}_i - \sum_{i \in I_+} l_j q_i \leq b_j, \text{ where } I_+ \triangleq \{i : \bar{q}_i - l_j q_i \geq 0\}.$$

Hence, $(\mathbf{q}, \bar{\mathbf{q}}, \mathbf{t})$ satisfies the constraints in (1.24). □

[13] The set $\mathcal{U}_{\phi,h}(\mathbf{p})$ is compact due to the lower-semicontinuity assumption made in Assumption 1.4.

Proof of Theorem 1.21. We reformulate the constraint

$$\max_{(\mathbf{q},\bar{\mathbf{q}},\mathbf{t})\in\bar{\mathcal{U}}} -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)) \leq c,$$

where

$$\bar{\mathcal{U}} \triangleq \left\{ (\mathbf{q},\bar{\mathbf{q}},\mathbf{t}) \in \mathbb{R}^{2m}_{\geq 0} \times \mathbb{R}^{mK}_{\geq 0} \;\middle|\; \begin{array}{l} \sum_{i=1}^m q_i = 1, \sum_{i=1}^m \bar{q}_i = 1, \sum_{i=1}^m p_i \phi\left(\frac{q_i}{p_i}\right) \leq r \\ \sum_{i=1}^m t_{i,j} \leq b_j, \ \forall j \in [K] \\ \bar{q}_i - l_j q_i \leq t_{i,j}, \ \forall i \in [m], \ \forall j \in [K] \end{array} \right\}.$$

We note that $(\mathbf{p},\mathbf{p},\{\max\{(l_j-1)p_i,0\}_{i,j}\})$ is a Slater point in $\bar{\mathcal{U}}$ and that the left-hand side of the constraint above constitutes a maximization problem upper bounded by $c \in \mathbb{R}$. Therefore, strong duality holds, and we examine the Lagrangian function

$$\begin{aligned}
&L(\mathbf{a},\mathbf{q},\bar{\mathbf{q}},\alpha,\beta,\gamma,\lambda_{ij},t_{i,j},\nu_j) \\
&= -\sum_{i=1}^{m} \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)) - \alpha\left(\sum_{i=1}^{m} q_i - 1\right) - \beta\left(\sum_{i=1}^{m} \bar{q}_i - 1\right) - \gamma\left(\sum_{i=1}^{m} p_i \phi\left(\frac{q_i}{p_i}\right) - r\right) \\
&\quad - \sum_{i=1}^{m}\sum_{j=1}^{K} \lambda_{ij}(\bar{q}_i - l_j q_i - t_{i,j}) - \sum_{j=1}^{K} \nu_j \left(\sum_{i=1}^{m} t_{i,j} - b_j\right) \\
&= \alpha + \beta + \gamma r + \sum_{j=1}^{K} \nu_j b_j - \sum_{i=1}^{m} \left(u(f(\mathbf{a},\mathbf{x}_i)) + \beta + \sum_{j=1}^{K} \lambda_{ij}\right) \bar{q}_i \\
&\quad + \sum_{i=1}^{m} p_i \left(\left(-\alpha + \sum_{j=1}^{K} \lambda_{ij} l_j\right) \frac{q_i}{p_i} - \gamma\phi\left(\frac{q_i}{p_i}\right)\right) + \sum_{i=1}^{m}\sum_{j=1}^{K} (\lambda_{ij} - \nu_j) t_{i,j}.
\end{aligned}$$

We have

$$\begin{aligned}
&\sup_{\bar{q}_1,\ldots,\bar{q}_m \geq 0} -\sum_{i=1}^{m} \left(u(f(\mathbf{a},\mathbf{x}_i)) + \beta + \sum_{j=1}^{K} \lambda_{ij}\right) \bar{q}_i \\
&= \begin{cases} 0 & \text{if } u(f(\mathbf{a},\mathbf{x}_i)) + \beta + \sum_{j=1}^{K} \lambda_{ij} \geq 0, \ \forall i \in [m] \\ \infty & \text{else,} \end{cases}
\end{aligned} \tag{1.50}$$

$$\sup_{q_1,\ldots,q_m \geq 0} \sum_{i=1}^{m} p_i \left(\left(-\alpha + \sum_{j=1}^{K} \lambda_{ij} l_j\right) \frac{q_i}{p_i} - \gamma\phi\left(\frac{q_i}{p_i}\right)\right) = \sum_{i=1}^{m} p_i \gamma \phi^*\left(\frac{-\alpha + \sum_{j=1}^{K} \lambda_{ij} l_j}{\gamma}\right),$$

and

$$\sup_{t_{1,1},\ldots,t_{m,K}\geq 0}\sum_{i=1}^{m}\sum_{j=1}^{K}(\lambda_{ij}-\nu_j)t_{i,j}=\begin{cases}0 & \text{if } \lambda_{ij}\leq\nu_j,\ \forall i\in[m],\ \forall j\in[K]\\ \infty & \text{else.}\end{cases} \tag{1.51}$$

Therefore, employing the same arguments as in the proof of Theorem 1.7, the reformulated robust counterpart is given by

$$\begin{cases}\alpha+\beta+\gamma r+\sum_{j=1}^{K}\nu_j b_j+\sum_{i=1}^{m}p_i\gamma\phi^*\left(\frac{-\alpha+\sum_{j=1}^{K}\lambda_{ij}l_j}{\gamma}\right)\leq c\\ -u(f(\mathbf{a},\mathbf{x}_i))-\beta-\sum_{j=1}^{K}\lambda_{ij}\leq 0,\ \forall i\in[m]\\ \lambda_{ij}\leq\nu_j,\ \forall i\in[m],\ \forall j\in[K]\\ \alpha,\beta\in\mathbb{R},\gamma,\lambda_{ij},\nu_j\geq 0.\end{cases}$$

In the nominal case where $\mathbf{q}=\mathbf{p}$, we have to reformulate the following constraint:

$$\max_{(\bar{\mathbf{q}},\mathbf{t})\in\bar{\mathcal{U}}_{\text{nom}}}-\sum_{i=1}^{m}\bar{q}_i u(f(\mathbf{a},\mathbf{x}_i))\leq c,$$

where

$$\bar{\mathcal{U}}_{\text{nom}}\triangleq\left\{(\bar{\mathbf{q}},\mathbf{t})\in\mathbb{R}^m_{\geq 0}\times\mathbb{R}^{mK}_{\geq 0}\;\middle|\;\begin{array}{l}\sum_{i=1}^{m}\bar{q}_i=1\\ \sum_{i=1}^{m}t_{i,j}\leq b_j,\ \forall j\in[K]\\ \bar{q}_i-l_jp_i\leq t_{i,j},\ \forall i\in[m],\ \forall j\in[K]\end{array}\right\}.$$

The above maximization problem is a linear programming problem bounded from above. Therefore, strong duality applies, and we examine the Lagrangian function

$$\begin{aligned}&L(\mathbf{a},\bar{\mathbf{q}},\beta,\lambda_{ij},t_{i,j},\nu_j)\\ &=-\sum_{i=1}^{m}\bar{q}_i u(f(\mathbf{a},\mathbf{x}_i))-\beta\left(\sum_{i=1}^{m}\bar{q}_i-1\right)-\sum_{i=1}^{m}\sum_{j=1}^{K}\lambda_{ij}(\bar{q}_i-l_jp_i-t_{i,j})-\sum_{j=1}^{K}\nu_j\left(\sum_{i=1}^{m}t_{i,j}-b_j\right)\\ &=\beta+\sum_{j=1}^{K}\nu_j b_j+\sum_{i=1}^{m}\sum_{j=1}^{K}\lambda_{ij}l_jp_i+\sum_{i=1}^{m}\bar{q}_i\left(-u(f(\mathbf{a},\mathbf{x}_i))-\beta-\sum_{j=1}^{K}\lambda_{ij}\right)+\sum_{i=1}^{m}\sum_{j=1}^{K}(\lambda_{ij}-\nu_j)t_{i,j}.\end{aligned}$$

By (1.50)–(1.51), we have that

$$\begin{aligned}&\inf_{\substack{\lambda_{ij},\nu_j\geq 0\\ \beta\in\mathbb{R}}}\sup_{\bar{\mathbf{q}},\mathbf{t}\geq\mathbf{0}}L(\mathbf{a},\bar{\mathbf{q}},\beta,\lambda_{ij},t_{i,j},\nu_j)\\ &=\inf_{\substack{\lambda_{ij},\nu_j\geq 0\\ \beta\in\mathbb{R}}}\left\{\beta+\sum_{j=1}^{K}\nu_j b_j+\sum_{i=1}^{m}\sum_{j=1}^{K}\lambda_{ij}l_jp_i\;\middle|\;-u(f(\mathbf{a},\mathbf{x}_i))-\beta-\sum_{j=1}^{K}\lambda_{ij}\leq 0,\forall i,\ \lambda_{ij}\leq\nu_j,\forall i,j\right\}.\end{aligned}$$

Hence, the reformulated robust counterpart is given by

$$\begin{cases} \beta + \sum_{j=1}^{K} \nu_j b_j + \sum_{i=1}^{m} \sum_{j=1}^{K} \lambda_{ij} l_j p_i \leq c \\ -u(f(\mathbf{a}, \mathbf{x}_i)) - \beta - \sum_{j=1}^{K} \lambda_{ij} \leq 0, \ \forall i \in [m] \\ \lambda_{ij} \leq \nu_j, \ \forall i \in [m], \ \forall j \in [K] \\ \lambda_{ij}, \nu_j \geq 0, \ \forall i \in [m], \ \forall j \in [K]. \end{cases}$$

□

Proof of Lemma 1.22. First, we have that

$$e_i(x_{i+1}) = \sup_{x \in [x_i, 1]} \left\{ h(x) - \frac{h(x_{i+1}) - h(x_i)}{x_{i+1} - x_i}(x - x_i) - h(x_i) \right\},$$

since, by concavity, the supremum is only taken in the interval $[x_i, x_{i+1}]$. Therefore, we can extend the feasibility region to $[x_i, 1]$.

Next, again by concavity, we have that $-\frac{h(x_{i+1})-h(x_i)}{x_{i+1}-x_i}(x - x_i)$ is an increasing function of x_{i+1}, for all $x \in (x_i, 1]$. Therefore, $e_i(x_{i+1})$ is increasing in x_{i+1}. It is also a continuous function in x_{i+1}. This is because the function

$$\tilde{e}_i(y) \triangleq \sup_{x \in [x_i, 1]} \{h(x) - y(x - x_i) - h(x_i)\},$$

is convex in y; indeed, it is a supremum of a linear function of y. Hence, $\tilde{e}_i$ is continuous on the interior of its domain, which is the whole of $\mathbb{R}$, because $\tilde{e}_i(y)$ is a supremum of a continuous function on a compact interval. Thus, $\tilde{e}_i(y)$ exists and is finite everywhere. This implies that e_i is continuous for all $x_{i+1} \in (x_i, 1]$, since $e_i(x_{i+1}) = \tilde{e}_i\left(\frac{h(x_{i+1})-h(x_i)}{x_{i+1}-x_i}\right)$ is a composition of continuous functions.

Finally, we show the existence of a x_{i+1} such that $e_i(x_{i+1}) = \epsilon$, for any given $\epsilon > 0$, under the assumption that $e_i(1) > \epsilon$. This is guaranteed if we can find a $z \in (x_i, 1)$ such that $e_i(z) < \epsilon$. Since h has decreasing slope and is increasing, we have that for any $y \in (x_i, 1)$ and all $x \in [x_i, y]$:

$$h(x) - \frac{h(y) - h(x_i)}{y - x_i}(x - x_i) - h(x_i) \leq h(y) - h(x_i).$$

We note that the maximization problem in $e_i(y)$ can be restricted to the interval $[x_i, y]$. Therefore, $e_i(y) \leq h(y) - h(x_i)$. Taking $y \downarrow x_i$, it follows by the continuity of h that such a point z must exist. □

Proof of Theorem 1.23 (Part I). We split the proof into two parts. We first treat the case of problem (P) and next the case of problem (P-constraint). Fix an $\mathbf{a} \in \mathcal{A}$ and consider the ranked realizations $u(f(\mathbf{a}, \mathbf{x}_{(1)})) \geq \ldots \geq u(f(\mathbf{a}, \mathbf{x}_{(m)}))$. Let h_1, h_2 be any two

concave distortion functions such that $\sup_{t\in[0,1]} |h_1(t) - h_2(t)| \leq \epsilon$. Set $u(f(\mathbf{a}, \mathbf{x}_{(0)})) \triangleq 0$. Then, we have

$$\begin{aligned}
\rho_{u,h_1,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) &= \sum_{i=1}^{m} h_1\left(\sum_{j=i}^{m} q_{(j)}\right)\left(u(f(\mathbf{a},\mathbf{x}_{(i-1)})) - u(f(\mathbf{a},\mathbf{x}_{(i)}))\right) \\
&\leq -u(f(\mathbf{a},\mathbf{x}_{(1)})) + \sum_{i=2}^{m}\left(h_2\left(\sum_{j=i}^{m} q_{(j)}\right) + \epsilon\right)\left(u(f(\mathbf{a},\mathbf{x}_{(i-1)})) - u(f(\mathbf{a},\mathbf{x}_{(i)}))\right) \\
&= \rho_{u,h_2,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) + \epsilon\cdot\left(\max_{i\in[m]} u(f(\mathbf{a},\mathbf{x}_i)) - \min_{i\in[m]} u(f(\mathbf{a},\mathbf{x}_i))\right) \\
&\leq \rho_{u,h_2,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) + \epsilon\cdot\left(M + \max_{i\in[m]} -u(f(\mathbf{a},\mathbf{x}_i))\right),
\end{aligned}$$

where $M \triangleq \sup_{\mathbf{a}\in\mathcal{A},i\in[m]} u(f(\mathbf{a},\mathbf{x}_i)) < \infty$. The idea now is to bound the term $\max_{i\in[m]}$ $-u(f(\mathbf{a},\mathbf{x}_i))$ on a subset $\mathcal{A}_0 \subset \mathcal{A}$ of the feasible region, for which the restriction of the minimization problem (P) on $\mathcal{A}_0$ does not change the original optimal value.

We take $h_1 \equiv h$ and $h_2 \in \{h_\epsilon, \tilde{h}_\epsilon\}$. By Assumption 1.2, there exists $\mathbf{a}_0 \in \mathcal{A}$ such that $\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}_0,\mathbf{X})) < \infty$. This implies $\rho_{u,h,\mathbf{p}}(f(\mathbf{a}_0,\mathbf{X})) < \infty$, hence $\max_{i\in[m]} -u(f(\mathbf{a}_0,\mathbf{x}_i)) < \infty$. Then, we also have that

$$\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_2,\mathbf{q}}(f(\mathbf{a}_0,\mathbf{X})) \leq \max_{i\in[m]} -u(f(\mathbf{a}_0,\mathbf{x}_i)) < \infty.$$

Therefore, we may define the finite number

$$c_0 \triangleq \max_{j=1,2}\left\{\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_j,\mathbf{q}}(f(\mathbf{a}_0,\mathbf{X}))\right\} < \infty.$$

Then, for all $\mathbf{a} \in \mathcal{A}$ such that $\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_j,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \leq c_0$ for any $j \in \{1,2\}$, we have that

$$-\sum_{i=1}^{m} p_i u(f(\mathbf{a},\mathbf{x}_i)) \leq \rho_{u,h_j,\mathbf{p}}(f(\mathbf{a},\mathbf{X})) \leq c_0, \tag{1.52}$$

due to concavity (which implies $h_j(x) \geq x, \forall x \in [0,1]$, for any $j \in \{1,2\}$). Therefore, for all such $\mathbf{a}$, we have,

$$\sup_{i\in[m]} -u(f(\mathbf{a},\mathbf{x}_i)) \leq \frac{c_0 + M}{p_{\min}}, \tag{1.53}$$

as shown in the proof of Theorem 1.14. Define,

$$\mathcal{A}_0 \triangleq \left\{\mathbf{a}\in\mathcal{A} \;\middle|\; \sup_{i\in[m]} -u(f(\mathbf{a},\mathbf{x}_i)) \leq \frac{c_0 + M}{p_{\min}}\right\}.$$

Then, we have that for any $j = 1, 2$:

$$\min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_j,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) = \min_{\mathbf{a}\in\mathcal{A}_0} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_j,\mathbf{q}}(f(\mathbf{a},\mathbf{X})),$$

since any potential minimizer $\mathbf{a}$ satisfies $\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_j,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \le c_0$ and thus (1.53). Therefore,

$$\begin{aligned}
\min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_1,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) &= \min_{\mathbf{a}\in\mathcal{A}_0} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_1,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \\
&\le \min_{\mathbf{a}\in\mathcal{A}_0} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_2,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) + \epsilon\cdot\left(M + \frac{c_0+M}{p_{\min}}\right) \\
&= \min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_2,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) + \epsilon\cdot\left(M + \frac{c_0+M}{p_{\min}}\right).
\end{aligned}$$

By symmetry, we thus have

$$\left|\min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_1,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) - \min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_2,\mathbf{q}}(f(\mathbf{a},\mathbf{X}))\right| \le \epsilon\cdot\left(M + \frac{c_0+M}{p_{\min}}\right),$$

which approaches zero as $\epsilon \to 0$. □

Proof of Theorem 1.23 (Part II). We define

$$\begin{aligned}
L_\epsilon &\triangleq \min_{\mathbf{a}\in\mathcal{A}} \left\{ g(\mathbf{a}) \;\middle|\; \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_\epsilon,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \le c \right\}, \\
U_\epsilon &\triangleq \min_{\mathbf{a}\in\mathcal{A}} \left\{ g(\mathbf{a}) \;\middle|\; \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,\tilde{h}_\epsilon,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \le c \right\},
\end{aligned}$$

and let P_0 denote the optimal objective value of (P-constraint). By definition, $L_\epsilon \le P_0 \le U_\epsilon$.

We first note that L_ϵ has a nonempty feasible set, which follows from Assumption 1.5 and the fact that $\rho_{u,h_\epsilon,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \le \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X}))$ for all $\mathbf{a} \in \mathcal{A}$ and any probability vector $\mathbf{q}$. Then, following the first part of the proof of Theorem 1.23, we can show that for any $\mathbf{a} \in \mathcal{A}$ that is feasible for L_ϵ, we have that, for all $\mathbf{q}$,

$$\rho_{u,h_\epsilon,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \ge \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) - \epsilon\left(M + \frac{c+M}{p_{\min}}\right).$$

Therefore, we also have the following implication for any $\mathbf{a} \in \mathcal{A}$:

$$\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_\epsilon,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \le c \Rightarrow \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \le c + \epsilon\left(M + \frac{c+M}{p_{\min}}\right).$$

Hence,

$$\begin{aligned}
L_\epsilon &\geq \min_{\mathbf{a}\in\mathcal{A}} \left\{ g(\mathbf{a}) \;\middle|\; \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \leq c + \epsilon\left(M + \frac{c+M}{p_{\min}}\right) \right\} \\
&\overset{(*)}{\geq} \sup_{\lambda\geq 0} \inf_{\mathbf{a}\in\mathcal{A}} g(\mathbf{a}) + \lambda\left(\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) - c - \epsilon\left(M + \frac{c+M}{p_{\min}}\right)\right) \\
&\geq \inf_{\mathbf{a}\in\mathcal{A}} g(\mathbf{a}) + \lambda^*\left(\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) - c - \epsilon\left(M + \frac{c+M}{p_{\min}}\right)\right) \\
&\overset{(**)}{=} P_0 - \lambda^*\cdot\epsilon\left(M + \frac{c+M}{p_{\min}}\right),
\end{aligned}$$

where in $(*)$ we used weak duality and in $(**)$ we used λ^*, the KKT-vector of (P-constraint) corresponding to the constraint $\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \leq c$, which is a strictly positive constant by Assumption 1.5 and does not depend on ϵ. Therefore, we have

$$0 \leq P_0 - L_\epsilon \leq \lambda^*\epsilon\left(M + \frac{c+M}{p_{\min}}\right),$$

which approaches zero as $\epsilon \to 0$.

Similarly, we have the following implication for any $\mathbf{a} \in \mathcal{A}$:

$$\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \leq c - \epsilon\left(M + \frac{c+M}{p_{\min}}\right) \Rightarrow \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,\tilde{h}_\epsilon,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \leq c.$$

Therefore,

$$U_\epsilon \leq \min_{\mathbf{a}\in\mathcal{A}} \left\{ g(\mathbf{a}) \;\middle|\; \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \leq c - \epsilon\left(M + \frac{c+M}{p_{\min}}\right) \right\}. \tag{1.54}$$

We note that the minimization problem on the right-hand side of (1.54) contains a Slater point, for ϵ sufficiently small. Indeed, by Assumption 1.5, there exists a point $\mathbf{a}_0 \in \text{int}(\mathcal{A})$, such that

$$\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}_0,\mathbf{X})) \triangleq c_0 < c.$$

Then, for $\epsilon \leq \frac{c-c_0}{2\left(M+\frac{c+M}{p_{\min}}\right)}$, we have that

$$\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}_0,\mathbf{X})) < c - \epsilon\left(M + \frac{c+M}{p_{\min}}\right). \tag{1.55}$$

Therefore, together with Assumption 1.5, as well as the reformulation in Theorem 1.6, we

may apply the strong duality theorem and obtain the upper estimation

$$\begin{aligned}
U_\epsilon &\leq \sup_{\lambda\geq 0}\min_{\mathbf{a}\in\mathcal{A}} g(\mathbf{a}) + \lambda\left(\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) - c + \epsilon\left(M + \frac{c+M}{p_{\min}}\right)\right)\\
&= \min_{\mathbf{a}\in\mathcal{A}} g(\mathbf{a}) + \lambda^*(\epsilon)\left(\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) - c + \epsilon\left(M + \frac{c+M}{p_{\min}}\right)\right)\\
&\leq \lambda^*(\epsilon)\cdot\epsilon\left(M + \frac{c+M}{p_{\min}}\right) + \sup_{\lambda\geq 0}\min_{\mathbf{a}\in\mathcal{A}} g(\mathbf{a}) + \lambda\left(\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) - c\right)\\
&= \lambda^*(\epsilon)\cdot\epsilon\left(M + \frac{c+M}{p_{\min}}\right) + P_0,
\end{aligned}$$

where $\lambda^*(\epsilon)$ is the KKT-vector of the minimization problem on the right-hand side of (1.54), corresponding to the supremum constraint, which depends on ϵ. As a final step, we will show that $\lambda^*(\epsilon)$ can be further bounded by a constant that does not depend on ϵ, for ϵ sufficiently small. Indeed, let $\mathbf{a}_0$ be the point in (1.55) and let $\epsilon \leq \frac{c-c_0}{2\left(M+\frac{c+M}{p_{\min}}\right)}$. Since $P_0 \leq U_\epsilon$,

$$\begin{aligned}
P_0 &\leq \min_{\mathbf{a}\in\mathcal{A}} g(\mathbf{a}) + \lambda^*(\epsilon)\left(\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) - c + \epsilon\left(M + \frac{c+M}{p_{\min}}\right)\right)\\
&\leq g(\mathbf{a}_0) + \lambda^*(\epsilon)\left(\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}_0,\mathbf{X})) - c + \epsilon\left(M + \frac{c+M}{p_{\min}}\right)\right)\\
&\leq g(\mathbf{a}_0) - \lambda^*(\epsilon)\frac{c-c_0}{2},
\end{aligned}$$

which implies that, for all $\epsilon \leq \frac{c-c_0}{2\left(M+\frac{c+M}{p_{\min}}\right)}$, we have

$$\lambda^*(\epsilon) \leq \frac{2(g(\mathbf{a}_0) - P_0)}{c - c_0} \triangleq C^*.$$

Hence, for ϵ sufficiently small,

$$0 \leq U_\epsilon - P_0 \leq C^*\cdot\epsilon\left(M + \frac{c+M}{p_{\min}}\right),$$

which converges to zero as $\epsilon \to 0$. □

Proof of Proposition 1.24. Choose the function $h(p) = p^2$. Then $h(0) = 0$, $h(1) = 1$ and $h(p) < p$, $\forall p \in (0,1)$. Due to the compactness of $\mathcal{U}$, there exists a maximizer $\mathbf{q}^*$ of $\sup_{\mathbf{q}\in\mathcal{U}} \rho_{h,\mathbf{q}}(X)$. Let $\hat{h}(p) = p$ be the concave envelope of h. By the assumptions on $\mathcal{U}$, we

have that $\mathbf{q}_1^* < 1$. Hence,

$$\rho_{\hat{h},\mathbf{q}^*}(X) - \rho_{h,\mathbf{q}^*}(X) = \sum_{i=2}^{m} \left(\hat{h}\left(\sum_{k=i}^{m} q_k^* \right) - h\left(\sum_{k=i}^{m} q_k^* \right) \right) (x_{i-1} - x_i) > 0,$$

since $\hat{h}(p) > h(p)$ for all $p \in (0,1)$ and $x_{i-1} - x_i > 0$ for all $i \in [m]$. Therefore, $\sup_{\mathbf{q}\in\mathcal{U}} \rho_{\hat{h},\mathbf{q}}(X) > \sup_{\mathbf{q}\in\mathcal{U}} \rho_{h,\mathbf{q}}(X)$. □

Proof of Theorem 1.25. It is sufficient to show that for any X, with outcomes $\{x_i\}_{i=1}^m$, we have

$$\rho_{h,\mathbf{p}}(X) = \sup_{\mathbf{q}\in M_h^{\mathrm{ca}}(\mathbf{p})} \sum_{i=1}^{m} -q_i x_i - \sup_{\bar{\mathbf{q}}\in N_{\bar{h}}^{\mathrm{cv}}(\mathbf{p})} \sum_{i=1}^{m} \bar{q}_i x_i,$$

since $\rho_{u,h,\mathbf{p}}(f(\mathbf{a},\mathbf{X})) = \rho_{h,\mathbf{p}}(u(f(\mathbf{a},\mathbf{X}))$. Let $-x_{(1)} \geq \ldots \geq -x_{(m)}$ be the ranked realizations of X. Let k^* be the index where $h\left(\sum_{s=1}^{k^*} p_{(s)}\right) \leq h(p^0) \leq h\left(\sum_{s=1}^{k^*+1} p_{(s)}\right)$. Then,

$$\begin{aligned}
&\rho_{h,\mathbf{p}}(X) \\
&= \sum_{i=1}^{k^*} \left(h\left(\sum_{s=1}^{i} p_{(s)} \right) - h\left(\sum_{s=1}^{i-1} p_{(s)} \right) \right) \cdot (-x_{(i)}) + \left(h(p^0) - h\left(\sum_{s=1}^{k^*} p_{(s)} \right) \right) \cdot (-x_{(k^*+1)}) \\
&\quad + \left(h\left(\sum_{s=1}^{k^*+1} p_{(s)} \right) - h(p^0) \right) \cdot (-x_{(k^*+1)}) + \sum_{i=k^*+2}^{m} \left(h\left(\sum_{s=1}^{i} p_{(s)} \right) - h\left(\sum_{s=1}^{i-1} p_{(s)} \right) \right) \\
&\quad \cdot (-x_{(i)}),
\end{aligned}$$

where an empty sum $\sum_{s=1}^{0}$ is zero by convention. We can also express the second line of the previous expression in terms of the dual function $\bar{h}(p) \triangleq 1 - h(1-p)$, and rearrange the indices to obtain:

$$\begin{aligned}
&\sum_{i=k^*+2}^{m} \left(h\left(\sum_{s=1}^{i} p_{(s)} \right) - h\left(\sum_{s=1}^{i-1} p_{(s)} \right) \right) \cdot (-x_{(i)}) + \left(h\left(\sum_{s=1}^{k^*+1} p_{(s)} \right) - h(p^0) \right) \cdot (-x_{(k^*+1)}) \\
&= \sum_{i=k^*+2}^{m} \left(\bar{h}\left(\sum_{s=i}^{m} p_{(s)} \right) - \bar{h}\left(\sum_{s=i+1}^{m} p_{(s)} \right) \right) \cdot (-x_{(i)}) + \left(\bar{h}(1-p^0) - \bar{h}\left(\sum_{s=k^*+2}^{m} p_{(s)} \right) \right) \\
&\quad \cdot (-x_{(k^*+1)}) \\
&= - \sum_{i=1}^{m-k^*-1} \left(\bar{h}\left(\sum_{s=1}^{i} p_{(m-s+1)} \right) - \bar{h}\left(\sum_{s=1}^{i-1} p_{(m-s+1)} \right) \right) \cdot x_{(m-i+1)} \\
&\quad - \left(\bar{h}(1-p^0) - \bar{h}\left(\sum_{s=1}^{m-k^*-1} p_{(m-s+1)} \right) \right) \cdot x_{(k^*+1)},
\end{aligned}$$

where the empty sum $\sum_{s=m+1}^{m}$ is again zero. It remains to show that the sums are equal

to their dual representations:

$$\begin{aligned}\sup_{\mathbf{q}\in M_h^{\mathrm{ca}}(\mathbf{p})} \sum_{i=1}^{m} -q_i x_i = \sum_{i=1}^{k^*} \left(h\left(\sum_{s=1}^{i} p_{(s)}\right) - h\left(\sum_{s=1}^{i-1} p_{(s)}\right)\right)\cdot(-x_{(i)}) \\ + \left(h(p^0) - h\left(\sum_{s=1}^{k^*} p_{(s)}\right)\right)\cdot(-x_{(k^*+1)}),\end{aligned} \tag{1.56}$$

and

$$\begin{aligned}\sup_{\bar{\mathbf{q}}\in N_{\bar{h}}^{\mathrm{cv}}(\mathbf{p})} \sum_{i=1}^{m} \bar{q}_i x_i = \sum_{i=1}^{m-k^*-1} \left(\bar{h}\left(\sum_{s=1}^{i} p_{(m-s+1)}\right) - \bar{h}\left(\sum_{s=1}^{i-1} p_{(m-s+1)}\right)\right)\cdot x_{(m-i+1)} \\ + \left(\bar{h}(1-p^0) - \bar{h}\left(\sum_{s=1}^{m-k^*-1} p_{(m-s+1)}\right)\right)\cdot x_{(k^*+1)}.\end{aligned} \tag{1.57}$$

We will first show (1.56); (1.57) follows similarly. Let h_0 be defined as follows:

$$h_0(p) \triangleq \begin{cases} h(p) & p \le p^0 \\ h(p^0) & p \ge p^0. \end{cases} \tag{1.58}$$

Then, h_0 is non-decreasing and concave. Let P be the probability measure induced by the vector $(p_i)_{i=1}^m$. Then, by Lemma 1.B.1, $\mu_1 \triangleq h_0 \circ P$ is a monotone, submodular set function. By the definition of a Choquet integral, we have

$$\begin{aligned}&\int -X \mathrm{d}\mu_1 \\ &= \sum_{i=1}^{k^*} \left(h_0\left(\sum_{s=1}^{i} p_{(s)}\right) - h_0\left(\sum_{s=1}^{i-1} p_{(s)}\right)\right)\cdot(-x_{(i)}) + \left(h_0(p^0) - h_0\left(\sum_{s=1}^{k^*} p_{(s)}\right)\right)\cdot(-x_{(k^*+1)}) \\ &= \sum_{i=1}^{k^*} \left(h\left(\sum_{s=1}^{i} p_{(s)}\right) - h\left(\sum_{s=1}^{i-1} p_{(s)}\right)\right)\cdot(-x_{(i)}) + \left(h(p^0) - h\left(\sum_{s=1}^{k^*} p_{(s)}\right)\right)\cdot(-x_{(k^*+1)}).\end{aligned}$$

Moreover, since μ_1 is monotone and submodular, we have by Proposition 10.3 of Denneberg [1994],

$$\int -X \mathrm{d}\mu_1 = \left\{ \sum_{i=1}^{m} -q_i x_i \;\middle|\; q_i \ge 0, \sum_{i\in J} q_i \le h_0\left(\sum_{i\in J} p_i\right),\ \forall J \subset [m],\ \sum_{i=1}^{m} q_i = h_0(p^0) \right\},$$

which involves the same feasible set as in (1.29).

Similarly, define the function

$$\bar{h}_0(p) \triangleq \begin{cases} \bar{h}(p) & p \le 1-p^0 \\ \bar{h}(1-p^0) & p \ge 1-p^0. \end{cases} \tag{1.59}$$

Then, $\bar{h}_0$ is concave and non-decreasing. Hence, by the same arguments as above, we have that $\mu_2 \triangleq \bar{h}_0 \circ P$ is monotone and submodular. Therefore,

$$\int X \mathrm{d}\mu_2 = \sum_{i=1}^{m-k^*-1} \left(\bar{h}\left(\sum_{s=1}^{i} p_{(m-s+1)} \right) - \bar{h}\left(\sum_{s=1}^{i-1} p_{(m-s+1)} \right) \right) \cdot x_{(m-i+1)} + \left(\bar{h}(1-p^0) - \bar{h}\left(\sum_{s=1}^{m-k^*-1} p_{(s)} \right) \right) \cdot x_{(k^*+1)}.$$

Moreover, we have

$$\int X \mathrm{d}\mu_2 = \sup \left\{ \sum_{i=1}^{m} \bar{q}_i x_i \;\middle|\; \bar{q}_i \geq 0, \sum_{i \in J} \bar{q}_i \leq \bar{h}_0 \left(\sum_{i \in J} p_i \right),\ \forall J \subset [m],\ \sum_{i=1}^{m} \bar{q}_i = \bar{h}_0(1-p^0) \right\},$$

which implies (1.30). □

Lemma 1.B.1. *Let Ω be a set and let $\mathcal{F} = 2^\Omega$ be the collection of all subsets of Ω. Furthermore, let $h : [0,1] \to \mathbb{R}$ be a non-decreasing concave function such that $h(0) = 0$ and let $P : \mathcal{F} \to [0, \infty)$ be an additive set function. Then, $\mu \triangleq h \circ P$ is a monotone, submodular set function.*

Proof. Monotonicity of μ is clear due to the additivity of P and the monotonicity of h. Let $A, B \in \mathcal{F}$. Additivity of P also implies that

$$P(A \cup B) + P(A \cap B) = P(A) + P(B).$$

Let $a \triangleq P(A),\ b \triangleq P(B)$. Then, we have

$$P(A \cap B) \triangleq i \leq a \leq b \leq u \triangleq P(A \cup B).$$

Since $b - i = u - a$, we have that by concavity of h,

$$\frac{h(b) - h(i)}{b - i} \geq \frac{h(u) - h(a)}{u - a},$$

which, via $h(i) + h(u) \leq h(a) + h(b)$, implies the submodularity of μ. □

Proof of Theorem 1.27. By Theorem 1.25, it is sufficient to show that the sets $M_h^{\mathrm{ca}}(\mathbf{p})$ and $N_{\bar{h}}^{\mathrm{cv}}(\mathbf{p})$ are the projection of the following sets on the coordinates $\mathbf{q}$ and $\bar{\mathbf{q}}$, respectively:

$$M_{h,l}^{\mathrm{ca}}(\mathbf{p}) = \left\{ \mathbf{q} \in \mathbb{R}^m, \mathbf{t}^{(1)} \in \mathbb{R}^{m \times K_1} \;\middle|\; \begin{array}{l} q_i \geq 0 \\ q_i \leq l_k^{(1)} p_i + t_{ik}^{(1)},\ \forall i \in [m],\ \forall k \in [K_1] \\ \sum_{i=1}^m t_{ik}^{(1)} \leq b_k^{(1)},\ \forall k \in [K_1] \\ \sum_{i=1}^m q_i = h(p^0) \end{array} \right\}, \quad (1.60)$$

and

$$N^{\mathrm{cv}}_{\bar{h},l}(\mathbf{p}) = \left\{ \bar{\mathbf{q}} \in \mathbb{R}^m, \mathbf{t}^{(2)} \in \mathbb{R}^{m\times K_2} \;\middle|\; \begin{array}{l} \bar{q}_i \geq 0 \\ \bar{q}_i \leq l_k^{(2)} p_i + t_{ik}^{(2)},\ \forall i \in [m],\ \forall k \in [K_2] \\ \sum_{i=1}^m t_{ik}^{(2)} \leq b_k^{(2)}, \forall k \in [K_2] \\ \sum_{i=1}^m \bar{q}_i = \bar{h}(1-p^0) \end{array} \right\}, \quad (1.61)$$

We will only show this for $M^{\mathrm{ca}}_{h,l}(\mathbf{p})$, since the case for $N^{\mathrm{cv}}_{\bar{h},l}(\mathbf{p})$ is identical, *mutatis mutandis*.

Let $(\mathbf{q}, \mathbf{t}^{(1)}) \in M^{\mathrm{ca}}_{h,l}(\mathbf{p})$. Then, $q_i \leq l_k^{(1)} \cdot p_i + t_{ik}^{(1)}$ for all $i \in [m]$ and $k \in [K_1]$. Thus, for any subset $I \subset [m]$, we also have that, for all $j \in [K_1]$,

$$\sum_{i\in I} q_i \leq l_k^{(1)} \sum_{i\in I} p_i + \sum_{i\in I} t_{ik}^{(1)} \leq l_k^{(1)} \sum_{i\in I} p_i + \sum_{i=1}^m t_{ik}^{(1)} \leq l_k^{(1)} \sum_{i\in I} p_i + b_k^{(1)}.$$

Hence, $\mathbf{q} \in M^{\mathrm{ca}}_h(\mathbf{p})$. Conversely, let $\mathbf{q} \in M^{\mathrm{ca}}_h(\mathbf{p})$. Then, $q_i \leq l_k^{(1)} \cdot p_i + t_{ik}^{(1)}$ for all $i \in [m]$ and $k \in [K_1]$ with $t_{ik}^{(1)} \triangleq \max\{q_i - l_k^{(1)} p_i, 0\}$. Moreover, we have, for all $k \in [K_1]$,

$$\sum_{i=1}^m t_{ik}^{(1)} = \sum_{i=1}^m \max\{q_i - l_k^{(1)} p_i, 0\} = \sum_{i\in I_+} q_i - \sum_{i\in I_+} l_k^{(1)} p_i \leq b_k^{(1)}, \text{ where } I_+ \triangleq \{i : q_i - l_k^{(1)} p_i \geq 0\}.$$

Indeed, the last inequality follows from the fact that, for any $\mathbf{q} \in M^{\mathrm{ca}}_h(\mathbf{p})$, we have that $\sum_{i\in I} \bar{q}_i \leq l_k^{(1)} \sum_{i\in I} q_i + b_k^{(1)}$ if $\sum_{i\in I} p_i \leq h(p^0)$. For any index set I such that $\sum_{i\in I} p_i > h(p^0)$, we also have that $l_k^{(1)} \sum_{i\in I} p_i + b_k^{(1)} > l_k^{(1)} h(p^0) + b_k^{(1)} \overset{(*)}{\geq} l_{K_1}^{(1)} h(p^0) + b_{K_1}^{(1)} = h(p^0) \geq \sum_{i\in I} q_i$, where $(*)$ follows from our ordering of the slopes and the intercepts $(l_k^{(1)}, b_k^{(1)})_{k=1}^{K_1}$ as described in (1.32). Hence, the inequality under consideration indeed holds for the index set I_+. Thus, $(\mathbf{q}, \mathbf{t}^{(1)}) \in M^{\mathrm{ca}}_{h,l}(\mathbf{p})$. The reformulation in (1.34) now follows as in Theorem 1.21. □

Proof of Theorem 1.28. We first note that, by Theorem 1.25,

$$\rho_{u,h,\mathbf{q}}(f(\mathbf{a}, \mathbf{X})) = \sup_{\bar{\mathbf{q}} \in M^{\mathrm{ca}}_h(\mathbf{q})} \sum_{i=1}^m -\bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)) - \sup_{\tilde{\mathbf{q}} \in N^{\mathrm{cv}}_{\bar{h}}(\mathbf{q})} \sum_{i=1}^m \tilde{q}_i u(f(\mathbf{a}, \mathbf{x}_i)).$$

Since $\mathcal{U}_t \subset \mathcal{D}_\phi(\mathbf{p}, r)$, we have that for each iteration t, the cutting-plane procedure yields a lower bound c^t on the true optimal objective value (P), which we denote by $\tilde{c}$. By Assumption 1.2, $\tilde{c}$ is finite. Hence, we may assume that $|c^s - c^t| \leq \frac{1}{2}\epsilon_{\mathrm{tol}}$, for all s, t sufficiently large. Otherwise, the lower bounds will attain or exceed $\tilde{c}$ after finitely many iterations. Assume now that $s > t$. Let $\mathbf{q}^t$ be the worst-case probability vector added to $\mathcal{U}_t$ and let $\mathbf{a}_t, \mathbf{a}_s$ be the optimal solutions obtained at the t, s-th iteration. By definition,

$$\sup_{\bar{\mathbf{q}} \in M^{\mathrm{ca}}_h(\mathbf{q}^t)} \sum_{i=1}^m -\bar{q}_i u(f(\mathbf{a}^t, \mathbf{x}_i)) - \sup_{\tilde{\mathbf{q}} \in N^{\mathrm{cv}}_{\bar{h}}(\mathbf{q}^t)} \sum_{i=1}^m \tilde{q}_i u(f(\mathbf{a}^t, \mathbf{x}_i)) > c^t + \epsilon_{\mathrm{tol}}, \quad (1.62)$$

and

$$\sup_{\bar{\mathbf{q}}\in M_h^{\mathrm{ca}}(\mathbf{q}^t)} \sum_{i=1}^m -\bar{q}_i u(f(\mathbf{a}^s,\mathbf{x}_i)) - \sup_{\tilde{\mathbf{q}}\in N_{\tilde{h}}^{\mathrm{cv}}(\mathbf{q}^t)} \sum_{i=1}^m \tilde{q}_i u(f(\mathbf{a}^s,\mathbf{x}_i)) \le c^s. \tag{1.63}$$

Therefore,

$$\begin{aligned}
&\sup_{\bar{\mathbf{q}}\in M_h^{\mathrm{ca}}(\mathbf{q}^t)} \sum_{i=1}^m -\bar{q}_i u(f(\mathbf{a}^t,\mathbf{x}_i)) - \sup_{\bar{\mathbf{q}}\in M_h^{\mathrm{ca}}(\mathbf{q}^t)} \sum_{i=1}^m -\bar{q}_i u(f(\mathbf{a}^s,\mathbf{x}_i)) \\
&\qquad + \sup_{\tilde{\mathbf{q}}\in N_{\tilde{h}}^{\mathrm{cv}}(\mathbf{q}^t)} \sum_{i=1}^m \tilde{q}_i u(f(\mathbf{a}^s,\mathbf{x}_i)) - \sup_{\tilde{\mathbf{q}}\in N_{\tilde{h}}^{\mathrm{cv}}(\mathbf{q}^t)} \sum_{i=1}^m \tilde{q}_i u(f(\mathbf{a}^t,\mathbf{x}_i)) \\
&> \epsilon_{\mathrm{tol}} - (c^s - c^t) \ge \frac{1}{2}\epsilon_{\mathrm{tol}}.
\end{aligned}$$

We can further upper bound the differences of the suprema as follows:

$$\begin{aligned}
&\sup_{\bar{\mathbf{q}}\in M_h^{\mathrm{ca}}(\mathbf{q}^t)} \sum_{i=1}^m -\bar{q}_i u(f(\mathbf{a}^t,\mathbf{x}_i)) - \sup_{\bar{\mathbf{q}}\in M_h^{\mathrm{ca}}(\mathbf{q}^t)} \sum_{i=1}^m -\bar{q}_i u(f(\mathbf{a}^s,\mathbf{x}_i)) \\
&\le \sup_{\bar{\mathbf{q}}\in M_h^{\mathrm{ca}}(\mathbf{q}^t)} \left| \sum_{i=1}^m -\bar{q}_i (u(f(\mathbf{a}^s,\mathbf{x}_i)) - u(f(\mathbf{a}^t,\mathbf{x}_i))) \right| \\
&\le \sup_{\bar{\mathbf{q}}\in M_h^{\mathrm{ca}}(\mathbf{q}^t)} \|\bar{\mathbf{q}}\|_2 \|(u(f(\mathbf{a}^s,\mathbf{x}_i)) - u(f(\mathbf{a}^t,\mathbf{x}_i)))_{i=1}^m\|_2 \\
&\le \|(u(f(\mathbf{a}^s,\mathbf{x}_i)) - u(f(\mathbf{a}^t,\mathbf{x}_i)))_{i=1}^m\|_2.
\end{aligned}$$

Similarly,

$$\sup_{\tilde{\mathbf{q}}\in N_{\tilde{h}}^{\mathrm{cv}}(\mathbf{q}^t)} \sum_{i=1}^m \tilde{q}_i u(f(\mathbf{a}^s,\mathbf{x}_i)) - \sup_{\tilde{\mathbf{q}}\in N_{\tilde{h}}^{\mathrm{cv}}(\mathbf{q}^t)} \sum_{i=1}^m \tilde{q}_i u(f(\mathbf{a}^t,\mathbf{x}_i)) \le \|(u(f(\mathbf{a}^s,\mathbf{x}_i)) - u(f(\mathbf{a}^t,\mathbf{x}_i)))_{i=1}^m\|_2.$$

Therefore,

$$\|(u(f(\mathbf{a}^s,\mathbf{x}_i)) - u(f(\mathbf{a}^t,\mathbf{x}_i)))_{i=1}^m\|_2 \ge \frac{1}{4}\epsilon_{\mathrm{tol}}.$$

However, we have the assumption that $\sup_{\mathbf{a}\in\mathcal{A},i\in[m]} |u(f(\mathbf{a},\mathbf{x}_i))| < \infty$. Hence, this leads to a similar contradiction as in the proof of Theorem 1.14. Therefore, the cutting-plane procedure must terminate. □

Proof of Theorem 1.29. Let $M \triangleq \sup_{\mathbf{a}\in\mathcal{A},i\in[m]} |u(f(\mathbf{a},\mathbf{x}_i))| < \infty$. Similar to the proof of Theorem 1.23, for any two distortion functions h_1, h_2 such that $\sup_{p\in[0,1]} |h_1(p) - h_2(p)| \le$

ϵ, we have that

$$\begin{aligned}\rho_{u,h_1,\mathbf{p}}(f(\mathbf{a},\mathbf{X})) &\leq \rho_{u,h_2,\mathbf{p}}(f(\mathbf{a},\mathbf{X})) + \epsilon \cdot \left(\max_{i\in[m]} u(f(\mathbf{a},\mathbf{x}_i)) - \min_{i\in[m]} u(f(\mathbf{a},\mathbf{x}_i))\right) \\ &\leq \rho_{u,h_2,\mathbf{p}}(f(\mathbf{a},\mathbf{X})) + 2M\epsilon.\end{aligned}$$

Since $2M\epsilon$ does not depend on both $\mathbf{a}$ and $\mathbf{q}$, we have that, in the nominal case,

$$\left|\min_{\mathbf{a}\in\mathcal{A}} \rho_{u,h_1,\mathbf{p}}(f(\mathbf{a},\mathbf{X})) - \min_{\mathbf{a}\in\mathcal{A}} \rho_{u,h_2,\mathbf{p}}(f(\mathbf{a},\mathbf{X}))\right| \leq 2M\epsilon,$$

and similarly, in the robust case,

$$\left|\min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_1,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) - \min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_2,\mathbf{q}}(f(\mathbf{a},\mathbf{X}))\right| \leq 2M\epsilon,$$

which both approach zero as $\epsilon \to 0$. Since Algorithm 3 yields a final objective value c^* such that

$$\left|c^* - \min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_\epsilon,\mathbf{q}}(f(\mathbf{a},\mathbf{X}))\right| \leq \epsilon_{\text{tol}},$$

it follows that

$$\left|c^* - \min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X}))\right| \leq \epsilon_{\text{tol}} + \epsilon 2M \to 0,$$

as $\epsilon_{\text{tol}}, \epsilon \to 0$. □

Proof of Theorem 1.30. The proof that Algorithm 3 terminates after finitely many steps for problem (P-constraint) is similar to the proof of Theorem 1.28, where we now take $c^s = c^t = c$, with c the constraint value in (P-constraint). To show convergence, we note that $\mathcal{U}_j \subset \mathcal{D}_\phi(\mathbf{p}, r)$. Hence, $g(\mathbf{a}_{\epsilon_{\text{tol}}})$ is always a lower bound of (P-constraint), for which we denote its optimal objective value by $P(0)$. We define

$$P(\epsilon_{\text{tol}}) \triangleq \min_{\mathbf{a}\in\mathcal{A}} \left\{ g(\mathbf{a}) \;\middle|\; \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \leq c + \epsilon_{\text{tol}} \right\}.$$

Then, by Algorithm 3, the final solution $\mathbf{a}_{\epsilon_{\text{tol}}}$ is feasible for the minimization problem of $P(\epsilon_{\text{tol}})$. Hence, $P(\epsilon_{\text{tol}}) \leq g(\mathbf{a}_{\epsilon_{\text{tol}}}) \leq P(0)$. It remains to show that $\lim_{\epsilon_{\text{tol}}\to 0} P(\epsilon_{\text{tol}}) = P(0)$, which follows from the proof of Theorem 1.31 below. □

Proof of Theorem 1.31. Let

$$L_\epsilon \triangleq \min_{\mathbf{a}\in\mathcal{A}} \left\{ g(\mathbf{a}) \;\middle|\; \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_\epsilon,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \leq c \right\}.$$

Then, by the proof of Theorem 1.29, for all $\mathbf{a} \in \mathcal{A}$,

$$\left| \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h_\epsilon,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) - \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \right| \leq \epsilon M,$$

where $M \triangleq 2\sup_{\mathbf{a}\in\mathcal{A},i\in[m]} |u(f(\mathbf{a},\mathbf{x}_i))| < \infty$, which follows from the continuity of $u(f(\mathbf{a},\mathbf{x}_i))$ in $\mathbf{a}$ and the compactness of $\mathcal{A}$. Hence, we conclude that $P(\epsilon) \leq L_\epsilon \leq P(0)$, where $P(0)$ is the optimal objective value of (P-constraint), and

$$P(\epsilon) \triangleq \min_{\mathbf{a}\in\mathcal{A}} \left\{ g(\mathbf{a}) \;\middle|\; \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \leq c + \epsilon M \right\}.$$

We will now show that $\lim_{\epsilon\downarrow 0} P(\epsilon) = P(0)$, hence concluding the proof. To show this, we use Berge's maximum theorem (Berge, 1963), for which we have to show that the set-valued function

$$\epsilon \mapsto G(\epsilon) \triangleq \left\{ \mathbf{a} \in \mathcal{A} : \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \leq c + \epsilon M \right\}$$

is a compact-valued, continuous correspondence at $\epsilon = 0$. We first examine compactness, which entails that for each $\epsilon \geq 0$, the set $G(\epsilon)$ must be compact. We note that since $\mathcal{A}$ is compact, we have that $G(\epsilon)$ is bounded for all $\epsilon \geq 0$. Hence, we only need to show that $G(\epsilon)$ is closed. This holds if $\mathbf{a} \mapsto \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X}))$ is continuous, which can be proven as follows: by Lemma 1.B.2, we have that $(\mathbf{q},\mathbf{a}) \mapsto \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X}))$ is jointly continuous in $(\mathbf{q},\mathbf{a})$. Since the set $\mathcal{D}_\phi(\mathbf{p},r)$ is compact and independent of $\mathbf{a}$, Berge's maximum theorem implies that $\mathbf{a} \mapsto \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X}))$ is continuous. Hence, $G(\epsilon)$ is compact.

We now show that $G(\epsilon)$ is lower-and upper-hemicontinuous at $\epsilon = 0$. For upper-hemicontinuity, we must show that $G(0)$ is non-empty, which holds due to Assumption 1.5, and that, for any $\epsilon_j \to 0$ and any sequence $\mathbf{a}_j \in G(\epsilon_j)$, there exists a convergent subsequence $\mathbf{a}_{j_k}$ such that $\mathbf{a}_{j_k} \to \mathbf{a}_0 \in G(0)$. Note that due to the compactness of $\mathcal{A}$, there is indeed a subsequence $\mathbf{a}_{j_k} \to \mathbf{a}_0 \in \mathcal{A}$. It remains to show that we have $\mathbf{a}_0 \in G(0)$. This follows from continuity: $\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}_0,\mathbf{X})) = \lim_{k\to\infty} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}_{j_k},\mathbf{X})) \leq \lim_{k\to\infty} c + \epsilon_{j_k} M = c$.

Finally, for lower-hemicontinuity at $\epsilon = 0$, we have to show that, for every $\mathbf{a}_0 \in G(0)$ and every sequence $\epsilon_j \to 0$, there exist a $J \geq 1$ and a sequence $\mathbf{a}_j$ such that, for all

$j \geq J$, we have $\mathbf{a}_j \in G(\epsilon_j)$ and $\mathbf{a}_j \to \mathbf{a}_0$. However, since $\epsilon_j > 0$ for all j, we can take the sequence $\mathbf{a}_j = \mathbf{a}_0 \in G(\epsilon_j)$. Hence, we may apply Berge's maximum theorem to conclude that $\lim_{\epsilon \downarrow 0} P(\epsilon) = P(0)$, which concludes the proof. □

Lemma 1.B.2. *Let $u(f(\mathbf{a}, \mathbf{x}_i))$ be continuous in $\mathbf{a}$ for all $i \in [m]$ and let h be a continuous distortion function. Then, the rank-dependent evaluation function*

$$(\mathbf{q}, \mathbf{a}) \mapsto \rho_{u,h,\mathbf{q}}(f(\mathbf{a}, \mathbf{X})),$$

is jointly continuous in $(\mathbf{q}, \mathbf{a})$.

Proof. For any $\mathbf{a} \in \mathcal{A}$, we denote the indices $1(\mathbf{a}), \ldots, m(\mathbf{a})$ such that

$$-u(f(\mathbf{a}, \mathbf{x}_{1(\mathbf{a})})) \leq \cdots \leq -u(f(\mathbf{a}, \mathbf{x}_{m(\mathbf{a})})),$$

and we define

$$\Delta u(f(\mathbf{a}, \mathbf{x}_{1(\mathbf{a})})) \triangleq -u(f(\mathbf{a}, \mathbf{x}_{1(\mathbf{a})})),\ \Delta u(f(\mathbf{a}, \mathbf{x}_{i(\mathbf{a})})) \triangleq u(f(\mathbf{a}, \mathbf{x}_{(i-1)(\mathbf{a})})) - u(f(\mathbf{a}, \mathbf{x}_{i(\mathbf{a})})),$$

for $i = 2, \ldots, m$. Denote the index set

$$\mathcal{I}_+(\mathbf{a}_0) = \{i \in \{2, \ldots, m\} : \Delta u(f(\mathbf{a}_0, \mathbf{x}_{i(\mathbf{a}_0)})) > 0\}.$$

We enumerate the elements in $\mathcal{I}_+(\mathbf{a}_0)$ as $i_1 < \cdots < i_K$, for $K = |\mathcal{I}_+(\mathbf{a}_0)| \leq m - 1$. Set $i_0 = 1$ and $i_{K+1} = m+1$. Then, the index set $[m]$ is partitioned into $K+1$ disjoint classes $\Pi_j \triangleq \{i(\mathbf{a}_0) : i_{j-1} \leq i < i_j\}$, with $j \in [K+1]$, that have the following two properties: (i) If $k, l \in [m]$ are indices of adjacent classes, e.g., $k \in \Pi_j$ and $l \in \Pi_{j+1}$ for some j, then $u(f(\mathbf{a}_0, \mathbf{x}_k)) - u(f(\mathbf{a}_0, \mathbf{x}_l)) = \Delta u(f(\mathbf{a}_0, \mathbf{x}_{i_{j+1}(\mathbf{a}_0)})) > 0$. (ii) If k, l belong to the same class Π_j, we have $u(f(\mathbf{a}_0, \mathbf{x}_k)) - u(f(\mathbf{a}_0, \mathbf{x}_l)) = 0$. Continuity of $u(f(\mathbf{a}, \mathbf{x}))$ in $\mathbf{a}_0$ implies that for all $\mathbf{a}$ sufficiently close to $\mathbf{a}_0$, the ranking indices $1(\mathbf{a}), \ldots, m(\mathbf{a})$ can also be partitioned such that $\{i(\mathbf{a}) : i_{j-1} \leq i < i_j\} = \{i(\mathbf{a}_0) : i_{j-1} \leq i < i_j\}$, for all $j \in [K+1]$. In particular, this means that within each class, the ranking indices of $\mathbf{a}$ constitute a permutation of that of $\mathbf{a}_0$.

Hence, we have

$$
\begin{aligned}
&\rho_{u,h,\mathbf{q}}(f(\mathbf{a},\mathbf{X})) \\
&= \sum_{i=1}^{m} h\left(\sum_{k=i}^{m} q_{k(\mathbf{a})}\right) \Delta u(f(\mathbf{a},\mathbf{x}_{i(\mathbf{a})})) = \sum_{j=1}^{K+1} \sum_{i=i_{j-1}}^{i_j-1} h\left(\sum_{k=i}^{m} q_{k(\mathbf{a})}\right) \Delta u(f(\mathbf{a},\mathbf{x}_{i(\mathbf{a})})) \\
&= \sum_{j=1}^{K+1} h\left(\sum_{k=i_{j-1}}^{m} q_{k(\mathbf{a})}\right) \Delta u(f(\mathbf{a},\mathbf{x}_{i_{j-1}(\mathbf{a})})) + \sum_{j=1}^{K+1} \sum_{i=i_{j-1}+1}^{i_j-1} h\left(\sum_{k=i}^{m} q_{k(\mathbf{a})}\right) \Delta u(f(\mathbf{a},\mathbf{x}_{i(\mathbf{a})})) \\
&\overset{(\mathbf{q},\mathbf{a})\to(\mathbf{q}_0,\mathbf{a}_0)}{=} \sum_{j=1}^{K+1} h\left(\sum_{k=i_{j-1}}^{m} q_{0,k(\mathbf{a}_0)}\right) \Delta u(f(\mathbf{a}_0,\mathbf{x}_{i_{j-1}(\mathbf{a}_0)})) + 0 \\
&= \rho_{u,h,\mathbf{q}_0}(f(\mathbf{a}_0,\mathbf{X})).
\end{aligned}
$$

This completes the proof. □

1.C Derivations of the Conjugate Functions and Their Epigraphs

In this appendix, we provide detailed derivations of the explicit conjugate functions, as well as the epigraphs of $\lambda(-h)^*(\frac{-\nu}{\lambda}) \le z, \lambda > 0$, and $\gamma\phi^*(\frac{s}{\gamma}) \le t, \gamma > 0$, for the collection of canonical examples in Tables 1.4–1.5; we only provide the details in case the derivations are considered to be non-trivial.

1.C.1 Distortion Functions

- For $h(p) = \begin{cases} (1+r)p & p < 1/2 \\ (1-r)p + r & p \ge 1/2 \end{cases}$ and $0 < r < 1$, we examine the conjugate

$$(-h)^*(y) = \max\{\sup_{t\in[0,\frac{1}{2})} (y + (1+r))t, \sup_{t\ge\frac{1}{2}} (y + (1-r))t + r\}.$$

  We have $(-h)^*(y) = \infty$ for $y > -(1-r)$. For $-(1+r) \le y \le -(1-r)$, we have $(-h)^*(y) = \max\{1/2(y + (1+r)), 1/2(y + (1-r)) + r\} = 1/2(y + (1+r))$. For $y < -(1+r)$, $(-h)^*(y) = \max\{0, 1/2(y + (1-r)) + r\} = 0$. Therefore, $(-h)^*(y) = \max\{(y + 1 + r)/2, 0\}$, for $y \le -(1-r)$.

- Let $h(p) = (1+r)p - rp^2$, $0 < r < 1$. For $y \le 0$, we have

$$(-h)^*(y) = \sup_{t\in[0,1]} \{ty + (1+r)t - rt^2\} = \sup_{t\in[0,1]} \{(y + (1+r))t - rt^2\}.$$

Differentiating the objective w.r.t. t yields

$$y + (1 + r) - 2rt,$$

which is non-negative for $t \leq \frac{y+1+r}{2r}$ and negative otherwise. If $\frac{y+1+r}{2r} < 0$, then the derivative is always negative, hence the maximum is obtained at $t = 0$ and thus we have $(-h)^*(y) = 0$. If $\frac{y+1+r}{2r} > 1$, then the derivative is always positive, hence $(-h)^*(y) = y + 1$. If $\frac{y+1+r}{2r} \in [0, 1]$, then the maximum is attained at $t = \frac{y+1+r}{2r}$. Hence, for $y \leq 0$, we have

$$(-h)^*(y) = \begin{cases} \frac{1}{4r} \max\{y + 1 + r, 0\}^2 & y + 1 \leq r \\ y + 1 & y + 1 > r. \end{cases}$$

The epigraph of $(-h)^*$ can be represented by

$$\begin{cases} t = t_1 + t_2 + r \\ y + 1 - r = y_1 + y_2 \\ t_1 \geq \frac{1}{4r} \max\{y_1 + 2r, 0\}^2 - r, \ t_2 \geq y_2 \\ y_1 \leq 0, y_2 \geq 0. \end{cases}$$

Indeed, let $(y, t) \in \text{Epi}((-h)^*)$. If $y + 1 - r \leq 0$, then we can choose $y_1 = y + 1 - r$, $t_1 = t - r$, $y_2 = t_2 = 0$. If $y + 1 - r \geq 0$, then choose $y_1 = t_1 = 0$, $y_2 = y + 1 - r$, $t_2 = t - r$.

Conversely, let $(y, t, y_1, y_2, t_1, t_2)$ satisfy the above constraints. Define

$$f(z) = \begin{cases} \frac{1}{4r} \max\{z + 2r, 0\}^2 - r & z \leq 0 \\ z & z \geq 0. \end{cases}$$

Then, $(-h)^*(y) = f(y+1-r)+r$. Hence, $(-h)^*(y) \leq t \Leftrightarrow f(y+1-r) \leq t-r$. Since $y_1 \leq 0$, $y_2 \geq 0$, we have that $y_1 \leq y_1 + y_2 \leq y_2$. Since f is convex and $f(0) = 0$, we have

$$\frac{f(y_1 + y_2) - f(y_1)}{y_2} \leq \frac{f(y_2) - f(0)}{y_2},$$

and thus $f(y_1 + y_2) \leq f(y_1) + f(y_2)$. Therefore, $(-h)^*(y) = f(y + 1 - r) + r = f(y_1 + y_2) \leq f(y_1) + f(y_2) + r \leq t_1 + t_2 + r = t$.

The epigraph of $\lambda(-h)^*(\frac{-\nu}{\lambda}) \leq z$ is then given by

$$\begin{cases} z = z_1 + z_2 \\ -\nu + \lambda(1-r) = \xi_1 + \xi_2 \\ (z_1 + \lambda) \geq \sqrt{\frac{w^2}{r} + (z_1 - \lambda)^2}, \ z_2 \geq \xi_2 \\ \xi_1 + 2r \leq w \\ \xi_1 \leq 0, \xi_2, w, \nu \geq 0. \end{cases}$$

- For $h(p) = \begin{cases} 1-(1-p)^n & 0 \leq p < 1 \\ 1 & p \geq 1 \end{cases}$, we will derive a tractable reformulation of the epigraph $(-h)^*$ using duality. We have

$$\text{Epi}(-h) = \{(p,t) \in \mathbb{R}_{\geq 0} \times \mathbb{R} : \exists\ (u_1, u_2) \in \mathbb{R}^2_{\geq 0} : u_1 - 1 \leq t, u_2 \leq u_1^{1/n}, 1 - p \leq u_2\}.$$

Indeed, let $(p,t) \in \text{Epi}(-h)$. If $0 \leq p < 1$, then $(1-p)^n - 1 \leq t$. We can choose $u_2 = 1 - p \geq 0$ and $u_1 = u_2^n \geq 0$. If $p \geq 1$, then $-1 \leq t$. We can choose $u_1 = u_2 = 0$. Conversely, let (p, t, u_1, u_2) satisfy the above constraints. If $p \geq 1$, we have $t \geq -1 + u_1 \geq -1$. Thus, $(p,t) \in \text{Epi}(-h)$. If $0 \leq p < 1$, then $(1-p)^n - 1 \leq u_2^n - 1 \leq u_1 - 1 \leq t$. Hence, we also have $(p,t) \in \text{Epi}(-h)$.

Consider the epigraph of $(-h)^*$. We have

$$\text{Epi}((-h)^*) = \{(y,s) : yp - (-h)(p) \leq s, \forall p \geq 0\} = \{(y,s) : yp - t \leq s, \forall (p,t) \in \text{Epi}(-h)\}.$$

Therefore, we have that $(y,s) \in \text{Epi}((-h)^*)$ if and only if the optimization problem

$$\min_{p,u_1,u_2 \geq 0, t \in \mathbb{R}} \{-yp + t | u_1 - 1 \leq t, u_2 \leq u_1^{1/n}, 1 - p \leq u_2\}$$

is bounded below by $-s$. Since this is a convex problem with a point $p = 1, u_2 = 0, u_1 = 1, t = 0$ that satisfies Slater's condition, we may apply the duality theorem and obtain that this optimization problem is equal to

$$\max_{\xi_1,\xi_2,\xi_3 \geq 0} \inf_{p,u_1,u_2 \geq 0, t \in \mathbb{R}} -yp + t + \xi_1(u_1 - 1 - t) + \xi_2(u_2 - u_1^{1/n}) + \xi_3(1 - p - u_2),$$

which, after some rewriting is equal to

$$\max_{\xi_2,\xi_3 \geq 0} \{-1 + \xi_3 + (n^{-\frac{n}{n-1}} - n^{-\frac{1}{n-1}})\xi_2^{\frac{n}{n-1}} \mid y + \xi_3 \leq 0, \xi_2 \geq \xi_3\}.$$

Therefore, we have

$$\begin{aligned}
&\mathrm{Epi}((-h)^*)\\
&= \{(y,s) : \exists\ \xi_2,\xi_3 \geq 0 : 1-\xi_3 + (n^{-\frac{1}{n-1}} - n^{-\frac{n}{n-1}})\xi_2^{\frac{n}{n-1}} \leq s, y+\xi_3 \leq 0, \xi_2 \geq \xi_3\}\\
&= \{(y,s) : \exists\ \xi_2,\xi_3,\xi_4 \geq 0 : 1-\xi_3 + (n^{-\frac{1}{n-1}} - n^{-\frac{n}{n-1}})\xi_4 \leq s, \xi_2 \leq \xi_4^{\frac{n-1}{n}} \cdot 1^{1-\frac{n-1}{n}},\\
&\qquad y+\xi_3 \leq 0, \xi_2 \geq \xi_3\}.
\end{aligned}$$

This gives the reformulation of the epigraph of the perspective $\lambda(-h)^*(\frac{-\nu}{\lambda}) \leq z$ as

$$\lambda(-h)^*(\frac{-\nu}{\lambda}) \leq z \Leftrightarrow \begin{cases} \lambda - \xi_3 + (n^{-\frac{1}{n-1}} - n^{-\frac{n}{n-1}})\xi_4 \leq z\\ \xi_2 \leq \xi_4^{\frac{n-1}{n}} \cdot \lambda^{1-\frac{n-1}{n}}\\ -\nu + \xi_3 \leq 0\\ \xi_2 \geq \xi_3\\ \xi_2,\xi_3,\xi_4 \geq 0. \end{cases}$$

- For $h(p) = \begin{cases} (1-(1-p)^n)^{1/n} & 0 \leq p < 1\\ 1 & p \geq 1 \end{cases}$, we will derive a tractable reformulation of the epigraph $(-h)^*$ using duality. We have

$$\begin{aligned}
\mathrm{Epi}(-h) = \{&(p,t) \in \mathbb{R}_{\geq 0} \times \mathbb{R} : \exists\ u_1,u_2 \geq 0 : u_1 \geq -t, u_2^{1/n} \geq u_1, (1-u_2)^{1/n} \geq 1-p\\
&, u_2 \leq 1\}.
\end{aligned}$$

Indeed, if $(p,t) \in \mathrm{Epi}(-h)$, then for $0 \leq p < 1$, we have $-(1-(1-p)^n)^{1/n} \leq t$. Choose $u_2 = (1-(1-p)^n)$ and $u_1 = u_2^{1/n}$ gives the right inclusion. If $p \geq 1$, then choose $u_2 = 1 = u_1$, which also gives the right inclusion. Conversely, let (p,t,u_1,u_2) satisfy the above constraints. If $0 \leq p < 1$, then, by construction, $(p,t) \in \mathrm{Epi}(-h)$. If $p \geq 1$, we have $u_1 \leq u_2^{1/n} \leq 1$. Hence, $t \geq -u_1 \geq -1$, thus $(p,t) \in \mathrm{Epi}(-h)$.

Consider the epigraph of $(-h)^*$. Again, we have

$$\mathrm{Epi}((-h)^*) = \{(y,s) : yp - t \leq s, \forall (p,t) \in \mathrm{Epi}(-h)\}.$$

Therefore, we have that $(y,s) \in \mathrm{Epi}((-h)^*)$ if and only if the optimization problem

$$\min_{p,u_1,\geq 0, u_2 \in [0,1], t \in \mathbb{R}} \{-yp + t | u_1 \geq -t, u_2^{1/n} \geq u_1, (1-u_2)^{1/n} \geq 1-p\}$$

is bounded below by $-s$. Since this is a convex problem with a point $p = 2, u_2 = 1, u_1 = 1/2, t = 1$ that satisfies Slater's condition, we may apply the duality theorem

and obtain that this optimization problem is equal to

$$\begin{aligned}&\max_{\xi_1,\xi_2,\xi_3\geq 0}\{\xi_3+\inf_{p,u_1\geq 0,u_2\in[0,1],t\in\mathbb{R}} -(y+\xi_3)p+(1-\xi_1)t+(\xi_2-\xi_1)u_1-\xi_2 u_2^{1/n}\\&\quad-\xi_3(1-u_2)^{1/n}\}\\&=\max_{\xi_2,\xi_3\geq 0}\{\xi_3+\inf_{u_2\in[0,1]} -\xi_2 u_2^{1/n}-\xi_3(1-u_2)^{1/n}|\xi_1=1, y+\xi_3\leq 0,\xi_2\geq\xi_1\}.\end{aligned}$$

The convex function $-\xi_2 u_2^{1/n}-\xi_3(1-u_2)^{1/n}$ has derivative

$$\frac{\mathrm{d}}{\mathrm{d}u_2}=\frac{1}{n}(\xi_3(1-u_2)^{\frac{1-n}{n}}-\xi_2 u_2^{\frac{1-n}{n}}),$$

which has a root at

$$\begin{aligned}\frac{\xi_2}{u_2^{\frac{n-1}{n}}}&=\frac{\xi_3}{(1-u_2)^{\frac{n-1}{n}}}\\ \xi_2(1-u_2)^{\frac{n-1}{n}}&=\xi_3 u_2^{\frac{n-1}{n}}\\ u_2&=\frac{\xi_2^{\frac{n}{n-1}}}{\xi_2^{\frac{n}{n-1}}+\xi_3^{\frac{n}{n-1}}}\in[0,1].\end{aligned}$$

Since we are examining a convex function, this is where the minimum is attained. Hence, we have

$$\begin{aligned}\inf_{u_2\in[0,1]} -\xi_2 u_2^{1/n}-\xi_3(1-u_2)^{1/n}&=-\xi_2\cdot\left(\frac{\xi_2^{\frac{n}{n-1}}}{\xi_2^{\frac{n}{n-1}}+\xi_3^{\frac{n}{n-1}}}\right)^{1/n}-\xi_3\cdot\left(\frac{\xi_3^{\frac{n}{n-1}}}{\xi_2^{\frac{n}{n-1}}+\xi_3^{\frac{n}{n-1}}}\right)^{1/n}\\&=-\frac{\xi_2^{\frac{n}{n-1}}+\xi_3^{\frac{n}{n-1}}}{(\xi_2^{\frac{n}{n-1}}+\xi_3^{\frac{n}{n-1}})^{1/n}}=-(\xi_2^{\frac{n}{n-1}}+\xi_3^{\frac{n}{n-1}})^{\frac{n-1}{n}}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Epi}((-h)^*)&=\{(y,s):\exists\ \xi_2,\xi_3\geq 0: y+\xi_3\leq 0,\xi_2\geq 1,-\xi_3+(\xi_2^{\frac{n}{n-1}}+\xi_3^{\frac{n}{n-1}})^{\frac{n-1}{n}}\leq s\}\\&=\{(y,s):\exists\ \xi_2,\xi_3\geq 0: y+\xi_3\leq 0,\xi_2\geq 1,-\xi_3+\|(\xi_2,\xi_3)\|_{\frac{n}{n-1}}\leq s\},\end{aligned}$$

where $\|.\|_p$ is the p-norm for $p\geq 1$. Thus, the reformulation of the epigraph of the perspective $\lambda(-h)^*(\frac{-\nu}{\lambda})\leq z$ is given by

$$\lambda(-h)^*(\frac{-\nu}{\lambda})\leq z\Leftrightarrow -\nu+\xi_3\leq 0,\ \xi_2\geq\lambda,\ -\xi_3+\|(\xi_2,\xi_3)\|_{\frac{n}{n-1}}\leq z,\ \xi_2,\xi_3\geq 0.$$

1.C.2 Divergence Functions

- For $\phi^*(y) = e^y - 1$, we have

$$\gamma(e^{\frac{s}{\gamma}} - 1) \leq t \Leftrightarrow w - \gamma \leq t,\ e^{\frac{s}{\gamma}} \leq \frac{w}{\gamma} \Leftrightarrow w - \gamma \leq t, \gamma \log(\frac{\gamma}{w}) + s \leq 0.$$

  Note that $\gamma \log(\frac{\gamma}{w})$ is the relative entropy.

- For $\phi^*(y) = 2 - 2\sqrt{1-y}$, $y < 1$, we have

$$\gamma(2 - 2\sqrt{1 - \frac{s}{\gamma}}) \leq t, s < \gamma \Leftrightarrow 2\gamma - 2\gamma\sqrt{\frac{\gamma - s}{\gamma}} \leq t, s < \gamma$$

$$\Leftrightarrow 2\gamma - 2\gamma\sqrt{\frac{v}{\gamma}} \leq t, v = \gamma - s, s < \gamma \Leftrightarrow 2\gamma - 2\sqrt{\gamma v} \leq t, v = \gamma - s, v > 0$$

$$\Leftrightarrow 2\gamma - 2w \leq t, w^2 \leq \gamma v, w \geq 0, v = \gamma - s, v > 0$$

$$\Leftrightarrow 2\gamma - 2w \leq t, \sqrt{w^2 + \frac{1}{4}(\gamma - v)^2} \leq \frac{1}{2}(\gamma + v),$$

$$w \geq 0, v = \gamma - s, v > 0,$$

  where in the last equivalence we used the equality $xy = \frac{1}{4}((x+y)^2 - (x-y)^2)$.

- For $\phi^*(y) = \frac{y}{1-y} = -1 + \frac{1}{1-y}, y < 1$, we have

$$-1 + \frac{1}{1-y} \leq t, y < 1 \Leftrightarrow -1 + v \leq t, \frac{1}{w} \leq v, w = 1 - y, w > 0$$

$$\Leftrightarrow -1 + v \leq t, \sqrt{1 + \frac{1}{4}(v - w)^2} \leq \frac{1}{2}(v + w), w = 1 - y, w > 0.$$

  The epigraph of the perspective function can then be obtained as

$$\gamma\phi^*(\frac{s}{\gamma}) \leq t \Leftrightarrow -\gamma + v \leq t, \sqrt{\gamma^2 + \frac{1}{4}(v - w)^2} \leq \frac{1}{2}(v + w), w = \gamma - s, w > 0.$$

- For $\phi^*(y) = y + (\theta - 1)(\frac{|y|}{\theta})^{\frac{\theta}{\theta-1}}$. We show that the epigraph can be represented by a power cone, which is $\mathcal{P}_3^{\alpha,1-\alpha} \triangleq \{x \in \mathbb{R}^3 : x_1^\alpha x_2^{1-\alpha} \geq |x_3|, x_1, x_2 \geq 0\}$, for $0 < \alpha < 1$. More on tractability of power cones can be found in Chares [2009]. We have that

$$y + (\theta - 1)(\frac{|y|}{\theta})^{\frac{\theta}{\theta-1}} \leq t \Leftrightarrow y + (\theta - 1)\theta^{\frac{\theta}{1-\theta}} w \leq t,\ |y|^{\frac{\theta}{\theta-1}} \leq w$$

$$\Leftrightarrow y + (\theta - 1)\theta^{\frac{\theta}{1-\theta}} w \leq t,\ |y| \leq w^{\frac{\theta}{\theta-1}} \cdot 1^{1-\frac{\theta}{\theta-1}},$$

is conic quadratic representable. Hence, so is the epigraph of the perspective

$$\gamma\phi^*(\frac{s}{\gamma}) \le t \Leftrightarrow s + (\theta-1)\theta^{\frac{\theta}{1-\theta}} w \le t,\ |s| \le w^{\frac{\theta}{\theta-1}} \cdot \gamma^{1-\frac{\theta}{\theta-1}}.$$

- Similarly, for $\phi^*(y) = \frac{1}{\theta}(1-y(1-\theta))^{\frac{\theta}{\theta-1}} - \frac{1}{\theta}, y < \frac{1}{1-\theta}$, we have that

$$\frac{1}{\theta}(1-y(1-\theta))^{\frac{\theta}{\theta-1}} - \frac{1}{\theta} \le t, y < \frac{1}{1-\theta} \Leftrightarrow |w| \le (t\theta+1)^{\frac{\theta-1}{\theta}} \cdot 1^{1-\frac{\theta-1}{\theta}}, w = 1-y(1-\theta), \\ y < \frac{1}{1-\theta},$$

is conic quadratic. Hence, so is the epigraph of the perspective

$$\gamma\phi^*(\frac{s}{\gamma}) \le t \Leftrightarrow |w| \le (t\theta+\gamma)^{\frac{\theta-1}{\theta}} \cdot \gamma^{1-\frac{\theta-1}{\theta}}, w = \gamma - s(1-\theta), s < \frac{\gamma}{1-\theta}.$$

1.D Robust Rank-Dependent Evaluation Using Penalization

This section investigates the comparison between

$$\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(\cdot), \tag{1.64}$$

and

$$\tilde{\rho}_{\text{rob}}(\cdot) \triangleq \sup_{\mathbf{q}\in\Delta_m} \rho_{u,h,\mathbf{q}}(\cdot) - \theta I_\phi(\mathbf{q},\mathbf{p}), \quad \theta > 0, \tag{1.65}$$

and associated optimization problems, where Δ_m denotes the set of all m-dimensional probability vectors.

Lemma 1.D.3. *Let $h : [0,1] \to [0,1]$ be a concave distortion function. Then, the robust risk measure $\tilde{\rho}_{rob}$ defined in* (1.65) *admits the dual representation*

$$\tilde{\rho}_{rob}(X) = \sup_{\bar{\mathbf{q}}\in\Delta_m} \mathbb{E}_{\bar{\mathbf{q}}}[-u(X)] - c(\bar{\mathbf{q}}),$$

with ambiguity index

$$c(\bar{\mathbf{q}}) \triangleq \inf_{\mathbf{q}\in\Delta_m} \theta I_\phi(\mathbf{q},\mathbf{p}) + \alpha(\bar{\mathbf{q}},\mathbf{q}), \qquad \alpha(\bar{\mathbf{q}},\mathbf{q}) \triangleq \begin{cases} 0, & \text{if } \bar{\mathbf{q}} \in M_h(\mathbf{q}); \\ \infty, & \text{else;} \end{cases}$$

where $M_h(\mathbf{q})$ is as defined in (1.15).

Proof. By Denneberg [1994], Proposition 10.3, we have the dual representation

$$\rho_{u,h,\mathbf{q}}(X) = \sup_{\bar{\mathbf{q}}\in\Delta_m} \mathbb{E}_{\bar{\mathbf{q}}}[-u(X)] - \alpha(\bar{\mathbf{q}}, \mathbf{q}).$$

Hence, we derive

$$\begin{aligned}
\tilde{\rho}_{\text{rob}}(X) &= \sup_{\mathbf{q}\in\Delta_m} \rho_{u,h,\mathbf{q}}(X) - \theta \sum_{i=1}^m p_i \phi\left(\frac{q_i}{p_i}\right) \\
&= \sup_{\mathbf{q}\in\Delta_m} \sup_{\bar{\mathbf{q}}\in\Delta_m} \mathbb{E}_{\bar{\mathbf{q}}}[-u(X)] - \alpha(\bar{\mathbf{q}}, \mathbf{q}) - \theta \sum_{i=1}^m p_i \phi\left(\frac{q_i}{p_i}\right) \\
&= \sup_{\bar{\mathbf{q}}\in\Delta_m} \sup_{\mathbf{q}\in\Delta_m} \mathbb{E}_{\bar{\mathbf{q}}}[-u(X)] - \alpha(\bar{\mathbf{q}}, \mathbf{q}) - \theta \sum_{i=1}^m p_i \phi\left(\frac{q_i}{p_i}\right) \\
&= \sup_{\bar{\mathbf{q}}\in\Delta_m} \mathbb{E}_{\bar{\mathbf{q}}}[-u(X)] - \inf_{\mathbf{q}\in\Delta_m} \left(\alpha(\bar{\mathbf{q}}, \mathbf{q}) + \theta \sum_{i=1}^m p_i \phi\left(\frac{q_i}{p_i}\right) \right) \\
&= \sup_{\bar{\mathbf{q}}\in\Delta_m} \mathbb{E}_{\bar{\mathbf{q}}}[-u(X)] - c(\bar{\mathbf{q}}).
\end{aligned}$$

□

We next show that a decision-maker who minimizes (1.64) obtains the same minimizer as when minimizing (1.65), for a specific θ.

Proposition 1.D.4. *Assume the existence of a minimizer* $\mathbf{a}^* \in \operatorname{argmin}_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}, \mathbf{X}))$. *Then for each* $r > 0$, *there exists a* θ^*, *such that*

$$\mathbf{a}^* \in \operatorname{argmin}_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\Delta_m} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}, \mathbf{X})) - \theta^* I_\phi(\mathbf{q}, \mathbf{p}).$$

Proof. As we have shown in Lemma 1.D.3, $\min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\Delta_m} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}, \mathbf{X})) - \theta I_\phi(\mathbf{q}, \mathbf{p})$ is equivalent to

$$\min_{\mathbf{a}\in\mathcal{A}} \left\{ \sup_{\substack{\mathbf{q}\geq\mathbf{0} \\ \mathbf{q}^T\mathbf{1}=1}} \sup_{\bar{\mathbf{q}}\in M_h(\mathbf{q})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)) - \theta \sum_{i=1}^m p_i \phi\left(\frac{q_i}{p_i}\right) \right\}, \tag{1.66}$$

for any fixed constant $\theta > 0$. Let $\mathbf{a}^*$ be a solution of $\min_{\mathbf{a}\in\mathcal{A}} \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \rho_{u,h,\mathbf{q}}(f(\mathbf{a}, \mathbf{X}))$. Then, strong duality implies the existence of a γ^* (depending on $\mathbf{a}^*$), such that

$$\begin{aligned}
&\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \sup_{\bar{\mathbf{q}}\in M_h(\mathbf{q})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}^*, \mathbf{x}_i)) \\
&= \gamma^* r + \sup_{\substack{\mathbf{q}\geq\mathbf{0} \\ \mathbf{q}^T\mathbf{1}=1}} \sup_{\bar{\mathbf{q}}\in M_h(\mathbf{q})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}^*, \mathbf{x}_i)) - \gamma^* \sum_{i=1}^m p_i \phi\left(\frac{q_i}{p_i}\right).
\end{aligned} \tag{1.67}$$

Since $\mathbf{a}^*$ is a minimizer, we also have

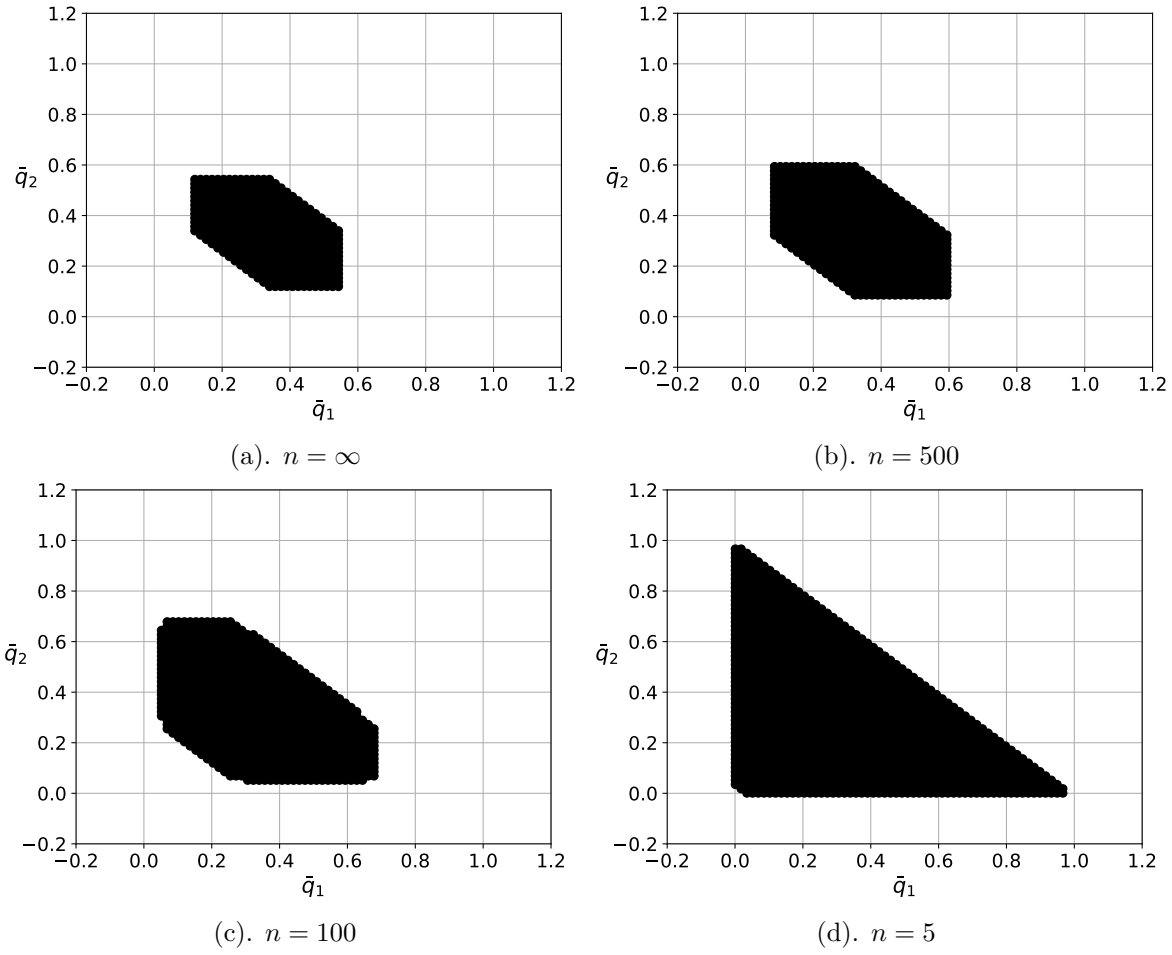
$$\begin{aligned}
&\sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \sup_{\bar{\mathbf{q}}\in M_h(\mathbf{q})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}^*,\mathbf{x}_i)) \\
&\leq \sup_{\mathbf{q}\in\mathcal{D}_\phi(\mathbf{p},r)} \sup_{\bar{\mathbf{q}}\in M_h(\mathbf{q})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)) \\
&= \inf_{\gamma\geq 0}\left\{\gamma r + \sup_{\substack{\mathbf{q}\geq\mathbf{0}\\ \mathbf{q}^T\mathbf{1}=1}} \sup_{\bar{\mathbf{q}}\in M_h(\mathbf{q})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)) - \gamma\sum_{i=1}^m p_i\phi\left(\frac{q_i}{p_i}\right)\right\} \\
&\leq \gamma^* r + \sup_{\substack{\mathbf{q}\geq\mathbf{0}\\ \mathbf{q}^T\mathbf{1}=1}} \sup_{\bar{\mathbf{q}}\in M_h(\mathbf{q})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)) - \gamma^*\sum_{i=1}^m p_i\phi\left(\frac{q_i}{p_i}\right).
\end{aligned}$$

Bringing $\gamma^* r$ to the left-hand side of the inequality, we obtain from (1.67) that

$$\begin{aligned}
&\sup_{\substack{\mathbf{q}\geq\mathbf{0}\\ \mathbf{q}^T\mathbf{1}=1}} \sup_{\bar{\mathbf{q}}\in M_h(\mathbf{q})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a}^*,\mathbf{x}_i)) - \gamma^*\sum_{i=1}^m p_i\phi\left(\frac{q_i}{p_i}\right) \\
&\leq \sup_{\substack{\mathbf{q}\geq\mathbf{0}\\ \mathbf{q}^T\mathbf{1}=1}} \sup_{\bar{\mathbf{q}}\in M_h(\mathbf{q})} -\sum_{i=1}^m \bar{q}_i u(f(\mathbf{a},\mathbf{x}_i)) - \gamma^*\sum_{i=1}^m p_i\phi\left(\frac{q_i}{p_i}\right),
\end{aligned}$$

for any $\mathbf{a}\in\mathcal{A}$. Hence, $\mathbf{a}^*$ is also a solution of (1.66), where $\theta^* = \gamma^*$. □

1.E A Visualization of the Various Shapes of the Uncertainty Set $\mathcal{U}_{\phi,h}(\mathbf{p})$



(a). $n = \infty$ (b). $n = 500$

(c). $n = 100$ (d). $n = 5$

Figure 1.5: Projections of the uncertainty set $\mathcal{U}_{\phi,h}(\mathbf{p})$ in (1.13) on the coordinates $(\bar{q}_1, \bar{q}_2)$, for $\mathbf{p} = (1/3, 1/3, 1/3)$ and $r = \frac{1}{n}\chi^2_{0.95,2}$, plotted for a range of values of the sample size n. We choose the modified chi-squared divergence function $\phi(t) = (t-1)^2$ and the second dual moment distortion function $h(p) = 1 - (1-p)^2$. As n approaches 0, we observe that the uncertainty set grows and the projection eventually approaches the entire probability simplex in $\mathbb{R}^2$, the case in which the decision-maker is completely ambiguous w.r.t. $\mathbf{p}$.

1.F The Optimistic Dual Counterpart

Beck and Ben-Tal [2009] and Gorissen et al. [2014] have shown that a robust minimization problem with a compact convex uncertainty set can also be reformulated by considering the optimistic dual counterpart, obtained by maximizing the dual of its uncertain problem over all uncertain variables. In this section, we derive and reformulate the optimistic dual

counterpart of the robust problem

$$\min_{\mathbf{a}\in\mathcal{A},c\in\mathcal{C}}\left\{\mathbf{a}^T\mathbf{d}+\zeta\cdot c\;\middle|\;\sup_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_{\phi,h}(\mathbf{p})}-\sum_{i=1}^m\bar{q}_i u(f(\mathbf{a},\mathbf{x}_i))\le c\right\},\tag{1.68}$$

where

$$\mathcal{U}_{\phi,h}(\mathbf{p})=\left\{(\mathbf{q},\bar{\mathbf{q}})\in\mathbb{R}^{2m}\;\middle|\;\begin{array}{l}\sum_{i=1}^m q_i=\sum_{i=1}^m\bar{q}_i=1\\ \sum_{i=1}^m p_i\phi\left(\frac{q_i}{p_i}\right)\le r\\ \sum_{i\in J}\bar{q}_i\le h\left(\sum_{i\in J}q_i\right),\ \forall J\subset[m]\\ q_i,\bar{q}_i\ge 0,\ \forall i\in[m]\end{array}\right\}.$$

Here, we have either $(\mathcal{C}=\mathbb{R},\mathbf{d}=\mathbf{0},\zeta=1)$ or $(\mathcal{C}=\{c_0\},\mathbf{d}\neq\mathbf{0},\zeta=0)$, where $\mathbf{d}\in\mathbb{R}^{n_a},c_0\in\mathbb{R}$. Note that $\mathcal{U}_{\phi,h}(\mathbf{p})$ is a compact set since it is bounded and closed due to Assumption 1.4 that the functions $\phi,-h$ are lower-semicontinuous convex functions.

To derive the optimistic dual counterpart of (1.68), we first consider its uncertain problem, which for a given $(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_{\phi,h}(\mathbf{p})$ is defined as

$$\begin{aligned}&\min_{\mathbf{a}\in\mathcal{A},c\in\mathcal{C}}\left\{\mathbf{a}^T\mathbf{d}+\zeta\cdot c\;\middle|\;-\sum_{i=1}^m\bar{q}_i u(f(\mathbf{a},\mathbf{x}_i))\le c\right\}\\ &=\min_{\mathbf{a}\in\mathbb{R}^I,c\in\mathcal{C}}\left\{\mathbf{a}^T\mathbf{d}+\zeta\cdot c\;\middle|\;-\sum_{i=1}^m\bar{q}_i u(f(\mathbf{a},\mathbf{x}_i))\le c, f_i(\mathbf{a})\le 0, i\in[L]\right\}.\end{aligned}\tag{1.69}$$

Here, recall that $\mathcal{A}$ is a set represented by convex inequalities, hence we can express it as $\mathcal{A}=\{\mathbf{a}\in\mathbb{R}^I|f_i(\mathbf{a})\le 0,i=1,\ldots,L\}$, for some convex functions f_i's. Assume that (1.69) satisfies Slater's condition and is bounded from below. Then, the dual of (1.69) is given by

$$\max_{y_0,\ldots,y_L\ge 0}\inf_{\mathbf{a}\in\mathbb{R}^I,c\in\mathcal{C}}\left\{\mathbf{a}^T\mathbf{d}+\zeta\cdot c-y_0\sum_{i=1}^m\bar{q}_i u(f(\mathbf{a},\mathbf{x}_i))-y_0c+\sum_{k=1}^L y_k f_k(\mathbf{a})\right\}.\tag{1.70}$$

The optimistic dual counterpart is defined by maximizing the dual over all uncertain variables in the uncertainty set, i.e.,

$$\sup_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_{\phi,h}(\mathbf{p})}\max_{y_0,\ldots,y_L\ge 0}\inf_{\mathbf{a}\in\mathbb{R}^I,c\in\mathcal{C}}\left\{\mathbf{a}^T\mathbf{d}+\zeta\cdot c-y_0\sum_{i=1}^m\bar{q}_i u(f(\mathbf{a},\mathbf{x}_i))-y_0c+\sum_{k=1}^L y_k f_k(\mathbf{a})\right\}.\tag{OD}$$

The following theorem states a reformulation of (OD).

Theorem 1.F.5. *The optimistic dual counterpart* (OD) *is equivalent to the following*

concave problem:

$$\begin{aligned}
\sup_{\substack{\mathbf{z},\bar{\mathbf{z}}\in\mathbb{R}^m_{\geq 0}\\ y_0,\ldots,y_L\geq 0\\ \boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_m,\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_L\in\mathbb{R}^I}} & -\sum_{i=1}^m \bar{z}_i(-u\circ f)^*\left(\frac{\boldsymbol{\lambda}_i}{y_0\bar{q}_i},\mathbf{x}_i\right)-\sum_{k=1}^L y_k(f_k)^*\left(\frac{\boldsymbol{\eta}_k}{y_k}\right)-y_0c_0\mathbb{1}_{\{\zeta=0\}}\\
\text{subject to } & \sum_{i=1}^m\boldsymbol{\lambda}_i=-\sum_{k=1}^L\boldsymbol{\eta}_k-\mathbf{d}\\
& \sum_{i=1}^m z_i=\sum_{i=1}^m\bar{z}_i=y_0\\
& \sum_{i\in I}\bar{z}_i-y_0h\left(\frac{\sum_{i\in I}z_i}{y_0}\right)\leq 0,\forall I\in 2^N_-,\\
& \sum_{i=1}^m y_0p_i\phi\left(\frac{z_i}{y_0p_i}\right)-y_0r\leq 0,
\end{aligned}\tag{1.71}$$

where $\mathbf{z}=y_0\mathbf{q}$ *and* $\bar{\mathbf{z}}=y_0\bar{\mathbf{q}}$*. If Slater's condition holds for problem* (1.71)*, then* (1.68) *is equal to* (1.71)*. Moreover, the KKT-vector of* (1.71) *corresponding to the dual equality constraints* $\sum_{i=1}^m\boldsymbol{\lambda}_i=-\sum_{k=1}^L\boldsymbol{\eta}_k-\mathbf{d}$ *gives the optimal solution of* (1.68)*.*

Proof. We have that

$$\begin{aligned}
&\inf_{\mathbf{a}\in\mathbb{R}^I,c\in\mathcal{C}}\left\{\mathbf{a}^T\mathbf{d}+\zeta\cdot c-y_0\sum_{i=1}^m\bar{q}_iu(f(\mathbf{a},\mathbf{x}_i))-y_0c+\sum_{k=1}^L y_kf_k(\mathbf{a})\right\}\\
&=\inf_{\mathbf{a}\in\mathbb{R}^I}\left\{\mathbf{a}^T\mathbf{d}-y_0\sum_{i=1}^m\bar{q}_iu(f(\mathbf{a},\mathbf{x}_i))+\sum_{k=1}^L y_kf_k(\mathbf{a})\right\}+\inf_{c\in\mathcal{C}}(\zeta-y_0)c\\
&=-\sup_{\mathbf{a}\in\mathbb{R}^I}\left\{-\mathbf{a}^T\mathbf{d}+y_0\sum_{i=1}^m\bar{q}_iu(f(\mathbf{a},\mathbf{x}_i))-\sum_{k=1}^L y_kf_k(\mathbf{a})\right\}+\inf_{c\in\mathcal{C}}(\zeta-y_0)c.
\end{aligned}$$

Note that $\mathcal{C}$ is either $\mathbb{R}$ (if $\zeta=1$) or $\{c_0\}$ for some $c_0\in\mathbb{R}$ (if $\zeta=0$). Hence, we have that

$$\inf_{c\in\mathcal{C}}(\zeta-y_0)c=\begin{cases}-y_0c_0 & \text{if } \mathcal{C}=c_0\\ 0 & \text{if } \mathcal{C}=\mathbb{R},\ y_0=1\\ -\infty & \text{else.}\end{cases}$$

We examine the supremum term. We have

$$
\begin{aligned}
&\sup_{\mathbf{a}\in\mathbb{R}^I}\left\{-\mathbf{a}^T\mathbf{d}+y_0\sum_{i=1}^m\bar{q}_i u(f(\mathbf{a},\mathbf{x}_i))-\sum_{k=1}^L y_k f_k(\mathbf{a})\right\}\\
&=\sup_{\substack{\mathbf{a},\mathbf{w}_1,\dots,\mathbf{w}_m,\\ \mathbf{v}_1,\dots,\mathbf{v}_L\in\mathbb{R}^I}}\left\{-\mathbf{a}^T\mathbf{d}+y_0\sum_{i=1}^m\bar{q}_i u(f(\mathbf{w}_i,\mathbf{x}_i))-\sum_{k=1}^L y_k f_k(\mathbf{v}_k)\;\middle|\;\mathbf{w}_i=\mathbf{a},\ \mathbf{v}_k=\mathbf{a}, i\in[m], k\in[L]\right\}\\
&=\inf_{\substack{\boldsymbol{\lambda}_1,\dots\boldsymbol{\lambda}_m,\\ \boldsymbol{\eta}_1,\dots,\boldsymbol{\eta}_L\in\mathbb{R}^I}}\sup_{\substack{\mathbf{a},\mathbf{w}_1,\dots,\mathbf{w}_m,\\ \mathbf{v}_1,\dots,\mathbf{v}_L\in\mathbb{R}^I}}\left\{-\mathbf{a}^T\mathbf{d}+\sum_{i=1}^m\boldsymbol{\lambda}_i^T(\mathbf{w}_i-\mathbf{a})+\sum_{k=1}^L\boldsymbol{\eta}_k^T(\mathbf{v}_k-\mathbf{a})+y_0\sum_{i=1}^m\bar{q}_i u(f(\mathbf{w}_i,\mathbf{x}_i))\right.\\
&\qquad\left.-\sum_{k=1}^L y_k f_k(\mathbf{v}_k)\right\}\\
&=\inf_{\substack{\boldsymbol{\lambda}_1,\dots\boldsymbol{\lambda}_m,\\ \boldsymbol{\eta}_1,\dots,\boldsymbol{\eta}_L\in\mathbb{R}^I}}\sup_{\substack{\mathbf{a},\mathbf{w}_1,\dots,\mathbf{w}_m,\\ \mathbf{v}_1,\dots,\mathbf{v}_L\in\mathbb{R}^I}}\left\{-\left(\sum_{i=1}^m\boldsymbol{\lambda}_i+\sum_{k=1}^L\boldsymbol{\eta}_k+\mathbf{d}\right)^T\mathbf{a}+\sum_{i=1}^m\boldsymbol{\lambda}_i^T\mathbf{w}_i-y_0\bar{q}_i(-u)(f(\mathbf{w}_i,\mathbf{x}_i))\right.\\
&\qquad\left.+\sum_{k=1}^L\boldsymbol{\eta}_k^T\mathbf{v}_k-y_k f_k(\mathbf{v}_k)\right\}\\
&=\inf_{\substack{\boldsymbol{\lambda}_1,\dots\boldsymbol{\lambda}_m,\\ \boldsymbol{\eta}_1,\dots,\boldsymbol{\eta}_L\in\mathbb{R}^I}}\left\{\sum_{i=1}^m y_0\bar{q}_i(-u\circ f)^*\left(\frac{\boldsymbol{\lambda}_i}{y_0\bar{q}_i},\mathbf{x}_i\right)+\sum_{k=1}^L y_k(f_k)^*\left(\frac{\boldsymbol{\eta}_k}{y_k}\right)\;\middle|\;\sum_{i=1}^m\boldsymbol{\lambda}_i=-\sum_{k=1}^L\boldsymbol{\eta}_k-\mathbf{d}\right\},
\end{aligned}
$$

where $(-u\circ f)$ denotes the composition of the two functions. A change of variables $\mathbf{z}=y_0\mathbf{q}$ and $\bar{\mathbf{z}}=y_0\bar{\mathbf{q}}$ and a multiplication of the constraints in $\mathcal{U}_{\phi,h}(\mathbf{p})$ by y_0 yields the desired statement. □

1.G Optimization of Rank-Dependent Models in the Constraint

We describe in detail the adaptation of Algorithm 1 to (P-constraint) in Algorithm 4.

1.H SOS2-Constraints Formulation

Let $u:[l_0,u_0]\to\mathbb{R}$ be a non-decreasing, piecewise-linear utility function defined on some interval $[l_0,u_0]$ that contains the image set $\{\mathbf{a}^T\mathbf{x}_i\mid\mathbf{a}\in\mathcal{A}, i\in[m]\}$. Let $\{t_j,u(t_j)\}_{j=1}^K$ be the support points of u, where $l_0=t_1<\cdots<t_K=u_0$. Then, the constraints in (1.34)

Algorithm 4 Cutting-Plane Method with RDU Constraint

1: Start with $\mathcal{U}_1 = \{(\mathbf{p}, \mathbf{p})\}$. Fix a tolerance parameter $\epsilon_{\text{tol}} > 0$.
2: At the j-th iteration, solve the following problem with the uncertainty set $\mathcal{U}_j$:

$$\min_{\mathbf{a}\in\mathcal{A}} \left\{ g(\mathbf{a}) \;\middle|\; \sup_{(\mathbf{q},\bar{\mathbf{q}})\in\mathcal{U}_j} \sum_{i=1}^{m} -\bar{q}_i u(f(\mathbf{a}, \mathbf{x}_i)) \le c \right\}. \tag{1.72}$$

3: Let $\mathbf{a}_j$ be the optimal solution of (1.20). Determine the ranking:

$$-u(f(\mathbf{a}_j, \mathbf{x}_{(1)})) \le \ldots \le -u(f(\mathbf{a}_j, \mathbf{x}_{(m)})).$$

Then, solve the optimization problem (1.19), which gives an optimal objective value v_j and a solution $(\mathbf{q}_j^*, \bar{\mathbf{q}}_j^*)$.
4: If $v_j - c \le \epsilon_{\text{tol}}$, then the solution is accepted and the process is terminated.
5: If not, set $\mathcal{U}_{j+1} = \mathcal{U}_j \cup \{(\mathbf{q}_j^*, \bar{\mathbf{q}}_j^*)\}$ and repeat steps 2–5.

can be formulated using the following bilinear and SOS2 constraints:

$$\begin{cases} \beta \cdot h(p^0) + \sum_{k=1}^{K_1} \nu_k b_k^{(1)} + \sum_{i=1}^{m}\sum_{k=1}^{K_1} \lambda_{ik} l_k^{(1)} p_i - \sum_{i=1}^{m} \bar{q}_i z_i \le c \\ -z_i - \beta - \sum_{k=1}^{K_1} \lambda_{ik} \le 0, \ \forall i \in [m] \\ \lambda_{ik} \le \nu_k, \ \forall i \in [m], \ \forall k \in [K_1] \\ \bar{q}_i \le l_k^{(2)} p_i + t_{ik}, \ \forall i \in [m], \ \forall k \in [K_2] \\ \sum_{i=1}^{m} t_{ik} \le b_k^{(2)}, \ \forall k \in [K_2] \\ \sum_{i=1}^{m} \bar{q}_i = \bar{h}(1 - p^0) \\ \mathbf{a}^T \mathbf{x}_i = \sum_{j=1}^{K} \tilde{\lambda}_{ij} t_j, \ \forall i \in [m] \\ z_i = \sum_{j=1}^{K} \tilde{\lambda}_{ij} u(t_j), \ \forall i \in [m] \\ \sum_{j=1}^{K} \tilde{\lambda}_{ij} = 1, \ \forall i \in [m] \\ \tilde{\lambda}_{ij} \ge 0, \ \text{SOS2}, \ \forall i \in [m], \ \forall j \in [K]. \end{cases} \tag{1.73}$$

1.I The Hit-and-Run Algorithm

In this appendix, we explain how the Hit-and-Run algorithm is applied to generate a uniform sample on the ϕ-divergence set. Let S be an open subset in $\mathbb{R}^d$. The general procedure of the Hit-and-Run algorithm is fairly simple. Start with an interior point $x_0 \in S$. Choose uniformly a random direction $u \in \partial\mathcal{D}$ where $\partial\mathcal{D} \triangleq \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. Draw uniformly a scalar $\lambda \in \{\lambda \in \mathbb{R} : x_0 + \lambda u \in S\}$ and then update x_0 with $x_0 + \lambda u$. It is shown by Bélisle et al. [1993] that for a convex set S, this sampling process will converge to a uniform sample on S.

To apply the Hit-and-Run algorithm to the ϕ-divergence uncertainty set, we first re-parametrize the set. Recall the definition of the ϕ-divergence set (1.8). Using the equality

$\mathbf{q}^T\mathbf{1} = 1$, it is clear that we can parametrize the ϕ-divergence set with the following set:

$$\tilde{\mathcal{D}}_\phi(\mathbf{q}|\mathbf{p}, r) \triangleq \left\{ \mathbf{q} \in \mathbb{R}^{m-1} | \mathbf{q} \geq \mathbf{0}, 1 - \sum_{i=1}^{m-1} q_i \geq 0, \sum_{i=1}^{m-1} p_i \phi\left(\frac{q_i}{p_i}\right) + p_m \phi\left(\frac{1 - \sum_{i=1}^{m-1} q_i}{p_m}\right) \leq r \right\}. \tag{1.74}$$

Since this is a set with convex inequalities, and we assume Slater's condition, its interior is an open convex set. Hence, we can apply the Hit-and-Run algorithm to obtain a uniform sample of the interior of $\tilde{\mathcal{D}}_\phi(\mathbf{q}|\mathbf{p}, r)$ and thus also the interior of $\mathcal{D}_\phi(\mathbf{p}, r)$ due to their one-to-one correspondence.

The precise procedure is the following.

- We start with an interior point $\mathbf{q}_k \in \text{int}\tilde{\mathcal{D}}_\phi(\mathbf{q}|\mathbf{p}, r)$, where we set $\mathbf{q}_0 = (p_1, \ldots, p_{m-1})$. Then, we draw a uniform element $\mathbf{u} \in \partial\mathcal{D}$ of dimension $d = m-1$. This can be done by drawing d standard normal i.i.d. samples $\mathbf{X} = (X_1, \ldots, X_d)$ and setting $\mathbf{u} = \frac{\mathbf{X}}{\|\mathbf{X}\|_2}$, where $\|.\|_2$ is the Euclidean 2-norm. Then, $\mathbf{u}$ follows a uniform distribution on $\partial\mathcal{D}$.

- Given $\mathbf{u}$ and $\mathbf{q}_k$, we determine the set $\Lambda \triangleq \{\lambda \in \mathbb{R} : \mathbf{q}_k + \lambda\mathbf{u} \in \tilde{\mathcal{D}}_\phi(\mathbf{q}|\mathbf{p}, r)\}$. Note that since $\tilde{\mathcal{D}}_\phi(\mathbf{q}|\mathbf{p}, r)$ is convex, bounded and closed, there exist $\lambda_{\min}, \lambda_{\max}$ such that $[\lambda_{\min}, \lambda_{\max}] = \Lambda$. We may find both values using a bisection search. Indeed, since $\mathbf{q}_k \in \text{int}\tilde{\mathcal{D}}_\phi(\mathbf{q}|\mathbf{p}, r)$, there exists a sufficiently small $\lambda_0 > 0$ such that $\lambda_0 \in \Lambda$. Take a sufficiently large $\lambda_1 > \lambda_0$ such that $\lambda_1 \notin \Lambda$, which also exists since Λ is bounded. Then, perform the bisection search on $[\lambda_0, \lambda_1]$ to find $\lambda_{\max}$. Do the same for $\lambda_{\min}$.

- We draw a random $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and update $\mathbf{q}_{k+1} = \mathbf{q}_k + \lambda\mathbf{u}$. We then repeat the process.

Chapter 2

Constructing Uncertainty Sets for Robust Risk Measures: A Composition of ϕ-Divergences Approach to Combat Tail Uncertainty

2.1 Introduction

Model misspecification, especially in the tails, is an inevitable issue that underlies the practice of financial and operational risk management. One way to safeguard against model uncertainty is distributionally robust optimization (DRO). The idea is to consider a set of plausible models, often referred to as the uncertainty/ambiguity set, and evaluate the worst case among all these models. DRO has its roots in classical robust optimization, pioneered by Ben-Tal et al. [2009], as well as in economics and decision theory, in the works of Gilboa and Schmeidler [1989], Maccheroni et al. [2006], and Hansen and Sargent [2008], which provide decision-theoretic foundations of the classical decision rule of Wald [1950]; see also Huber [1981].

An essential part of distributional robust optimization is to choose a proper uncertainty set. In data-driven optimization, DRO is often used to address sampling errors of empirical risk minimization, employing uncertainty sets that are calibrated to offer certain statistical guarantees. For this line of research, we refer the readers to works such as Ben-Tal et al. [2013], Esfahani and Kuhn [2018], Lam [2019], Duchi et al. [2021], van Parys et al. [2021]. In risk management, such as the calculation of capital requirements, one typically works with a pre-specified nominal model (e.g., an internal model proposed by the regulator) instead of the empirical distribution. In this case, the goal is to construct uncertainty sets that are sufficiently large — yet still plausible — to protect against model misspecification.

In much of the DRO literature, the respective optimization problem is studied in

a risk-neutral environment, where a random loss is evaluated under its expected value. In finance, insurance, and economics, it is more natural to model a decision maker's preferences using a risk measure that is nonlinear. This complicates the choice of a suitable uncertainty set, since in some cases the linearity assumption is crucial for deriving a tractable reformulation of the (primal) robust problem. Moreover, the sensitivity of a risk measure to the tails of a distribution also determines the suitability of the specification of an uncertainty set. For example, an evaluation of the entropic risk measure (Föllmer and Schied, 2016, Example 4.13) induced by expected exponential utility preferences is only finite for light-tailed distributions. This implies that any uncertainty set containing a heavy-tailed distribution is inadequate for the entropic risk measure. Therefore, the following question naturally arises: given a risk measure and a pre-specified nominal model, how should one properly specify an uncertainty set that is not too conservative (and not too restrictive), and that in addition admits a tractable reformulation?

In this paper, we provide an answer to the above question by proposing a simple, yet unifying framework to specify uncertainty sets for a broad class of risk measures, based on ϕ-divergences. Introduced by Csiszár [1975], ϕ-divergences are statistical measures for probability models that have received much attention in the DRO literature since the work of Ben-Tal et al. [2013]. Uncertainty sets that are defined as a ball around a nominal model, measured using a ϕ-divergence, have been extensively studied by Kruse et al. [2019, 2021] in the context of model risk assessment. Interestingly, ϕ-divergences are also strongly connected to risk measures. As shown by Ben-Tal et al. [1991], Ben-Tal and Teboulle [2007], the four major classes of risk measures: Expected utility, optimized certainty equivalents, shortfall risk measures (a.k.a. u-Mean certainty equivalents), and distortion risk measures, all admit a robust representation with a ϕ-divergence penalty function.

Our main contributions can be summarized as follows. By utilizing the pivotal role of ϕ-divergences in both robust optimization and risk theory, we characterize a wide range of distributionally robust risk measures in terms of two ϕ-divergences (ϕ_1, ϕ_2): a ϕ_1-divergence that specifies the risk measure through an *inner* robust representation, and a ϕ_2-divergence that controls an *outer* ambiguity set to address model uncertainty. We call this the *composition* approach, since the parametrization with two ϕ-divergences is formulated as a composite robust optimization problem with two layers of uncertainty. This approach allows us to translate the problem of choosing an uncertainty set for a risk measure into the task of specifying a ϕ_2 function given ϕ_1. By deriving a tractable reformulation of the composite robust problem, we show that our (ϕ_1, ϕ_2) characterization is not only computationally tractable, but also provides a blueprint for how to calibrate the ϕ_2-divergence uncertainty set to address tail uncertainty, given a risk measure and a nominal model. For risk measures such as the Conditional Value-at-Risk and the entropic risk measure, we provide examples showing the inadequacy of standard divergences to

evaluate model risk for certain nominal models, and provide explicit constructions of new ϕ_2-divergences that address these shortcomings, using our suggested framework. In addition, we show that our composition representation also offers other advantages, such as the elicitation of the divergences (ϕ_1, ϕ_2) through suitable queries to decision makers, and a natural extension to include a higher level of uncertainty using *globalized robust optimization* (as introduced in Ben-Tal et al. 2017), that also offers more flexibility in the construction of uncertainty sets. These results serve as extra tools for selecting proper divergences that are customized to a decision maker's risk and ambiguity preferences.

Finally, we provide a brief review of other related literature. Lam and Mottet [2017] studied the robust optimization approach to address tail uncertainty by constructing uncertainty sets that impose shape constraints on probability densities. The tractability of their formulation relies on an explicit representation of the risk measure in terms of density functions, which are not always obtainable for risk measures that are outside the expected utility framework. On the other hand, the tractability of our composite divergences formulation can be achieved via convex duality. Breuer and Csiszár [2016] used ϕ-divergences to measure model risk and provided a tractable reformulation of the corresponding worst-case expectation problem. Building upon this work, Kruse et al. [2019, 2021] demonstrated how specific ϕ-divergences can be designed to control the tail behaviors of the distributions in the divergence ball. Our work extends and unifies the results of Breuer and Csiszár [2016] and Kruse et al. [2019, 2021] to risk measures that are not simple expected values, such as the optimized certainty equivalents and shortfall risk measures. Glasserman and Xu [2014], Schneider and Schweizer [2015], Laeven and Stadje [2013] have used the Kullback-Leibler divergence to measure model risk for many financial problems, such as option hedging and portfolio credit risk management. Bertsimas and Brown [2009] provided a method for constructing uncertainty sets using coherent risk measures. Robust risk measures with uncertainty sets that are defined using the Wasserstein distance have also been studied by works such as Bartl et al. [2020] and Bernard et al. [2024], where the focus is more on tractability than on calibration of uncertainty sets.

The remaining parts of the paper are organized as follows: Section 2.2 formalizes our composite divergences approach in more detail. Sections 2.3 and 2.4 establish the conditions that we impose on ϕ_2 for different types of risk measures and provide explicit examples of the construction of new divergences, tailored to a given risk measure and nominal model. Section 2.5 shows how to elicit divergences from robust risk measures, and Section 2.6 discusses the computational aspects of robust risk measures, as well as an extension of our framework using globalized robust optimization to address higher levels of uncertainty. Numerical examples of robust options hedging and inventory planning problems are presented in Section 2.7. Conclusions are in Section 2.8. Proofs and additional technical details are contained in the Appendix. Computer code to implement the procedure

developed in this paper is available from: https://github.com/GuanJinNL/Uncertainty-Sets-for-Robust-Risk-Measures.git.

2.2 ϕ-Divergences and Risk Measures

Let Φ_0 denote the set of all non-negative functions $\phi : \mathbb{R} \to [0, \infty)$ that are convex and normalized, i.e., $\phi(1) = 0$, with an effective domain $\mathrm{dom}(\phi) \triangleq \{t \in \mathbb{R} \mid \phi(t) < \infty\} \subset [0, \infty)$ that also contains a neighborhood around 1. The ϕ-divergence between any two measures (not necessarily probability measures) $\mathbb{P}$ and $\mathbb{Q}$ on an event space $(\Omega, \mathcal{F})$ is defined as

$$I_\phi(\mathbb{Q}, \mathbb{P}) \triangleq \begin{cases} \int_\Omega \phi\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right) \mathrm{d}\mathbb{P} & \text{if } \mathbb{Q} \ll \mathbb{P} \\ +\infty & \text{else,} \end{cases} \tag{2.1}$$

where $\mathbb{Q} \ll \mathbb{P}$ denotes absolute continuity with respect to $\mathbb{P}$ (i.e., $\mathbb{P}(A) = 0 \Rightarrow \mathbb{Q}(A) = 0, \forall A \in \mathcal{F}$), and $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$ denotes the Radon-Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$. As in Pardo [2006], we adopt the conventions: $\phi(0) \triangleq \lim_{t \downarrow 0} \phi(t)$, $0\phi\,(0/0) = 0$, $0\phi\,(a/0) = \lim_{\epsilon \downarrow 0} \epsilon \phi\,(a/\epsilon) = a \lim_{t \to \infty} \phi(t)/t$, $a > 0$. We note that for any probability measures $\mathbb{Q}$, $\mathbb{P}$, we have $I_\phi(\mathbb{Q}, \mathbb{P}) \geq 0$ and $I_\phi(\mathbb{P}, \mathbb{P}) = 0$. If $\mathbb{P}$ and $\mathbb{Q}$ have density functions f, g, respectively, with respect to some σ-finite measure $\mathrm{d}x$ on $\mathbb{R}$,[1] then (2.1) can be expressed more explicitly in terms of real-valued density functions:

$$I_\phi(g, f) = \int_{\mathbb{R}} \phi\left(\frac{g(x)}{f(x)}\right) f(x) \mathrm{d}x.$$

Let $X : \Omega \to \mathbb{R}$ be a random payoff.[2] As shown in Ben-Tal and Teboulle [1987, 2007] and Ben-Tal et al. [1991], many types of risk measures/certainty equivalents admit a robust representation with a ϕ-divergence penalization. This is a consequence of the dual relationship between a utility function and a ϕ-divergence function through the Legendre-Fenchel transformation (a.k.a. convex/concave conjugate: for any ψ, its convex and concave conjugates are $\psi^*(\mathbf{y}) \triangleq \sup_{\mathbf{x} \in \mathbb{R}^d} \{\mathbf{y}^T \mathbf{x} - \psi(\mathbf{x})\}$ and $\psi_*(\mathbf{y}) \triangleq -(-\psi)^*(-\mathbf{y})$). For example, the optimized certainty equivalent (OCE) defined with respect to a utility function u and a probability measure $\mathbb{P}$, can be obtained and represented by the following robust problem, where all probability measures $\mathbb{Q} \ll \mathbb{P}$ are penalized by $\phi = -u_*$:

$$\begin{aligned} \rho_{\mathrm{oce},\mathbb{P}}(X) &\triangleq \sup_{\{\mathbb{Q}:\mathbb{Q} \ll \mathbb{P}\}} \mathbb{E}_\mathbb{Q}[-X] - I_\phi(\mathbb{Q}, \mathbb{P}) \\ &= \inf_{\eta \in \mathbb{R}} \{\eta - \mathbb{E}_\mathbb{P}[u(X + \eta)]\}. \end{aligned} \tag{2.2}$$

Therefore, the OCE is also known as a *divergence risk measure*, which includes examples

[1] Such as the Lebesgue measure or the counting measure or a combination of both.

[2] Negative payoffs are interpreted as losses.

such as:

- Conditional Value-at-Risk (CVaR_α) at $\alpha \in (0,1)$, for which $\phi(t) = \delta_{[0,1/(1-\alpha)]}(t)$ is the indicator function on $[0, 1/(1-\alpha)]$.[3]
- Entropic risk measure $\rho_{e,\gamma}(X) \triangleq \log\left(\mathbb{E}[\exp(-\gamma X)]\right)/\gamma$, $\gamma > 0$, for which $\phi(t) = (t\log(t) - t + 1)/\gamma$, for $t \geq 0$.

The class of utility-based shortfall risk measures (Föllmer and Schied, 2016, Section 4.9), a.k.a. u-Mean certainty equivalents, can also be represented in terms of a ϕ-divergence function:

$$\begin{aligned}\rho_{\text{sf},\mathbb{P}}(X) &\triangleq \sup_{\lambda>0} \sup_{\{\mathbb{Q}:\mathbb{Q}\ll\mathbb{P}\}} \mathbb{E}_{\mathbb{Q}}[-X] - I_{\phi^\lambda}(\mathbb{Q},\mathbb{P}) \\ &= \inf_{\eta\in\mathbb{R}}\{\eta \mid \mathbb{E}_{\mathbb{P}}[-u(X+\eta)] \leq 0\},\end{aligned} \tag{2.3}$$

where $\phi^\lambda(t) = \lambda\phi\left(t/\lambda\right)$ and $\phi(t) = -u_*(t)$.

Furthermore, for expected utility, the ϕ-divergence based robust representation takes the form of a maximization over all finite measures, instead of only probability measures:

$$\mathbb{E}_{\mathbb{P}}[-u(X)] = \sup_{\{\mathbb{Q}:\mathbb{Q}\ll\mathbb{P},\mathbb{Q}(\Omega)<\infty\}} \mathbb{E}_{\mathbb{Q}}[-X] - I_\phi(\mathbb{Q},\mathbb{P}), \tag{2.4}$$

where again $\phi = -u_*$.

Therefore, ϕ-divergences constitute basic building blocks for expected utility, utility-based shortfall risk measures, and optimized certainty equivalents. Using a second ϕ_2-divergence, we can parametrize robust risk measures of these classes by a composition of ϕ-divergences. We can address ambiguity in two ways: Either by means of the *multiple prior* approach (in the spirit of Gilboa and Schmeidler, 1989) via an uncertainty set — which we denote using ρ^s (where s comes from sets), or by means of the *multiplier preference* approach (in the spirit of Hansen and Sargent, 2008, Maccheroni et al., 2006, Laeven and Stadje, 2024) through penalization — which we denote using ρ^l (where l comes from penalization). In the case of OCE risk measures, this leads to the following composite robust formulations, where the maximization is taken over the two measures/variables $(\mathbb{Q}, \bar{\mathbb{Q}})$:

$$\rho^s_{\text{oce},\mathbb{P}_0}(X) \triangleq \sup_{\{\mathbb{Q}:I_{\phi_2}(\mathbb{Q},\mathbb{P}_0)\leq r\}} \sup_{\{\bar{\mathbb{Q}}:\bar{\mathbb{Q}}\ll\mathbb{Q}\}} \mathbb{E}_{\bar{\mathbb{Q}}}[-X] - I_{\phi_1}(\bar{\mathbb{Q}},\mathbb{Q}), \tag{2.5}$$

[3]For a set $\mathcal{C} \in \mathbb{R}^n$, the indicator function is defined as $\delta_{\mathcal{C}}(x) \triangleq \begin{cases} 0 & x \in \mathcal{C} \\ +\infty & x \notin \mathcal{C}. \end{cases}$

and

$$\rho^{l}_{\mathrm{oce},\mathbb{P}_0}(X) \triangleq \sup_{\{\mathbb{Q}:\mathbb{Q}\ll\mathbb{P}_0\}} \sup_{\{\bar{\mathbb{Q}}:\bar{\mathbb{Q}}\ll\mathbb{Q}\}} \mathbb{E}_{\bar{\mathbb{Q}}}[-X] - I_{\phi_1}(\bar{\mathbb{Q}},\mathbb{Q}) - I_{\phi_2}(\mathbb{Q},\mathbb{P}_0). \tag{2.6}$$

Other classes of robust risk measures are defined similarly. We note that for any constant $\lambda > 0$, we have $\lambda I_\phi(\mathbb{Q},\mathbb{P}) = I_{\lambda\phi}(\mathbb{Q},\mathbb{P})$. Hence, the representations in (2.5) and (2.6) also include penalization constants that are not equal to one. Throughout this paper, we fix $\mathbb{P}_0$ as a *given* nominal distribution. $\mathbb{P}_0$ may be a distribution that is already specified in an economic model, or a distribution calibrated from a parametric family. Furthermore, we assume that $X \in L^1(\mathbb{P}_0)$ (i.e., $\int |X| \mathrm{d}\mathbb{P}_0 < \infty$).

In the following section, we extensively study the robust OCE risk measures as the prime example of this paper. We show that by reformulating the problems (2.5) and (2.6) into finite-dimensional dual problems, we can obtain a blueprint for constructing ϕ_2 that is specifically tailored to a given risk measure associated with ϕ_1 and to the nominal model $\mathbb{P}_0$.

2.3 Divergence Choices for Robust OCE Risk Measures

2.3.1 Necessary and Sufficient Conditions for Finite Robust OCEs

To illustrate the importance of choosing a proper ϕ_2-divergence for a robust risk measure, consider an indicator divergence function $\phi_{2,\alpha}(t) = \delta_{[0,1/(1-\alpha)]}(t)$, for some $\alpha \in (0,1)$. Then, any alternative distribution with a density function g that has a heavier tail than the nominal density f_0 (i.e., $\lim_{x\to\infty} g(x)/f_0(x) = \infty$) will induce an infinite divergence value with respect to $\phi_{2,\alpha}$. Indeed, the likelihood ratio $g(x)/f_0(x)$ exceeds $1/(1-\alpha)$ for sufficiently large values of x, eventually lying outside the effective domain of $\phi_{2,\alpha}$. Therefore, the ϕ-divergence uncertainty set induced by $\phi_{2,\alpha}$ does not include any distributions that have a heavier right-side tail than the nominal distribution, while heavy-tailed losses are often the real threats in financial risk management that one wishes to be robust against.

The choice of ϕ_2-divergence also depends on the sensitivity of a given risk measure to tail losses. For example, the entropic risk measure $\rho_{e,1}(X) = \log(\mathbb{E}[\exp(-X)])$ is only finite for distributions for which the tail of their density functions decays faster than $\exp(-|x|)$. However, if the modified chi-squared divergence $\phi_2(x) = (x-1)^2$ is used to construct an uncertainty set where the nominal model for $Y \triangleq -X$ is the exponential distribution $f_0(y) = \lambda\exp(-\lambda y)$ for $y \geq 0, \lambda > 1$, then this can lead to an infinitely pessimistic robust risk evaluation. Indeed, any exponential distribution $g(y) = \eta\exp(-\eta y)$ with $\eta < 1 < \lambda < 2\eta$ has a finite modified chi-squared divergence with respect to f_0. Therefore, the modified chi-squared divergence ball will contain the distribution g for

sufficiently large radius r. However, since $1-\eta>0$, the entropic risk measure is infinite under the distribution of g. In fact, for any radius $r>0$, one can construct a density $\tilde{g}$, that has the same asymptotic tail behavior as $\exp(-\eta y)$, and such that $I_{\phi_2}(\tilde{g}, f_0)<r$ (see Kruse et al. 2021). Therefore, in this particular example, the modified chi-squared divergence ball will always lead to an infinite robust entropic risk measure evaluation, for any radius $r>0$.

Motivated by these examples, it is thus important to establish the necessary and sufficient conditions that one must impose on ϕ_2 to ensure a finite evaluation of the resulting robust risk measures. For robust OCE risk measures $\rho^s_{\mathrm{oce},\mathbb{P}_0}$ and $\rho^l_{\mathrm{oce},\mathbb{P}_0}$, these conditions can be identified by reformulating $\rho^s_{\mathrm{oce},\mathbb{P}_0}$ and $\rho^l_{\mathrm{oce},\mathbb{P}_0}$ into finite-dimensional problems using convex duality. This is stated in the following theorem, where for any divergence function ϕ, we define $0\phi^*(s/0)\triangleq 0$ if $s\leq 0$ and $0\phi^*(s/0)\triangleq\infty$ if $s>0$.

Theorem 2.1. *Let $X\in L^1(\mathbb{P}_0)$ and $\phi_1,\phi_2\in\Phi_0$ be lower-semicontinuous. Then, we have the following equalities:*

$$\rho^l_{\mathrm{oce},\mathbb{P}_0}(X)=\inf_{\theta_1,\theta_2\in\mathbb{R}} -\theta_1-\theta_2+\mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*(\phi_1^*(\theta_2-X)+\theta_1)\right] \tag{2.7}$$

$$\rho^s_{\mathrm{oce},\mathbb{P}_0}(X)=\inf_{\substack{\lambda\geq 0\\ \theta_1,\theta_2\in\mathbb{R}}} -\theta_1-\theta_2+\lambda r+\mathbb{E}_{\mathbb{P}_0}\left[\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2-X)+\theta_1}{\lambda}\right)\right]. \tag{2.8}$$

Furthermore, if the following conditions are satisfied:

1. *ϕ_1^* and ϕ_2^* are differentiable on $\mathbb{R}$,*
2. *Each domain set $\mathcal{F}(\boldsymbol{\theta})\triangleq\{\boldsymbol{\theta}\in\mathbb{R}^2 \mid \int_\Omega \beta^*(\omega,\boldsymbol{\theta})\mathrm{d}\mathbb{P}_0(\omega)<\infty\}$ and $\tilde{\mathcal{F}}(\boldsymbol{\theta},\lambda)\triangleq\{(\boldsymbol{\theta},\lambda)\in\mathbb{R}^2\times[0,\infty) \mid \int_\Omega \tilde{\beta}^*(\omega,\boldsymbol{\theta},\lambda)\mathrm{d}\mathbb{P}_0(\omega)<\infty\}$ has non-empty interior, where $\beta^*(\omega,\boldsymbol{\theta})\triangleq\phi_2^*(\phi_1^*(\theta_2-X(\omega))+\theta_1)$ and $\tilde{\beta}^*(\omega,\boldsymbol{\theta},\lambda)\triangleq\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2-X(\omega))+\theta_1}{\lambda}\right)$.*

Then, there exist dual solutions $\boldsymbol{\theta}^*$ and $(\tilde{\boldsymbol{\theta}}^*,\lambda^*)$ that attain the minimum in (2.7) and (2.8). Moreover, if $\boldsymbol{\theta}^*\in\mathrm{int}(\mathcal{F}(\boldsymbol{\theta}))$ and $(\tilde{\boldsymbol{\theta}}^*,\lambda^*)\in\mathrm{int}(\tilde{\mathcal{F}}(\boldsymbol{\theta},\lambda))$, then the partial derivatives of $\beta^*(\omega,.)$ and $\tilde{\beta}^*(\omega,.,\lambda)$ with respect to the variable $\boldsymbol{\theta}$, evaluated at the dual solutions $\boldsymbol{\theta}^*$ and $(\tilde{\boldsymbol{\theta}}^*,\lambda^*)$, are the worst-case densities of the measures $\mathbb{Q}^*$, $\bar{\mathbb{Q}}^*$ with respect to $\mathbb{P}_0$ that attain the maximum of the primal problems (2.6) and (2.5).

As a direct consequence of Theorem 2.1, we derive the following necessary and sufficient conditions for the finiteness of $\rho^s_{\mathrm{oce},\mathbb{P}_0}(X)$ and $\rho^l_{\mathrm{oce},\mathbb{P}_0}(X)$.

Corollary 2.2. *Let $X\in L^1(\mathbb{P}_0)$ and $\phi_1,\phi_2\in\Phi_0$ be lower-semicontinuous. Suppose there exists a real-valued density function $f_0(x)$ for the distribution of X under $\mathbb{P}_0$. Then, we*

have that $\rho^s_{\mathrm{oce},\mathbb{P}_0}(X) < \infty$*, if and only if there exist* $\theta_1, \theta_2 \in \mathbb{R}, \lambda \geq 0$*, such that*

$$\int_{\mathbb{R}} \lambda \phi_2^* \left(\frac{\phi_1^*(\theta_2 - x) + \theta_1}{\lambda} \right) f_0(x) \mathrm{d}x < \infty. \tag{2.9}$$

Similarly, we have that $\rho^l_{\mathrm{oce},\mathbb{P}_0}(X) < \infty$*, if and only if there exist* $\theta_1, \theta_2 \in \mathbb{R}$*, such that*

$$\int_{\mathbb{R}} \phi_2^* \left(\phi_1^*(\theta_2 - x) + \theta_1 \right) f_0(x) \mathrm{d}x < \infty. \tag{2.10}$$

The integral conditions (2.9) and (2.10) displayed in Corollary 2.2 show that the finiteness of a robust OCE risk measure is completely determined by the tail behaviors of three components: $(\phi_2^*, \phi_1^*, f_0)$, where ϕ_2^*, f_0 control the content of the ϕ_2-uncertainty set, and ϕ_1^* encodes the sensitivity of the risk measure to losses in the tail. This provides a guideline on how to calibrate an uncertainty set to a given risk measure and nominal model. Namely, one must choose the ϕ_2-divergence function such that its conjugate ϕ_2^* satisfies the integral condition in (2.9) or (2.10).

Remark 2.3. In practice, the nominal distribution of X might not be explicit since it can depend on many risk factors $Z_1, \ldots, Z_I$, where only the nominal distributions of the marginals Z_i's are specified and have an explicit form. In Section 2.C, we provide a sufficient condition under which one can choose the divergence based on the nominal distributions of Z_i's.

Using Corollary 2.2, we perform several tail analyses and illustrate how most standard choices of ϕ-divergences are not suitable to address tail uncertainty for some commonly used risk measures and nominal distributions. Throughout this paper, we use the big O notation $f_1(x) = O(f_2(x))$ if $\limsup_{x\to\infty} |f_1(x)|/|f_2(x)| < \infty$, and the small o notation $f_1(x) = o(f_2(x))$, if we have $\lim_{x\to\infty} |f_1(x)|/|f_2(x)| = 0$.

Example 2.4. Consider the CVaR_α risk measure with a log-normal nominal model. A common choice of a ϕ-divergence is the Kullback-Leibler divergence (a.k.a. relative entropy, see Csiszár, 1975 and Hansen and Sargent, 2008), where $\phi(t) = t\log(t) - t + 1$. Its conjugate is given by $\phi^*(s) = \exp(s) - 1$. The CVaR_α risk measure (1.6) is an OCE risk measure with $\phi_1^*(s) = \max\{s/(1-\alpha), 0\}$. However, examining the integral condition (2.9) reveals that for any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$, the integrand has a tail behavior of $O\left(\exp\left\{|x|/(\lambda\alpha)\right\}\right)$, which diverges to $+\infty$ as $x \to -\infty$. Hence, the uncertainty set induced by the Kullback-Leibler divergence is too conservative for the CVaR_α risk measure under a log-normal nominal model. In Table 2.1, we give an overview of the finiteness status of robust CVaR under other nominal distributions and ϕ-divergences. Moreover, we note that the same results as in Table 2.1 also apply to robust OCE risk measures with a picewise-linear utility function, as given in [Ben-Tal and Teboulle, 2007, Example 2.3].

Example 2.5. Another popular ϕ-divergence is the χ^2-distance given by $\phi(t) = (t-1)^2/t$. However, its conjugate $\phi^*(s) = 2-2\sqrt{1-s},\ s < 1$, is only finite on the domain on $(-\infty, 1)$, i.e., $\phi^*(s) = +\infty$, for $s > 1$. Hence, the integral condition (2.9) can never be satisfied if the χ^2-distance is used to define a robust OCE risk measure, for any nominal distribution that has unbounded support towards $-\infty$.

Similarly, other canonical examples of ϕ-divergences such as the Burg entropy $\phi(t) = -\log(t) + t - 1$, total variation distance $\phi(t) = |t-1|$, and polynomial divergence $\phi(t) = (t^p - p(t-1) - 1)/(p(p-1))$ with degree $p < 1$, all have a conjugate function ϕ^* that is infinite on a certain interval $[a, \infty)$ (see Table 2 of Ben-Tal et al. 2013 and page 6 of Pardo 2006). Hence, they all lead to an infinite robust OCE risk evaluation for a nominal distribution with unbounded support towards $-\infty$.

Example 2.6. The entropic risk measure $\rho_{e,\gamma}(X) = \log\left(\mathbb{E}[\exp(-\gamma X)]\right)/\gamma$ is an OCE risk measure with $\phi_1^*(x) = (\exp(\gamma x) - 1)/\gamma$. Due to its exponential growth behavior, the entropic risk measure has an infinite evaluation for heavy-tailed distributions and, therefore, the same applies to its robust counterpart. For the Gaussian distribution, the entropic risk measure is finite and equal to $\mu + \gamma\sigma^2/2$. However, if one chooses the KL-divergence to define a robust entropic risk measure with a Gaussian nominal model, then condition (2.9) is *not* satisfied for any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$, since the integrand has tail asymptotics of the order

$$O\left(\exp\left\{\frac{1}{\lambda}\left(\frac{1}{\gamma}\left(\exp(\gamma(\theta_2 - x)) - 1\right) + \theta_1\right) - \left(\frac{x-\mu}{\sigma}\right)^2\right\}\right),$$

which diverges to ∞ as $x \to -\infty$. Therefore, the KL-divergence is not adequate to address model uncertainty for the entropic risk measure, under many nominal distributions such as the Gaussian and Weibull distributions, as shown in Table 2.2.

Example 2.7. The OCE risk measure with a quadratic utility function

$$u(x) = \begin{cases} x - \frac{1}{2}x^2 & x < 1 \\ \frac{1}{2} & x \geq 1, \end{cases}$$

is equivalent to the mean-variance evaluation $\mu - \sigma^2/2$, as shown in Ben-Tal and Teboulle [2007, Example 2.2]. In Table 2.3, we also provide an overview of the finiteness status of this robust OCE risk measure under different nominal distributions and ϕ-divergences.

From these examples, we observe that many standard choices of ϕ-divergence uncertainty sets often lead to infinitely conservative robust risk evaluations. On the other hand, the polynomial divergence $\phi(t) = (t^p - p(t-1) - 1)/(p(p-1))$ can often be made to satisfy the finiteness conditions (2.9) and (2.10), by choosing a sufficiently large p. Indeed, an examination of its conjugate function $\phi^*(s) = \max\{1 + s(p-1), 0\}^{\frac{p}{p-1}}/p - 1/p$ shows

Divergence $\phi(t)$	$N(\mu,\sigma^2)$	$W(\lambda_0,k)$	$\ln(\mu,\sigma^2)$	$P(x_m,p_0)$	$t(\nu)$
$t\log t - t + 1$ *(Kullback-Leibler)*	$<\infty$	$<\infty \Leftrightarrow k \geq 1$	∞	∞	∞
$(t^p - pt + p - 1)/(p(1-p))$ *(Polynomial, $p>1$)*	$<\infty$	$<\infty$	$<\infty$	$<\infty \Leftrightarrow \frac{p}{p-1} < p_0$	$<\infty \Leftrightarrow \frac{p}{p-1} < \nu$
$(t-1)^2$ *(Modified χ^2-distance)*	$<\infty$	$<\infty$	$<\infty$	$<\infty \Leftrightarrow p_0 > 2$	$<\infty \Leftrightarrow \nu > 2$
$(t-1)^2/t$ *(χ^2-distance)*	∞	∞	∞	∞	∞
$\lvert t-1\rvert$ *(Variation distance)*	∞	∞	∞	∞	∞
$(\sqrt{t}-1)^2$ *(Hellinger distance)*	∞	∞	∞	∞	∞
$-\log t + t - 1$ *(Burg entropy)*	∞	∞	∞	∞	∞

Table 2.1: Finiteness status of robust CVaR_α risk measures (2.5) and (2.6) for different divergence measures versus nominal distributions. $N(\mu,\sigma^2)$ and $\ln(\mu,\sigma^2)$ are respectively the normal and lognormal distribution with mean μ and variance σ^2. $W(\lambda_0,k)$ denotes the Weibull distribution with scale λ_0 and shape k. $P(x_m,p_0)$ is the Pareto distribution with scale x_m and shape p_0. $t(\nu)$ denotes the Student's t-distribution with ν degrees of freedom. Details of derivations can be found in Section 2.E.

that as $p \to \infty$, ϕ^* approaches the linear function $\phi^*(s) = s$, which is the least conservative conjugate function of the ϕ-divergence that only considers the nominal model (i.e., $\phi(1) = 0$, $\phi(t) = +\infty$ elsewhere). However, the drawback of a polynomial divergence uncertainty set is that it can be very restrictive. For example, consider two Weibull distributions $g(x) = kx^{k-1}\exp(-x^k)$ and $f(x) = lx^{l-1}\exp(-x^l)$ with $k < l$. Then, any polynomial divergence $\phi(t) = O(t^p)$ with $p > 1$ would give $I_\phi(g,f) = \infty$, since $\phi\left(g(x)/f(x)\right) f(x) = O(x^{k-l}\exp((p-1)x^l - px^k))$. Therefore, when the nominal distribution belongs to the Weibull class, an uncertainty set induced by the polynomial divergence does not include any other Weibull distribution that has a heavier tail. Hence, apart from the conditions in Corollary 2.2, we need to impose additional constraints on the conjugate function ϕ_2^*, such that the corresponding divergence ϕ_2 is not too restrictive. We will address this issue in the next subsection.

2.3.2 Controlling the Tail Properties of Distributions in a Divergence Ball

As illustrated by the previous examples, many classical divergences fail to guarantee the finiteness of a robust risk measure, due to improper matching of tail behaviors between ϕ_2^*, ϕ_1^*, and f_0. Therefore, to construct an alternative ϕ_2-divergence that addresses this issue, we can specify a new tail function $\tilde{\psi}(s), s \geq 0$ that is compatible with (ϕ_1^*, f_0) in the sense of Corollary 2.2, and then set $\phi_2^*(s) = \tilde{\psi}(s)$, for $s \geq 0$. Since the integral conditions in Corollary 2.2 depend only on the tail behavior of $\phi_2^*(s)$ as $s \to \infty$, we may simply set $\phi_2^*(s) = \exp(s) - 1$ on the domain $s \leq 0$. Finally, the new tail function $\tilde{\psi}(s)$ must be

Divergence $\phi(t)$	$N(\mu,\sigma^2)$	$W(\lambda_0,k)$	$\ln(\mu,\sigma^2)$	$P(x_m,p_0)$	$t(\nu)$
$t\log t - t + 1$ *(Kullback-Leibler)*	∞	∞	∞	∞	∞
$(t^p - pt + p - 1)/(p(1-p))$ *(Polynomial, $p > 1$)*	$< \infty$	$< \infty \Leftrightarrow \begin{cases} k > 1 \\ k = 1 \text{ and } \frac{p}{p-1} < \frac{\gamma}{\lambda_0} \end{cases}$	∞	∞	∞
$(t-1)^2$ *(Modified χ^2-distance)*	$< \infty \Leftrightarrow \gamma > 2\sigma^2$	$< \infty \Leftrightarrow \begin{cases} k > 2 \\ k = 2 \text{ and } \gamma > \lambda_0 \end{cases}$	∞	∞	∞
$(t-1)^2/t$ *(χ^2-distance)*	∞	∞	∞	∞	∞
$\lvert t-1\rvert$ *(Variation distance)*	∞	∞	∞	∞	∞
$(\sqrt{t}-1)^2$ *(Hellinger distance)*	∞	∞	∞	∞	∞
$-\log t + t - 1$ *(Burg entropy)*	∞	∞	∞	∞	∞

Table 2.2: Finiteness status of robust entropic risk measures (2.5) and (2.6) with parameter $\gamma > 0$, for different divergence measures versus nominal distributions, with exponential utility OCE. $N(\mu,\sigma^2)$ and $\ln(\mu,\sigma^2)$ are respectively the normal and lognormal distribution with mean μ and variance σ^2. $W(\lambda_0,k)$ denotes the Weibull distribution with scale λ_0 and shape k. $P(x_m,p_0)$ is the Pareto distribution with scale x_m and shape p_0. $t(\nu)$ denotes the Student's t-distribution with ν degrees of freedom. Details of derivations can be found in Section 2.E.

normalized such that $(\phi_2^*)^*$ becomes a ϕ-divergence function (i.e., a convex, non-negative function satisfying $\phi(1) = 0$ and $\mathrm{dom}(\phi) \subset [0,\infty)$). This can be done by choosing $\tilde{\psi}(s)$ to be increasing and convex with $\tilde{\psi}(0) = 0$ and $\tilde{\psi}'(0) = 1$. To conclude, we propose the following construction: let ψ be an increasing, convex differentiable function on $[0,\infty)$ such that $\psi''(0) \neq 0$. Set, for $s \geq 0$,

$$\tilde{\psi}(s) = \frac{1}{\psi''(0)} \left(\psi(s) + (\psi''(0) - \psi'(0))\, s - \psi(0)\right), \tag{2.11}$$

and define the function

$$\phi_2^*(s) = \begin{cases} \tilde{\psi}(s) & s \geq 0 \\ \exp(s) - 1 & s \leq 0. \end{cases} \tag{2.12}$$

By construction, we have $(\phi_2^*)'(0) = 1$ and $(\phi_2^*)''(0) = 1$. These are normalization conditions that ensure $\phi_2(1) = 0, \phi_2'(1) = 0$ and $\phi_2''(1) = 1$, where $\phi_2 = \phi_2^{**}$ (the first two conditions are needed for ϕ_2 to be a ϕ-divergence, and $\phi_2''(1) = 1$ is often desirable in the statistical application of ϕ-divergences, see Remark 1 of Kruse et al. 2019 and Pardo 2006). The following proposition, which is an application of Corollary 23.5.1 of Rockafellar [1970], verifies that ϕ_2^*, as defined in (2.12), is indeed the conjugate function of a ϕ-divergence.

Proposition 2.8. *The function $\phi_2 \triangleq (\phi_2^*)^*$, where ϕ_2^* is defined in* (2.12)*, is a convex, non-negative function on $[0,\infty)$ with $\phi_2(1) = 0$. If ψ in* (2.11) *is strictly convex and continuously differentiable with $\lim_{s\to\infty} \psi'(s) = \infty$, then ϕ_2 is finite on $[0,\infty)$ and we*

Divergence $\phi(t)$	$N(\mu,\sigma^2)$	$W(\lambda_0,k)$	$\ln(\mu,\sigma^2)$	$P(x_m,p_0)$	$t(\nu)$
$t\log t - t + 1$ *(Kullback-Leibler)*	$<\infty$	$<\infty \Leftrightarrow k \geq 2$	∞	∞	∞
$(t^p - pt + p - 1)/(p(1-p))$ *(Polynomial, $p>1$)*	$<\infty$	$<\infty$	$<\infty$	$<\infty \Leftrightarrow \frac{2p}{p-1} < p_0$	$<\infty \Leftrightarrow \frac{2p}{p-1} < \nu$
$(t-1)^2$ *(Modified χ^2-distance)*	$<\infty$	$<\infty$	$<\infty$	$<\infty \Leftrightarrow p_0 > 4$	$<\infty \Leftrightarrow \nu > 4$
$(t-1)^2/t$ *(χ^2-distance)*	∞	∞	∞	∞	∞
$\lvert t-1\rvert$ *(Variation distance)*	∞	∞	∞	∞	∞
$(\sqrt{t}-1)^2$ *(Hellinger distance)*	∞	∞	∞	∞	∞
$-\log t + t - 1$ *(Burg entropy)*	∞	∞	∞	∞	∞

Table 2.3: Finiteness status of robust quadratic OCE risk measures (i.e., (2.5) and (2.6) with $\phi_1^*(s) = s + s^2/2$ for $s > -1$ and $\phi_1^*(s) = -1/2$ for $s \leq -1$, see Ben-Tal and Teboulle, 2007, Example 2.2) for different ϕ_2-divergences versus nominal distributions. $N(\mu,\sigma^2)$ and $\ln(\mu,\sigma^2)$ are respectively the normal and lognormal distribution with parameters μ and σ^2. $W(\lambda_0,k)$ denotes the Weibull distribution with scale λ_0 and shape k. $P(x_m,p_0)$ is the Pareto distribution with scale x_m and shape p_0. $t(\nu)$ denotes the Student's t-distribution with ν degrees of freedom. Details of the derivations can be found in Section 2.E.

have that $\phi_2'(t) = ((\phi_2^*)')^{-1}(t)$*, for* $t > 0$.

Besides the finiteness of a robust risk measure, it is also desirable to have a divergence that is not too restrictive, and allows for distributions with heavier tails than the nominal one in the divergence ball.[4] Intuitively, the degree of penalization induced by ϕ_2 is determined by its growth rate as $t \to \infty$, which is governed by the derivative $\phi_2'(t)$. Therefore, one must control $\phi_2'(t)$ to adjust the content of a divergence ball. However, if the divergence is constructed through the conjugate as in (2.12), then there is not always an explicit form of ϕ_2' available that allows us to examine its growth behavior as $t \to \infty$. Indeed, although the equality $\phi_2' = ((\phi_2^*)')^{-1}$ holds, the inverse of $(\phi_2^*)'$ cannot always be calculated explicitly. Therefore, it would seem more convenient to specify ϕ_2 through its derivative instead of the conjugate, and to derive sufficient conditions for (2.9) and (2.10) in terms of ϕ_2' (e.g., by bounding the conjugate using $\phi_2^*(s) \leq s(\phi_2')^{-1}(s)$, similarly as in Kruse et al. 2019). This creates a dilemma, since the latter construction does not give an analytical expression for the conjugate function and its derivatives, which is important for the use of optimization techniques to compute the robust risk measures via their dual problems, as we elaborate in Section 2.6. Nevertheless, we can resolve this dilemma as follows: first, we use the divergence construction as given in (2.12) to obtain an explicit conjugate function ϕ_2^*. Second, we calculate the derivative $(\phi_2^*)'$, which gives $(\phi_2')^{-1}$ by Proposition 2.8. Finally, we may find functions with similar tail properties as $(\phi_2^*)'$ that provide upper and lower bounds on $(\phi_2^*)'$, for which the inverse can be explicitly calcu-

[4] A divergence ball is a set of the form $\{\mathbb{Q} \mid I_\phi(\mathbb{Q},\mathbb{P}) \leq r\}$.

lated for examining the content of a divergence ball. We state this more precisely in the following proposition.

Proposition 2.9. *Let ϕ_2 be a continuous divergence function with a strictly increasing derivative ϕ_2'. Suppose ψ_1' is a continuous function such that its inverse satisfies $(\psi_1')^{-1}(y) \leq (\phi_2')^{-1}(y)$ for all $y \geq y_0$, for some $y_0 \in \mathbb{R}$. If for some $d > 1$, we have that*

$$\limsup_{|x|\to\infty} \psi_1'\left(\frac{1}{f_0(x)}\right)|x|^{-d} < \infty, \tag{2.13}$$

then for any probability density g such that the likelihood g/f_0 is bounded on any compact subset of $(-\infty, 0]$, we have that if g has a finite d-th moment: $\int_{-\infty}^{\infty} g(x)|x|^d \mathrm{d}x < \infty$, then we also have $I_{\phi_2}(g, f_0) < \infty$.

On the other hand, if there exists a divergence function $\psi_2 \in \Phi_0$ such that $(\phi_2')^{-1}(y) \leq (\psi_2')^{-1}(y)$ for all $y \geq 0$, we have that $I_{\phi_2}(g, f_0) = \infty$, if $I_{\psi_2}(g, f_0) = \infty$, for any density function g.

2.3.3 Examples of Tailored Divergences

In this subsection, we provide examples of new ϕ_2-divergences that are tailored to a given risk measure and nominal model. We will only specify $\phi_2^*(s)$ for $s \geq 0$ in these examples, since we always set $\phi_2^*(s) = \exp(s) - 1$, for $s \leq 0$ as in (2.12).

Example 2.10. Consider the CVaR_α risk measure with the generalized log-normal nominal distribution, for which the density function is given by

$$f_0(x) = \frac{1}{C\sigma|x|} \exp\left\{-\frac{1}{p\sigma^p}(\log|x| - \mu)^p\right\}, \ x \leq 0,$$

where $\mu \in \mathbb{R}, \sigma > 0, p \geq 2$, and $C > 0$ is a normalization constant. As shown in Example 2.4, the KL-divergence is not suitable for this combination of nominal model and risk measure. Hence, using the construction method in (2.12), we propose the following divergence specified by its conjugate function:

$$\phi_2^*(s) = c_1(s + e)\exp\left\{\frac{1}{p\,(\sigma d)^p}\log^p(s + e)\right\} + c_2 s + c_3, \ s \geq 0, \tag{2.14}$$

where the constants are $c_1 = 1/(p^2(a^2 + a)\exp(a - 1))$, $c_2 = 1 - \exp(a)(ap + 1)c_1$, $c_3 = -\exp(a + 1)c_1$ with $a = 1/(p\,(\sigma d)^p)$ and $d > 1$. We note that the proposed divergence (2.14) contains the parameters (σ, p), due to its dependence on the log-normal nominal model. It also contains a parameter $d > 1$, which controls the moments of the distributions in its divergence ball. Notably, the divergence (2.14) is independent of the parameter α of the Conditional Value-at-Risk, which is due to the linearity of the ϕ_1^* function corre-

sponding to the CVaR_α risk measure, which does not determine the asymptotic behavior of the integrand in (2.10).

Lemma 2.11. *The function ϕ_2^* defined in* (2.14) *is strictly convex and twice continuously differentiable on $\mathbb{R}$. Its conjugate $\phi_2 \triangleq (\phi_2^*)^*$ belongs to Φ_0. Moreover, ϕ_2^* satisfies* (2.9) *and* (2.10)*, for the* CVaR_α *risk measure and the generalized log-normal nominal model.*

In addition, for any continuous probability density g defined on $(-\infty, 0]$ that has finite d-th moment: $\int_{-\infty}^0 g(x)|x|^d < \infty$, we have that $I_{\phi_2}(g, f_0) < \infty$. Conversely, for any density function g such that $\liminf_{x\to-\infty} |x|^{t+1} g(x) > 0$, where $t \in (1, d)$, we have $I_{\phi_2}(g, f_0) = \infty$.

Example 2.12. Consider now the CVaR_α risk measure with a Weibull nominal distribution, for which the density f_0 is given by:

$$f_0(x) = \frac{k}{\lambda}\left(\frac{|x|}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{|x|}{\lambda}\right)^k\right\},\ x \le 0,$$

where $\lambda,\ k > 0$. If $k < 1$, the Weibull distribution becomes heavy-tailed, and the robust CVaR_α risk measure defined by the KL-divergence will lead to an infinite evaluation. To match the exponential tail of f_0, we propose the following divergence:

$$\phi_2^*(s) = c_1(s+1)\exp\{(s+1)^{k/d}\} + c_2 s + c_3,\ s \ge 0, \tag{2.15}$$

where $d > 1$, and the normalization constants are $c_1 = 1/(ep(2p+1))$, $c_2 = 1 - (1+p)/(p(2p+1))$, $c_3 = -1/(p(2p+1))$ with $p = k/d$.

Lemma 2.13. *The function ϕ_2^* defined in* (2.15) *is strictly convex and twice continuously differentiable on $\mathbb{R}$. Its conjugate $\phi_2 \triangleq \phi_2^{**}$ belongs to Φ_0. Moreover, ϕ_2^* satisfies the finiteness conditions* (2.9) *and* (2.10)*, for the* CVaR_α *risk measure and the Weibull nominal model.*

In addition, for any continuous probability density g defined on $(-\infty, 0]$ that has finite d-th moment: $\int_{-\infty}^0 g(x)|x|^d < \infty$, we have that $I_{\phi_2}(g, f_0) < \infty$. Conversely, for any density function g such that $\liminf_{x\to-\infty} |x|^{t+1} g(x) > 0$, where $t \in (1, d)$, we have $I_{\phi_2}(g, f_0) = \infty$.

Example 2.14. Consider the entropic risk measure $\rho_{e,\gamma}$ with the Weibull nominal distribution, where $k > 1$. Table 2.2 shows that the KL-divergence is generally not suitable for the entropic risk measure. Moreover, as mentioned previously, the polynomial divergence is too restrictive since its uncertainty set does not include any Weibull distributions with a heavier tail than the nominal one. Therefore, we propose the following divergence:

$$\phi_2^*(s) = c_1(s+e)\exp\left\{\frac{1}{(2\gamma\lambda)^k}\log^k(s+e)\right\} + c_2 s + c_3,\ s \ge 0, \tag{2.16}$$

where the constants are $c_1 = 1/(k^2(a^2+a)\exp(a-1))$, $c_2 = 1 - \exp(a)(ak+1)/c_1$, $c_3 = -\exp(a+1)/c_1$ with $a = 1/(2\gamma\lambda)^k$.

Lemma 2.15. *The function ϕ_2^* defined in* (2.16) *is strictly convex and twice continuously differentiable on $\mathbb{R}$. Its conjugate $\phi_2 \triangleq \phi_2^{**}$ belongs to Φ_0. Moreover, ϕ_2^* satisfies the finiteness conditions* (2.9) *and* (2.10)*, for the entropic risk measure (with parameter $\gamma > 0$) and the Weibull nominal distribution f_0 with parameters (λ, k). Furthermore, for any Weibull density function g with parameters (λ, l) such that $1 < l < k$, we have $I_{\phi_2}(g, f_0) < \infty$.*

2.4 Other Types of Robust Risk Measures

In this section, we show that the formulation with two ϕ-divergences (ϕ_1, ϕ_2) provides a general framework that also encompasses other families of robust risk measures. Similar to the previous section, we establish the necessary and sufficient conditions for finiteness and derive the finite-dimensional dual problems of the corresponding robust risk measures.

In this section, we focus on the class of robust shortfall risk measures,[5] which can be characterized as follows:

$$\rho^s_{\mathrm{sf},\mathbb{P}}(X) \triangleq \sup_{\lambda>0} \sup_{\{\mathbb{Q}: I_{\phi_2}(\mathbb{Q},\mathbb{P}_0)\le r\}} \sup_{\{\bar{\mathbb{Q}}:\bar{\mathbb{Q}}\ll\mathbb{Q}\}} \mathbb{E}_{\bar{\mathbb{Q}}}[-X] - I_{\phi_1^\lambda}(\bar{\mathbb{Q}}, \mathbb{Q}). \tag{2.17}$$

By examining the dual of (2.17), we can again establish the finiteness conditions for (2.17) in terms of the tail behavior of ϕ_1^*, ϕ_2^* and f_0.

Theorem 2.16. *Let $X \in L^1(\mathbb{P}_0)$. Let $\phi_1, \phi_2 \in \Phi_0$ be lower-semicontinuous. Then, we have that $\rho^s_{\mathrm{sf},\mathbb{P}}(X) < \infty$, if and only if there exist $\theta_1, \theta_2 \in \mathbb{R}, \eta \ge 0$, such that*

$$\int_{\mathbb{R}} \eta \phi_2^* \left(\frac{\phi_1^*(-\theta_2 - x) + \theta_1}{\eta} \right) f_0(x)\mathrm{d}x \le \theta_1 - \eta r. \tag{2.18}$$

Moreover, we have the duality relation:

$$\rho^s_{\mathrm{sf},\mathbb{P}}(X) = \inf_{\substack{\eta\ge 0 \\ \theta_1,\theta_2\in\mathbb{R}}} \left\{ \theta_2 \mid \mathbb{E}_{\mathbb{P}_0}\left[\beta^*(X, \theta_1, \theta_2, \eta)\right] \le \theta_1 - \eta r \right\}, \tag{2.19}$$

where $\beta^*(X, \theta_1, \theta_2, \eta) \triangleq \eta \phi_2^* \left(\frac{\phi_1^*(-\theta_2 - X) + \theta_1}{\eta} \right)$.

A similar duality result cannot be obtained if the robust shortfall risk measure is defined by adding a second ϕ_2 penalization (such as in (2.6)), due to the non-convex product term $\lambda\mathbb{Q}$ in $I_{\phi_1^\lambda}(\bar{\mathbb{Q}}, \mathbb{Q})$.[6] Nevertheless, as an alternative, we note that one can

[5] We briefly note that the robust expected utility risk measure can also be formulated in terms of two ϕ-divergences (ϕ_1, ϕ_2). Indeed, as shown in (2.4), expected utility can be represented by the divergence $\phi_1 = -u_*$. The dual reformulation of robust expected utility risk measures has already been derived in Breuer and Csiszár [2016], and its implication for divergence choices has been studied in Kruse et al. [2019, 2021].

[6] This is also the case for (2.17), but the set formulation in (2.17) allows us to circumvent the non-convexity by a change of variable.

instead also define a robust shortfall risk measure by adding the ϕ_2 penalization to the expectation constraint in (2.3), and then apply the reformulation in Breuer and Csiszár [2016] for the robust expectation. This would lead to the dual formulation in (2.26).

The necessary and sufficient conditions for $\rho^s_{\mathrm{sf},\mathbb{P}}(X) < \infty$ as established in (2.18) are not easy to directly verify analytically. Indeed, the integral on the left-hand side of (2.18) requires an explicit calculation, whereas in (2.9), for example, one only needs to perform an asymptotic tail analysis. To work around this, we note that the right-hand side of (2.18) is independent of θ_2, whereas the left-hand side is monotonically decreasing as $\theta_2 \to \infty$. This allows us to develop a set of tests presented in the following corollary, where (2.18) can be verified without the need for an explicit calculation of the integral in (2.18).

Corollary 2.17. *Let the assumptions of Theorem 2.16 be fulfilled, and suppose that* $\phi_1^*(-\infty) \triangleq \lim_{s\to-\infty} \phi_1^*(s)$ *exists, either finite or equal to* $-\infty$*. Then, the following holds:*

1. *If there exist no* $\theta_1, \theta_2 \in \mathbb{R}$, $\eta \geq 0$, *such that*

$$\int_{\mathbb{R}} \eta\phi_2^*\left(\frac{\phi_1^*(-\theta_2 - x) + \theta_1}{\eta}\right) f_0(x)\mathrm{d}x < \infty, \tag{2.20}$$

   then $\rho^s_{\mathrm{sf},\mathbb{P}}(X) = \infty$.

2. *If for all* $\theta_1, \theta_2 \in \mathbb{R}$, $\eta \geq 0$ *that satisfy* (2.20), *we have*

$$\eta\phi_2^*\left(\frac{\phi_1^*(-\infty) + \theta_1}{\eta}\right) > \theta_1 - \eta r, \tag{2.21}$$

   then $\rho^s_{\mathrm{sf},\mathbb{P}}(X) = \infty$ *(if* $\phi_1^*(-\infty) = -\infty$*, we define* $\eta\phi_2^*\left(\frac{-\infty+\theta_1}{\eta}\right) \triangleq \lim_{s\to-\infty} \eta\phi_2^*\left(\frac{s}{\eta}\right)$*, for any* $\eta \geq 0, \theta_1 \in \mathbb{R}$*).*

3. *If there exist some* $\theta_1, \theta_2 \in \mathbb{R}$, $\eta \geq 0$, *such that* (2.20) *holds, and*

$$\eta\phi_2^*\left(\frac{\phi_1^*(-\infty) + \theta_1}{\eta}\right) < \theta_1 - \eta r, \tag{2.22}$$

   then, $\rho^s_{\mathrm{sf},\mathbb{P}}(X) < \infty$.

Example 2.18. Consider a shortfall risk measure with a quadratic $\phi_1(t) = (t-1)^2$. Then,

$$\phi_1^*(s) = \begin{cases} s + s^2/4 & s \geq -2 \\ -1 & s < -2. \end{cases}$$

We show that the robust shortfall risk measure with the following combinations of divergence ϕ_2 and nominal distribution f_0 is finite for $r \leq 1$:

(1) Polynomial divergence with $p = 3$ and Log-normal distribution

(2) KL-divergence and Gaussian distribution (with $\sigma \leq 1$)

(3) Modified χ^2-divergence and Student-t distribution (with $\nu > 4$)

(1): We apply Corollary 2.17. First, for any $\theta_1, \theta_2 \in \mathbb{R}$, $\eta > 0$, as $x \to -\infty$,

$$\eta\phi_2^*\left(\frac{\phi_1^*(-\theta_2 - x) + \theta_1}{\eta}\right) f_0(x) = O\left(|x|^{2p/(p-1)} \exp\left(-\log^2|x|/(2\sigma^2)\right)\right),$$

which is integrable for any $p > 1$. Hence, (2.20) is satisfied for any $\theta_1, \theta_2 \in \mathbb{R}$, $\eta > 0$. It remains to verify (2.22). This can be verified numerically. We find that for $\theta_1 = 0.1098$, $\eta = 0.01$, (2.22) is satisfied for all $r \leq 1$.

(2): For any $\theta_1, \theta_2 \in \mathbb{R}$ and $\eta > \sigma^2/2$, we have that for $x \leq 2 - \theta_2$ (note that the remaining part of the integrand below for is integrable $x > 2 - \theta_2$):

$$\eta\phi_2^*\left(\frac{\phi_1^*(-\theta_2 - x) + \theta_1}{\eta}\right) f_0(x) = O\left(\exp\left(\left(\frac{1}{4\eta} - \frac{1}{2\sigma^2}\right)x^2\right)\right).$$

We note that $1/(4\eta) - 1/(2\sigma^2) < 0$ since $\eta > \sigma^2/2$. Hence, (2.20) is satisfied for any $\theta_1, \theta_2 \in \mathbb{R}$ and $\eta > \sigma^2/2$. It remains to verify (2.22). We find numerically that for $\theta_1 = 0.2098$, $\eta = 0.501$, (2.22) is satisfied for all $\sigma \leq 1, r \leq 1$.

(3): For any $\theta_1, \theta_2 \in \mathbb{R}$, $\eta > 0$, we have that as $x \to -\infty$,

$$\eta\phi_2^*\left(\frac{\phi_1^*(-\theta_2 - x) + \theta_1}{\eta}\right) f_0(x) = O\left(|x|^{4-\nu-1}\right),$$

which is only integrable if $\nu > 4$. It remains to verify (2.22). We find numerically that for $\theta_1 = 0.1098$, $\eta = 0.501$, (2.22) is satisfied for all $r \leq 1$.

As a final remark, we briefly mention distortion risk measures, which are defined by non-additive integrals associated with a distortion function $h : [0,1] \to [0,1]$ (see Denneberg, 1994, Föllmer and Schied, 2016, Section 4.6, and Föllmer and Knispel, 2013). If a robust distortion risk measure is defined by a ϕ_2-divergence uncertainty set as in (2.5) (i.e., $\sup_{\{\mathbb{Q}: I_{\phi_2}(\mathbb{Q},\mathbb{P}) \leq r\}} \rho_h(X)$, where ρ_h is a distortion risk measure) then a sufficient condition on its finiteness can be established through that of CVaR_α, whenever the distortion function is concave and has a finite derivative at zero (i.e., $h'(0) < \infty$). Indeed, a distortion risk measure with a concave distortion function h and finite $h'(0)$ can always be upper bounded by the CVaR_α risk measure, where the level α satisfies $1/(1-\alpha) > h'(0)$. This follows from the fact that CVaR_α is itself a distortion risk measure with $h(p) = \min\{p/(1-\alpha), 1\}$.

2.5 Elicitation of Divergences from Robust Risk Measures

In the previous sections, we have shown that the robust risk measures characterized by (ϕ_1, ϕ_2) can be reformulated as finite-dimensional dual problems in terms of the conjugates (ϕ_1^*, ϕ_2^*). In this section, we study the inverse problem of recovering the (ϕ_1^*, ϕ_2^*) from the robust risk measures. In the non-robust case where ambiguity is not present, it is shown by Ben-Tal and Ben-Israel [1991], Ben-Tal and Teboulle [2007] that utility functions that are normalized and strongly risk-averse (i.e., $u(0) = 0, u(x) < x, \forall x \in \mathbb{R}$) can be elicited by querying the certainty equivalent values (EU, OCE, and u-Mean) of random variables X_p, which for any $x \in \mathbb{R}$ are defined as

$$X_p = \begin{cases} x & \text{with probability } p \\ 0 & \text{with probability } 1-p. \end{cases} \tag{2.23}$$

More precisely, for any certainty equivalent $\mathrm{CE} \in \{\mathrm{EU}, \mathrm{OCE}, \text{u-Mean}\}$, we have the limit relation

$$\lim_{p \downarrow 0} \frac{\mathrm{CE}(X_p)}{p} = u(x). \tag{2.24}$$

Our goal is to extend the identity (2.24) to robust risk measures, which we show to be possible for the robust risk measures $\{\rho^l_{\mathrm{eu},\mathbb{P}}, \rho^l_{\mathrm{oce},\mathbb{P}}, \rho^l_{\mathrm{sf},\mathbb{P}}\}$ that are defined through penalization. In the case of robust expected utility $\rho^l_{\mathrm{eu},\mathbb{P}}$, such an extension is already evident due to its duality relation to OCE (see (2.2)):

$$\rho^l_{\mathrm{eu},\mathbb{P}}(X) \triangleq \inf_{\theta \in \mathbb{R}} -\theta + \mathbb{E}_{\mathbb{P}}[\phi_2^*\left(\phi_1^*(-X) + \theta\right)],$$

where the right-hand side is equal to $-\mathrm{OCE}_{v,\mathbb{P}}(u(X))$, for $u(x) \triangleq -\phi_1^*(-x)$ and $v(x) \triangleq -\phi_2^*(-x)$. Since identity (2.24) holds for the OCE, it follows that if $\phi_2^*(x) < x$, $\forall x \in \mathbb{R} \setminus \{0\}$, then we have that

$$\lim_{p \downarrow 0} \frac{\rho^l_{\mathrm{eu},\mathbb{P}}(X_p)}{p} = \phi_2^*(\phi_1^*(-x)). \tag{2.25}$$

In Proposition 2.19, we state that the same relation (2.25) can also be derived for the robust OCE risk measure $\rho^l_{\mathrm{oce},\mathbb{P}}$ and the robust shortfall risk measure $\rho^l_{\mathrm{sf},\mathbb{P}}$ through their dual formulation. Recall that the dual of $\rho^l_{\mathrm{oce},\mathbb{P}}$ is given in (2.7). The robust risk measure $\rho^l_{\mathrm{sf},\mathbb{P}}$, which is defined by replacing the expected utility constraint in (2.3) with its robust counterpart, has the following dual formulation:

$$\rho^l_{\mathrm{sf},\mathbb{P}}(X) = \inf_{\theta_1, \theta_2 \in \mathbb{R}} \left\{\theta_2 \mid \mathbb{E}_{\mathbb{P}}[\phi_2^*(\theta_1 - u(X + \theta_2))] \leq \theta_1\right\}. \tag{2.26}$$

Proposition 2.19. *Let $\phi_1, \phi_2 \in \Phi_0$ be lower-semicontinuous, twice continuously differentiable, and* $\mathrm{dom}(\phi_1^*) = \mathrm{dom}(\phi_2^*) = \mathbb{R}$. *Assume furthermore that $\phi_1^*(s) > s$, $\phi_2^*(s) > s$ for all $s \neq 0$. Then, for X_p defined in* (2.23) *with any $x \in \mathbb{R}$, we have*

$$\lim_{p \downarrow 0} \frac{\rho^l_{\mathrm{oce},\mathbb{P}}(X_p)}{p} = \phi_2^*(\phi_1^*(-x)), \qquad \lim_{p \downarrow 0} \frac{\rho^l_{\mathrm{sf},\mathbb{P}}(X_p)}{p} = \phi_2^*(\phi_1^*(-x)). \tag{2.27}$$

Therefore, Proposition 2.19 suggests a method for measuring the divergences ϕ_1^*, ϕ_2^*, by conducting the following experiment. One can first measure ϕ_1^*, by retrieving the certainty equivalent for a sequence of X_p, where p is known and tends to zero. In this case, the decision maker is not ambiguous and (2.24) gives a measurement of $u(x) = -\phi_1^*(-x)$. In a second experiment, we determine the robust certainty equivalent for X_p, where p is now hidden from the decision maker. This will yield $\phi_2^*(\phi_1^*(-x))$. Of course, this only allows us to recover the values of ϕ_2^* on the image set of ϕ_1^*. On the other hand, the values of ϕ_2^* outside this image set do not contribute to the values of the robust risk measure and can thus be taken arbitrarily.

Finally, we note that besides the elicitation of divergences, the formulation of the robust OCE and robust shortfall risk measures in terms of (ϕ_1^*, ϕ_2^*) also allows us to characterize risk aversion by simple convexity conditions imposed on ϕ_1^*, ϕ_2^*. We refer the readers to Section 2.D for more details.

2.6 Computing the Robust OCE Risk Measures

In this section, we discuss the computation of a general robust OCE risk measure using the sample average approximation (SAA). Theorem 2.1 states that robust OCE risk measures can be reformulated into the finite-dimensional dual problems (2.7) and (2.8), which are stochastic programming problems. The SAA method approximates these stochastic optimization problems by replacing $\mathbb{P}_0$ with the empirical distribution constructed from samples $X_1, \ldots, X_N$ generated from $\mathbb{P}_0$. This leads to the following optimization problems:

$$\min_{\theta_1, \theta_2 \in \mathbb{R}} -\theta_1 - \theta_2 + \frac{1}{N} \sum_{i=1}^{N} \phi_2^* \left(\phi_1^*(\theta_2 - X_i) + \theta_1 \right), \tag{2.28}$$

and

$$\min_{\substack{\lambda \geq 0 \\ \theta_1, \theta_2 \in \mathbb{R}}} -\theta_1 - \theta_2 + \lambda r + \frac{1}{N} \sum_{i=1}^{N} \lambda \phi_2^* \left(\frac{\phi_1^*(\theta_2 - X_i) + \theta_1}{\lambda} \right). \tag{2.29}$$

The formulations in (2.28) and (2.29) are convex optimization problems with only two or three variables. Therefore, they can be efficiently solved using the ellipsoid method (see, e.g., Bland et al. 1981). This is particularly useful if ϕ_1^* is only subdifferentiable, such

as for the CVaR_α risk measure where $\phi_1^*(x) = \max\{x/\alpha, 0\}$. If both ϕ_1^* and ϕ_2^* are conic representable, then (2.28) and (2.29) can also be solved using conic optimization software. Otherwise, one can consider using solvers such as CVXOPT and IPOPT. In practice, we observed that the ellipsoid method and conic solvers are the most promising ones for the problems (2.28) and (2.29).

The consistency of the SAA method, both in the objective value and the associated solutions, has been established in Theorem 5.4 of Shapiro et al. [2009], which requires the assumption that the optimal solutions of the true stochastic problems (2.7) and (2.8) are contained in a compact set, for any nominal distribution $\mathbb{P}$. The following proposition states that this is indeed the case.

Proposition 2.20. *Let $X \in L^1(\mathbb{P}_0)$ and $\phi_1, \phi_2 \in \Phi_0$ be lower semi-continuous. Suppose that condition* (2.10) *holds. Then, the decision variables θ_1, θ_2 in the dual problems* (2.7) *and* (2.8) *can be restricted to a compact set, without changing the optimum.*

Together with Proposition 2.20 and Theorem 5.4 of Shapiro et al. [2009], we can now state the consistency of the SAA method for robust OCE risk measures.

Theorem 2.21. *Let $X \in L^1(\mathbb{P}_0)$ and ϕ_1^*, ϕ_2^* be proper, non-decreasing convex functions that satisfy the following conditions:*

1. $\text{dom}(\phi_1^*) = \text{dom}(\phi_2^*) = \mathbb{R}$
2. $\phi_2^*(s) > s$, $\phi_1^*(s) > s$, *for all* $s \neq 0$ *and* $\phi_1^*(0) = \phi_2^*(0) = 0$.

Assume furthermore that there exists a $(\theta_1^0, \theta_2^0, \lambda_0)$ such that for all $(\theta_1, \theta_2, \lambda)$ in a neighborhood of $(\theta_1^0, \theta_2^0, \lambda_0)$, we have

$$\mathbb{E}_{\mathbb{P}_0}\left[\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2 - X) + \theta_1}{\lambda}\right)\right] < \infty.$$

Then, as the sample size $N \to \infty$, the SAA approximations (2.28) *and* (2.29) *converge both in the optimal objective value and in the optimal solutions to those of the true problems* (2.7) *and* (2.8)*, with probability 1.*

Although the SAA method converges theoretically, it might be very slow in practice due to a lack of data in the tails of the nominal distribution $\mathbb{P}_0$ of X. As noted by Smith and Winkler [2006], empirical risk minimization often underestimates the true risk. If we have a closed-form expression of the nominal density function f_0, then importance sampling can also be employed to improve the performance of SAA through a change of measure. For example, let g be the density function of a measure $\tilde{\mathbb{P}}_0$ such that $\mathbb{P}_0 \ll \tilde{\mathbb{P}}_0$, and let $Y_1, \ldots, Y_N$ be i.i.d. samples of $\tilde{\mathbb{P}}_0$. Then, through a multiplication of a likelihood

factor, we can also express the SAA of (2.29) as

$$\min_{\substack{\lambda\geq 0\\ \theta_1,\theta_2\in\mathbb{R}}} -\theta_1-\theta_2+\lambda r+\frac{1}{N}\sum_{i=1}^{N}\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2-Y_i)+\theta_1}{\lambda}\right)\frac{f(Y_i)}{g(Y_i)}. \tag{2.30}$$

Note that the likelihood factor $f(Y_i)/g(Y_i)$ does not depend on the decision variables, and thus does not affect the tractability of (2.30).

2.6.1 Addressing Sampling Errors and Model Misspecification Using Globalized Robust Optimization

Another way to address the underestimation of the true value of the robust OCE risk measure by SAA is to add an extra uncertainty set to the nominal probability from which the SAA of the robust OCE risk measure is sampled, using a third divergence ϕ_3. This is formulated as the following *globalized* robust OCE risk measure:

$$\rho^g_{\mathrm{oce},\mathbb{P}_0}(X)=\sup_{\{\mathbb{P}:I_{\phi_3}(\mathbb{P},\mathbb{P}_0)\leq r\}}\sup_{\{\mathbb{Q}:\mathbb{Q}\ll\mathbb{P}\}}\sup_{\{\bar{\mathbb{Q}}:\bar{\mathbb{Q}}\ll\mathbb{Q}\}}\mathbb{E}_{\bar{\mathbb{Q}}}[-X]-I_{\phi_1}(\bar{\mathbb{Q}},\mathbb{Q})-I_{\phi_2}(\mathbb{Q},\mathbb{P}). \tag{2.31}$$

We call (2.31) the globalized robust OCE risk measure since it utilizes the idea of globalized robust optimization introduced by Ben-Tal et al. [2017]. We note that (2.31) provides an upper bound on (2.5), due to the extra uncertainty set that is added to address the sampling error induced by the SAA approximation for computing a robust risk measure. In the following theorem, we show that (2.31) can also be reformulated as a finite-dimensional dual optimization problem.

Theorem 2.22. *Let $X\in L^1(\mathbb{P}_0)$ and $\phi_1,\phi_2,\phi_3\in\Phi_0$ be lower-semicontinuous. Then, we have the following duality:*

$$\rho^g_{\mathrm{oce},\mathbb{P}_0}(X)=\inf_{\substack{\lambda\geq 0\\ \theta_1,\theta_2,\theta_3\in\mathbb{R}}} -\theta_1-\theta_2-\theta_3+\lambda r+\mathbb{E}_{\mathbb{P}_0}\left[\gamma^*(X,\theta_1,\theta_2,\theta_3,\lambda)\right], \tag{2.32}$$

where $\gamma^*(X,\theta_1,\theta_2,\theta_3,\lambda)=\lambda\phi_3^*\left(\frac{\phi_2^*(\phi_1^*(\theta_3-X)+\theta_2)+\theta_1}{\lambda}\right)$.

We note that the globalized robust OCE risk measure also offers a new alternative for constructing uncertainty sets to address tail misspecification. Indeed, the duality relation (2.32) leads to the following integral condition that guarantees the finiteness of the globalized robust OCE risk measure under the nominal $\mathbb{P}_0$, namely that there must exist $\theta_1,\theta_2,\theta_3\in\mathbb{R}$ and $\lambda\geq 0$ such that

$$\int_{\mathbb{R}}\lambda\phi_3^*\left(\frac{\phi_2^*(\phi_1^*(\theta_3-x)+\theta_2)+\theta_1}{\lambda}\right)f_0(x)\mathrm{d}x<\infty. \tag{2.33}$$

The presence of an extra divergence ϕ_3^* in condition (2.33), compared to the non-globalized case (2.9), offers certain flexibility advantages. Instead of defining a single, more complicated divergence to tailor the tails of the conjugate functions and the nominal density, we can also use a composition of two elementary divergences $\phi_3^* \circ \phi_2^*$. For example, a composition of the KL-divergence $\phi_3^*(x) = O(e^x)$ and the polynomial divergence with $\phi_2^*(x) = O(x^{\tilde{p}})$ for some $\tilde{p} > 1$, would result in a tail behavior $\phi_3^*(\phi_2^*(-x)) = O(\exp(|x|^{\tilde{p}}))$ as $x \to -\infty$, which can be a suitable divergence for CVaR_α risk measures and a Weibull nominal distribution with shape parameter $k > 1$. The advantage of using a combination of a KL-divergence ϕ_3 and a polynomial divergence ϕ_2 is that both divergences have an explicit form of ϕ and conic representable ϕ^* functions.

2.7 Numerical Illustrations

In this section, we present three numerical experiments in which we utilize the results established in Section 2.3 to construct uncertainty sets for robust CVaR_α risk measures. The Python codes for each experiment are provided at https://github.com/GuanJinNL/Uncertainty-Sets-for-Robust-Risk-Measures.git.

2.7.1 Toy Examples

We first investigate a toy example where a payoff variable X follows a distribution supported on $[-\infty, -1)$ with density $g(x) = 2/|x|^3$ for $x \leq -1$ (i.e., $-X$ is a Pareto random variable with scale 1 and shape 2). We aim to measure its $\text{CVaR}_{0.975}$ at level 0.975, aligning with the Basel III regulation. The true $\text{CVaR}_{0.975}(X)$ value under g is equal to 12.649 (see Norton et al. 2021). Since the true distribution $g(x)$ is often unknown, we compute the robust $\text{CVaR}_{0.975}$, which we define as in (2.5) by means of a ϕ-divergence uncertainty set with a radius $r > 0$. We let the nominal distribution be $f_0(x) = 2.2/|x|^{3.2}$ for $x \leq -1$, which is slightly different from $g(x)$ and thus there is model misspecification. We investigate two polynomial divergences with $p = 3$ and $p = 11$, and their ability to address this model misspecification. We note that according to Table 2.1, both divergences will ensure a finite robust $\text{CVaR}_{0.975}$ evaluation. However, the polynomial divergence set with $p = 3$ contains $g(x)$ for r sufficiently large, whereas for $p = 11$ it does not contain $g(x)$ for any $r > 0$.[7] Therefore, we expect the robust $\text{CVaR}_{0.975}$ risk measure with $p = 11$ tends to underestimate the true $\text{CVaR}_{0.975}$.

For each radius r considered, we generate 10 times 500 samples from the nominal distribution f_0 and calculate the SAA of the robust $\text{CVaR}_{0.975}$ value for the two polynomial divergences with $p = 3$ and $p = 11$. The SAA of the robust $\text{CVaR}_{0.975}$ with polynomial

[7] Indeed, $g(x)/f_0(x) = O(|x|^{0.2})$. For $\phi_2(t) = O(t^p)$, $\int_\infty^{-1} \phi_2\left(g(x)/f_0(x)\right) f_0(x)dx < \infty$ if and only if $p < 11$.

divergence is given by the conic optimization problem (2.29), where $\phi_1^*(s) = \max\{s/(1-\alpha), 0\}$ and $\phi_2^*(s) = \max\{1 + s(p-1), 0\}^{\frac{p}{p-1}}/p - 1/p$. The results are displayed in Table 2.7.1. As we can observe, the polynomial divergence uncertainty set with $p = 3$ is able to account for model misspecification when $r > 0.02$, where the robust $\text{CVaR}_{0.975}$ value provides an upper estimate on the true $\text{CVaR}_{0.975}$. On the other hand, the polynomial divergence with $p = 11$ is too restrictive and still underestimates the true risk. This illustrates that a suitable uncertainty set needs to be sufficiently large to protect against model misspecification, yet still plausible.

Divergence Radius (r)	Robust $\text{CVaR}_{0.975}$ ($p = 3$)	Robust $\text{CVaR}_{0.975}$ ($p = 11$)
0.001	9.445	9.043
0.005	10.589	9.448
0.01	11.320	9.652
0.02	12.212	9.874
0.05	13.278	10.180
0.1	14.190	10.417

Table 2.4: SAA of Robust $\text{CVaR}_{0.975}$ risk measure with polynomial divergences of degrees $p = 3$ and $p = 11$, calculated for 500 samples drawn from the nominal f_0, for each radius r. The SAA values are taken as the median value over 10 repetitions, for each r.

In a second experiment, we investigate the effect on the SAA of the robust $\text{CVaR}_{0.975}$ risk measure when we choose a ϕ_2-divergence that does *not* ensure a finite evaluation. According to Table 2.1, we have an infinite evaluation under f_0 when we use the KL-divergence. Therefore, we calculate the SAA (2.29) for both the polynomial divergence with $p = 3$ and the KL-divergence across different sample sizes generated from f_0, where we fix $r = 0.05$. The results are shown in Table 2.5. As we can see, the SAA of the robust $\text{CVaR}_{0.975}$ risk measure induced by the KL-divergence provides a much more conservative upper estimation of the true $\text{CVaR}_{0.975}(X)$ than the one induced by the polynomial divergence. To make this difference even more visible, we use importance sampling to draw more data from the tail by sampling under the Pareto distribution $\tilde{g}(x) = 1/x^2$. The results are given in Table 2.6. Since we are sampling from the tails, it is expected that the values in Table 2.6 are larger than the ones in Table 2.5. However, in contrast to the polynomial divergence, the SAA of the robust $\text{CVaR}_{0.975}$ risk measure induced by the KL-divergence now provides overly and "implausibly" conservative upper bounds on the exact $\text{CVaR}_{0.975}(X)$ value. Moreover, it is sensitive to extreme values, as is displayed by the drastic increase in SAA values when the sample size changes from 2500 to 3000. This is because the robust $\text{CVaR}_{0.975}(X)$ risk measure under the KL-divergence is, in fact, infinite.

Sample Size	Robust $\text{CVaR}_{0.975}$ Polynomial	Robust $\text{CVaR}_{0.975}$ KL
500	11.388	14.765
1000	11.452	14.650
1500	12.748	18.349
2000	13.298	18.775
2500	13.250	19.264
3000	17.439	44.461
3500	17.202	43.343
4000	16.713	42.267
4500	16.611	41.507
5000	16.161	40.680
5500	16.306	40.219
6000	15.967	39.580

Table 2.5: SAA for the robust $\text{CVaR}_{0.975}$ risk measure with divergence radius $r = 0.05$, computed for a range of sample sizes drawn from the nominal distribution $f_0(x) = 2.2/|x|^{3.2},\ x \leq -1$. Results are displayed for ϕ_2 in (2.29) chosen to be the polynomial divergence with $p = 3$ and the KL-divergence.

Sample Size	Robust $\text{CVaR}_{0.975}$ Polynomial	Robust $\text{CVaR}_{0.975}$ KL
500	21.870	122.423
1000	21.471	116.577
1500	21.911	175.000
2000	22.271	178.624
2500	21.377	186.414
3000	21.548	1400.889
3500	21.693	1384.170
4000	21.649	1369.866
4500	21.666	1356.501
5000	21.669	1346.487
5500	21.703	1336.832
6000	21.796	1328.153

Table 2.6: SAA for the Robust $\text{CVaR}_{0.975}$ risk measure with divergence radius $r = 0.05$, computed for samples drawn using importance sampling under the sampling distribution $\tilde{g}(x) = 1/x^2,\ x \leq -1$. Results are displayed for ϕ_2 in (2.29) being the polynomial divergence with $p = 3$ and the KL-divergence.

2.7.2 Discrete Delta Hedging in the Black-Scholes Model

We consider a discrete delta hedging strategy for a call option under the Black-Scholes framework. In this model, the stock price S follows a geometric Brownian motion with drift μ_S and volatility $\sigma_S > 0$, i.e.:

$$dS_t/S_t = \mu_S dt + \sigma_S dW_t,$$

with W_t a standard Brownian motion. The call option with strike price K and a maturity time T has a final payoff given by $C(S_T) \triangleq \max\{S_T - K, 0\}$.

The discrete delta hedging strategy is implemented by constructing a self-financing portfolio that aims to optimally replicate the payoff of the call option. At each time step $t_i = iT/n$ (for some $n \in \mathbb{N}$ and $i \in \{1, \ldots, n\}$), the value of the portfolio is the sum of the stock value, Stock(i), and the cash value, Cash(i), that grows with an interest rate r_f. The portfolio is self-financing, which means that any rebalancing of the amount invested in one asset is only achieved by buying or selling the other asset. Let δ_{S,t_i} denote the number of shares of the stock held at time $[t_{i-1}, t_i)$, which we calculate using the Black-Scholes option pricing formula as in the continuous-time setting. The dynamics of the stock value and the cash value are determined by the following recursive relation:

$$\begin{aligned} \text{Stock}(t_i) &= S_{t_i} \cdot \delta_{S,t_i} \\ \text{Cash}(t_i) &= \exp\left(\frac{r_f}{n}\right) \text{Cash}(t_{i-1}) - S_{t_i} \cdot (\delta_{S,t_i} - \delta_{S,t_{i-1}}) \\ &\quad - k_0 - k \cdot |\delta_{S,t_i} - \delta_{S,t_{i-1}}| \cdot S_{t_i}, \end{aligned} \tag{2.34}$$

where we have added a transaction cost, $k_0 + k \cdot |\delta_{S,t_i} - \delta_{S,t_{i-1}}| \cdot S_{t_i}$ with $k_0, k > 0$, which consists of a fixed amount k_0 and a cost that is proportional to the volume of the trade executed. The initial values of Stock(0), Cash(0) are determined by the price of the call option at time zero according to the Black-Scholes option pricing formula. The hedging error, which we aim to measure, is defined as the absolute difference between the call option payoff and the payoff of this hedging strategy, evaluated at the maturity date T:

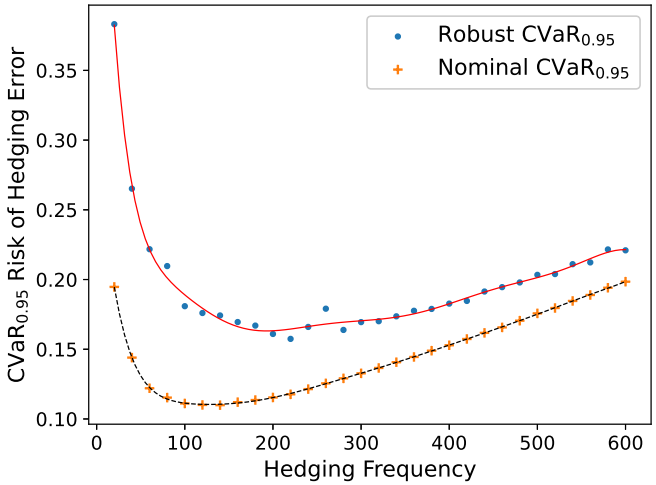
$$X_{\text{error}} \triangleq |C(S_T) - \text{Stock}(T) - \text{Cash}(T)|\,.$$

In the absence of transaction costs and with rebalancing occurring over infinitesimally small time intervals (i.e., in continuous time), the hedging error equals zero. In this experiment, we will measure the risk associated with the hedging error X_{error} both in the nominal (when the divergence radius $r = 0$) and the robust setting (when $r > 0$). Measurements of model risk of hedging error have been previously studied by Kruse et al. [2019] and Glasserman and Xu [2014]. In contrast to these studies, where they examined the expected hedging error, we measure the nominal and robust Conditional Value-at-Risk of X_{error} at level $\alpha = 0.95$.

To choose a proper ϕ_2-divergence for measuring ambiguity, we first note that $0 < \delta_{S_{t_i}} < 1$ by the Black-Scholes formula. Hence, according to the dynamics given in (2.34), one can show that there exists a constant $C > 0$, such that X_{error} satisfies (2.51):

$$|X_{\text{error}}| \leq C \left(1 + \sum_{i=1}^{n} S_{t_i}\right).$$

Figure 2.1: Absolute hedging error measured by the nominal and robust Conditional Value-at-Risk at confidence level $\alpha = 0.95$, as a function of the hedging frequency.



Therefore, we can choose the divergence according to the nominal distribution of the risk factor S_{t_i}, which, for each i, follows a log-normal distribution with $\sigma_i = i\sigma_S/\sqrt{n}$. Hence, we can set the volatility parameter of the divergence in (2.14) to be $\sigma = \max_i \sigma_i = \sigma_S$. Furthermore, we choose $d = 2$, so that our ϕ_2-divergence set may contain distributions with tails as heavy as $|x|^{-2}$.

As a comparison, we choose the same parameters as in Kruse et al. [2019], where $\mu_S = 0.05$, $\sigma_S = 0.3$, $r_f = 0.01$, $T = 1$, $S_0 = 1$, $K = 1$, $k_0 = 0.0002$ and $k = 0.005$. We vary the hedging frequency n and simulate for each frequency 50000 hedging error data points. Then, we calculate the SAA approximation of the $\text{CVaR}_{0.95}$ value of the hedging error, as well as that of the robust $\text{CVaR}_{0.95}$ value, with the radius in (2.5) set to be $r = 0.1$. The results are given in Figure 2.1. As we can see, both the nominal and the robust tail risk of the hedging error display a U-shape as a function of the hedging frequency, suggesting a tradeoff between transaction cost and hedging error. We see that the optimal hedging frequency in the nominal case is around 100, slightly larger than the case investigated in Kruse et al. [2019], where the expected hedging error was measured. In the robust case, the optimal frequency is increased to around 200, which is twice as large as in the nominal case. Moreover, the difference between the robust risk value and the nominal risk value is much larger at low hedging frequencies, and starts decreasing at higher frequencies until reaching near stability for frequencies higher than 400. This shows that for high hedging frequencies, the hedging portfolio is more robust against model uncertainty.

2.7.3 Risk-Averse Newsvendor Optimization

In the newsvendor problem, a stock owner must decide *a priori* about the amount of products that need to be ordered, before the demand is realized. Let d be a realized demand value of the uncertain demand variable D. Furthermore, we let c be the ordering cost per unit, $v > c$ be the selling price per unit, $s < c$ be the salvage value per unsold item returned to the factory, and l be the loss per unit of unmet demand. If the stock owner decides to order y number of items of the product, then the profit function is given by:

$$\pi(y, d) \triangleq v \min\{d, y\} + s \max\{y - d, 0\} - l \max\{d - y, 0\} - cy.$$

In the classical newsvendor problem, one minimizes the expected loss, $\mathbb{E}[-\pi(y, D)]$, with respect to the choice variable y. In a risk-averse setting, the expectation operator is replaced, and a risk measure such as Conditional Value-at-Risk, $\text{CVaR}_\alpha(\pi(y, D))$, is minimized, where α controls the degree of risk aversion of the decision maker. In the absence of ambiguity, the risk-averse newsvendor CVaR minimization problem admits a closed-form solution (see Gotoh and Takano 2007), which is given by:

$$y^*_{\text{nom}} = \frac{E+V}{E+U} F^{-1}\left(\frac{U(1-\alpha)}{E+U}\right) + \frac{U-V}{E+U} F^{-1}\left(\frac{E\alpha+U}{E+U}\right), \tag{2.35}$$

where F^{-1} is the quantile function of the random demand variable, and $E \triangleq c - s$, $U \triangleq v + l - c$, $V \triangleq v - c$.

We investigate the effect of model uncertainty on the optimal number of items ordered, by solving the following *robust* $\text{CVaR}_{0.95}(\pi(y, D))$ newsvendor minimization problem using the ellipsoid method, for $N = 50{,}000$:

$$\min_{\substack{\lambda \geq 0 \\ y, \theta_1, \theta_2 \in \mathbb{R}}} \; -\theta_1 - \theta_2 + \lambda r + \frac{1}{N} \sum_{i=1}^{N} \lambda \phi_2^* \left(\frac{\phi_1^*(\theta_2 - \pi(y, D_i)) + \theta_1}{\lambda} \right), \tag{2.36}$$

where $D_1, \ldots, D_{50000}$ are demand samples drawn from either the heavy-tailed log-normal distribution with parameters $\mu_D = 0$ and $\sigma_D = 1$, or the Weibull distribution with shape parameter $k = 1/2$ and scale parameter $\lambda = 1$. Moreover, we have $\phi_1^*(s) = \max\{s/(1-\alpha), 0\}$, with $\alpha = 0.95$.

We solve (2.36) for a range of radii $r > 0$, where a larger r represents a higher degree of model uncertainty. To choose a proper ϕ_2-divergence, we note that $\pi(y, D)$ is a piecewise-linear concave function in D, and thus $|\pi(y, D)|$ also satisfies (2.51) for some constant $C > 0$.[8] Hence, we may choose the ϕ_2-divergence according to the nominal

[8] Specifically, for $C = \max\{l, |v - s|\} \cdot \max\left\{1, \frac{\max\{|v+l-c|, |s-c|\}}{\max\{l, |v-s|\}} |y|\right\}$.

distribution of the demand D. If the nominal distribution is log-normal, then we choose (2.14) with $\sigma = \sigma_D = 1$, $p = 2$ and $d = 2$ similarly as in the Black-Scholes example of the previous subsection. If the nominal distribution is Weibull, then we choose the divergence (2.15) with the shape parameter k and $d = 2$. For the remaining parameters, we choose $v = 8,\ c = 4,\ s = 2$ and $l = 4$. Then, one can calculate using (2.35) that the optimal decision without model uncertainty under the log-normal and the Weibull nominal distributions are given by $y^*_{\text{nom}} = 4.2$ and $y^*_{\text{nom}} = 8.5$. The robust results are displayed in Table 2.7, which show that a higher degree of model uncertainty increases the optimal number of items ordered. This is because for each unit of unmet demand, the stock owner pays a price of $l(y - d)$. When the demand is heavy-tailed, a large loss in profit can occur when extremely large demand values are realized. Therefore, the optimal number of items ordered in the robust newsvendor problem (2.36) is larger than the nominal solution (2.35), since it must hedge against the potential large loss due to unmet demand.

Nominal Distribution: Log-normal($\mu_D = 0,\ \sigma_D = 1$)

Divergence radius r	Optimal Number of Items Ordered (y)
0.001	4.514
0.005	5.093
0.01	5.488
0.02	6.168
0.05	7.557
0.1	8.894
0.2	11.506
0.4	16.244
0.5	19.115

Nominal Distribution: Weibull ($k = 1/2$, $\lambda = 1$)

Divergence radius r	Optimal Number of Items Ordered (y)
0.001	9.796
0.005	11.157
0.01	12.094
0.02	13.262
0.05	15.350
0.1	17.590
0.2	20.943
0.4	25.776
0.5	26.949

Table 2.7: The effect of model uncertainty on the optimal number of items ordered in the robust newsvendor optimization problem (2.36). The larger the parameter $r > 0$, the higher the degree of model uncertainty. The demand samples $D_1, \ldots, D_{50000}$ are drawn from either the heavy-tailed log-normal($\mu_D = 0$,$\sigma_D = 1$) distribution, or the Weibull distribution ($k = 1/2$, $\lambda = 1$). The divergence ϕ_1 is $\delta_{[0,1/(1-\alpha)]}$ with $\alpha = 0.95$. The divergence ϕ_2 is given by (2.14) with $\sigma = \sigma_D = 1$, $p = 2$ and $d = 2$, when the nominal distribution is log-normal. When the nominal distribution is Weibull, the divergence ϕ_2 is given by (2.15) with $d = 2$. The remaining parameters of the newsvendor problem are given by $v = 8,\ c = 4,\ s = 2$ and $l = 4$.

2.8 Concluding Remarks

In this paper, we have proposed a unifying framework to construct suitable uncertainty sets for robust risk measures. It parametrizes a broad class of robust risk measures in terms of two ϕ-divergences, (ϕ_1, ϕ_2), where ϕ_1 specifies the risk measure and ϕ_2 specifies the ambiguity set. We have shown that this framework is not only computationally tractable, but also offers a blueprint to systematically construct ϕ-divergence uncertainty sets that are tailored to a given risk measure and nominal model. In addition, this framework extends naturally to include higher levels of uncertainty using globalized robust optimization. For many robust risk measures, we can also elicit the divergence functions (ϕ_1, ϕ_2), which serves as an extra tool for calibrating the uncertainty set.

Appendix

2.A Proofs

Proof of Theorem 2.1. Due to the absolute continuity relations $\bar{\mathbb{Q}} \ll \mathbb{Q} \ll \mathbb{P}_0$, there exist probability density functions $g, \bar{g}$ with respect to the measure $\mathbb{P}_0$, such that the optimization problem:

$$\sup_{\substack{\{\mathbb{Q}:\mathbb{Q}\ll\mathbb{P}_0\}\\ \{\bar{\mathbb{Q}}:\bar{\mathbb{Q}}\ll\mathbb{Q}\}}} \mathbb{E}_{\bar{\mathbb{Q}}}[-X] - I_{\phi_1}(\bar{\mathbb{Q}}, \mathbb{Q}) - I_{\phi_2}(\mathbb{Q}, \mathbb{P}_0),$$

can be formulated as an optimization problem over density functions:

$$-\inf_{\substack{g,\bar{g}\geq 0,\\ \int_\Omega g \mathrm{d}\mathbb{P}_0=1, \int_\Omega \bar{g}\mathrm{d}\mathbb{P}_0=1}} \int_\Omega \beta(\omega, g(\omega), \bar{g}(\omega))\mathrm{d}\mathbb{P}_0(\omega),$$

where

$$\beta(\omega, s, t) \triangleq X(\omega)t + s\phi_1\left(\frac{t}{s}\right) + \phi_2(s), \qquad s, t \geq 0.$$

Note that $\beta(\omega, s, t) = +\infty$ for $s < 0$ or $t < 0$ since both ϕ_1 and ϕ_2 take the value $+\infty$ on the negative real line. By assumption, ϕ_1 and ϕ_2 are convex lower-semicontinuous functions that have an effective domain with non-empty interior. Therefore, $\beta(\omega, s, t)$ is a normal convex integrand (in the sense of Lemma 2 of Rockafellar 1968 and Proposition 14.39 of Rockafellar and Wets 2004). Furthermore, there exists a feasible point, namely the constant function $(g, \bar{g}) \equiv (1, 1)$: $\int_\Omega \beta(\omega, 1, 1)\mathrm{d}\mathbb{P}_0(\omega) = \mathbb{E}_{\mathbb{P}_0}[X] + \phi_1(1) + \phi_2(1) < \infty$.

Therefore, applying Theorem 2.B.3 yields

$$\inf_{\substack{g,\bar{g}\geq 0,\\ \int_\Omega g\mathrm{d}\mathbb{P}_0=1,\int_\Omega \bar{g}\mathrm{d}\mathbb{P}_0=1}} \int_\Omega \beta(\omega, g(\omega), \bar{g}(\omega))\mathrm{d}\mathbb{P}_0(\omega) = \sup_{\theta_1,\theta_2\in\mathbb{R}} \theta_1+\theta_2 - \int_\Omega \beta^*(\omega,\boldsymbol{\theta})\mathrm{d}\mathbb{P}_0(\omega),$$

where

$$\begin{aligned}
\beta^*(\omega,\boldsymbol{\theta}) &\triangleq \sup_{s,t\geq 0} \theta_1 s+\theta_2 t - X(\omega)t - s\phi_1\left(\frac{t}{s}\right) - \phi_2(s) \\
&= \sup_{s\geq 0} \theta_1 s - \phi_2(s) + \sup_{t\geq 0}(\theta_2 - X(\omega))t - s\phi_1\left(\frac{t}{s}\right) \\
&= \sup_{s\geq 0} \theta_1 s - \phi_2(s) + s\phi_1^*(\theta_2 - X(\omega)) \\
&= \phi_2^*(\theta_1 + \phi_1^*(\theta_2 - X(\omega))).
\end{aligned}$$

Finally, to obtain the worst-case density, we have that $\mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*(\phi_1^*(\theta_2 - X)+\theta_1)\right] < \infty$ for some (θ_1,θ_2). Hence, the dual problem (2.7) is finite and there exists a dual solution (θ_1^*,θ_2^*) by Theorem 2.B.2. If furthermore $(\theta_1^*,\theta_2^*)\in \mathrm{int}(\mathcal{F}(\boldsymbol{\theta}))$, then $\mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*(\phi_1^*(\theta_2-X)+\theta_1)\right]$ is differentiable at (θ_1^*,θ_2^*) by Theorem 2.B.4. Therefore, the partial derivatives $(g^*(\omega),\bar{g}^*(\omega))$ $\triangleq (\partial_{\theta_1}\beta^*(\omega,\boldsymbol{\theta}),\partial_{\theta_2}\beta^*(\omega,\boldsymbol{\theta}))$ evaluated at $\boldsymbol{\theta}^*$, for each $\omega\in\Omega$, satisfy the first-order condition of the dual problem (2.7):

$$\int_\Omega g^*(\omega)\mathrm{d}\mathbb{P}_0(\omega) = 1, \quad \int_\Omega \bar{g}^*(\omega)\mathrm{d}\mathbb{P}_0(\omega) = 1.$$

Hence, it follows from Theorem 2.B.5 that the worst-case densities are given by the partial derivatives.

Similarly, the optimization problem

$$\sup_{\{\mathbb{Q}:I_{\phi_2}(\mathbb{Q},\mathbb{P}_0)\leq r\}} \sup_{\{\bar{\mathbb{Q}}:\bar{\mathbb{Q}}\ll\mathbb{Q}\}} \mathbb{E}_{\bar{\mathbb{Q}}}[-X] - I_{\phi_1}(\bar{\mathbb{Q}},\mathbb{Q})$$

can be rewritten as the following optimization problem over density functions with respect to the measure $\mathbb{P}_0$:

$$-\inf_{\substack{g,\bar{g}\geq 0,\\ \int_\Omega g\mathrm{d}\mathbb{P}_0=1,\int_\Omega \bar{g}\mathrm{d}\mathbb{P}_0=1\\ \int_\Omega \phi_2(g(\omega))\mathrm{d}\mathbb{P}_0(\omega)\leq r}} \int_\Omega X(\omega)\bar{g}(\omega) + g(\omega)\phi_1\left(\frac{\bar{g}(\omega)}{g(\omega)}\right)\mathrm{d}\mathbb{P}_0(\omega) \triangleq -J_0. \tag{2.37}$$

Assume without loss of generality that $J_0 > -\infty$. Then, we also have that $J_0 < \infty$, since $(g,\bar{g})\equiv(1,1)$ is a feasible solution. Hence, Theorem 2.B.2 implies that

$$J_0 = \sup_{\lambda\geq 0} -\lambda r + \inf_{\substack{g,\bar{g}\geq 0,\\ \int_\Omega g\mathrm{d}\mathbb{P}_0=1,\int_\Omega \bar{g}\mathrm{d}\mathbb{P}_0=1}} \int_\Omega X(\omega)\bar{g}(\omega) + g(\omega)\phi_1\left(\frac{\bar{g}(\omega)}{g(\omega)}\right) + \lambda\phi_2(g(\omega))\mathrm{d}\mathbb{P}_0(\omega).$$

We first examine the case when $\lambda = 0$, for which the above would become

$$\begin{aligned} J_0 &= \inf_{\substack{g,\bar{g}\geq 0,\\ \int_\Omega g \mathrm{d}\mathbb{P}_0 = 1, \int_\Omega \bar{g}\mathrm{d}\mathbb{P}_0 = 1}} \int_\Omega X(\omega)\bar{g}(\omega) + g(\omega)\phi_1\left(\frac{\bar{g}(\omega)}{g(\omega)}\right)\mathrm{d}\mathbb{P}_0(\omega) \\ &= \operatorname{ess\,inf}(X). \end{aligned}$$

On the other hand, since $0\phi_2^*(s/0) = 0$ for $s \leq 0$ and $+\infty$ otherwise, we have

$$\begin{aligned} &\sup_{\theta_1,\theta_2\in\mathbb{R}} \theta_1 + \theta_2 - \mathbb{E}_{\mathbb{P}_0}\left[0\phi_2^*\left(\frac{\phi_1^*(\theta_2 - X) + \theta_1}{0}\right)\right] \\ &= \sup_{\theta_1,\theta_2\in\mathbb{R}} \{\theta_1 + \theta_2 \mid \operatorname{ess\,sup}\phi_1^*(\theta_2 - X) + \theta_1 \leq 0\} \\ &= \sup_{\theta_2\in\mathbb{R}} \theta_2 - \phi_1^*(\theta_2 - \operatorname{ess\,inf}(X)) \\ &= \operatorname{ess\,inf}(X). \end{aligned}$$

Hence, the duality formula holds in the case of $\lambda = 0$. Assume now $\lambda > 0$. Define $\beta_\lambda(\omega, s, t) \triangleq X(\omega)t + s\phi_1\left(\frac{t}{s}\right) + \lambda\phi_2(s)$. Note that β_λ is only finite when $s, t \geq 0$. Then, applying Theorem 2.B.2 yields

$$J_0 = \sup_{\substack{\lambda\geq 0\\ \theta_1,\theta_2\in\mathbb{R}}} -\lambda r + \theta_1 + \theta_2 - \int_\Omega \beta_\lambda^*(\omega, \theta_1, \theta_2)\mathrm{d}\mathbb{P}_0(\omega).$$

It remains to compute $\beta_\lambda^*(\omega, \theta_1, \theta_2)$, which is equal to $\beta_\lambda^*(\omega, \theta_1, \theta_2) = \lambda\phi_2^*\left(\frac{\theta_1 + \phi_1^*(\theta_2 - X(\omega))}{\lambda}\right)$.

Now for the worst-case densities, we note that the domain set $\mathcal{F}(\boldsymbol{\theta}, \lambda)$ is assumed to be non-empty, thus a dual solution $(\theta_1^*, \theta_2^*, \lambda^*)$ exists by Theorem 2.B.2. If the dual solution also lies in the interior of $\mathcal{F}(\boldsymbol{\theta}, \lambda)$, then we have differentiability by Theorem 2.B.4, and thus the dual solution must satisfy the gradient conditions

$$\begin{aligned} &\int_\Omega (\phi_2^*)'\left(\frac{\theta_1^* + \phi_1^*(\theta_2^* - X(\omega))}{\lambda^*}\right)\mathrm{d}\mathbb{P}_0(\omega) = 1 \\ &\int_\Omega (\phi_2^*)'\left(\frac{\theta_1^* + \phi_1^*(\theta_2^* - X(\omega))}{\lambda^*}\right)(\phi_1^*)'(\theta_2^* - X(\omega))\mathrm{d}\mathbb{P}_0(\omega) = 1 \\ &\int_\Omega \frac{\theta_1^* + \phi_1^*(\theta_2^* - X(\omega))}{\lambda^*}(\phi_2^*)'\left(\frac{\theta_1^* + \phi_1^*(\theta_2^* - X(\omega))}{\lambda^*}\right) - \phi_2^*\left(\frac{\theta_1^* + \phi_1^*(\theta_2^* - X(\omega))}{\lambda^*}\right)\mathrm{d}\mathbb{P}_0(\omega) = r, \end{aligned}$$

where the third condition is equivalent to $\int_\Omega \phi_2\left((\phi_2^*)'\left(\frac{\theta_1^* + \phi_1^*(\theta_2^* - X(\omega))}{\lambda^*}\right)\right)\mathrm{d}\mathbb{P}_0(\omega) = r$, due to the relation $\phi_2 = \phi_2^{**}$. Therefore, the partial derivatives $(g^*(\omega), \bar{g}^*(\omega)) \triangleq (\partial_{\theta_1}\beta_\lambda^*(\omega, \boldsymbol{\theta}), \partial_{\theta_2}\beta_\lambda^*(\omega, \boldsymbol{\theta}))$ evaluated at $(\boldsymbol{\theta}^*, \lambda^*)$ are indeed feasible density functions of the primal problem (2.37).

Finally, the partial derivatives are indeed also the worst-case densities, since

$$\begin{aligned} J_0 &= -\lambda^* r + \theta_1^* + \theta_2^* - \int_\Omega \beta_{\lambda^*}^*(\omega, \theta_1^*, \theta_2^*) \mathrm{d}\mathbb{P}_0(\omega) \\ &\overset{(*)}{=} -\lambda^* r + \int_\Omega X(\omega)\bar{g}^*(\omega) + g^*(\omega)\phi_1\left(\frac{\bar{g}^*(\omega)}{g^*(\omega)}\right) + \lambda^*\phi_2(g^*(\omega))\mathrm{d}\mathbb{P}_0(\omega) \\ &= \int_\Omega X(\omega)\bar{g}^*(\omega) + g^*(\omega)\phi_1\left(\frac{\bar{g}^*(\omega)}{g^*(\omega)}\right)\mathrm{d}\mathbb{P}_0(\omega), \end{aligned}$$

where $(*)$ follows from the worst-case densities of (2.6). □

Proof of Corollary 2.2. Recall that from Theorem 2.1 we have

$$\sup_{\{\mathbb{Q}:\mathbb{Q}\ll\mathbb{P}_0\}} \mathbb{E}_{\bar{\mathbb{Q}}}[-X] - I_{\phi_1}(\bar{\mathbb{Q}}, \mathbb{Q}) - I_{\phi_2}(\mathbb{Q}, \mathbb{P}_0) = \inf_{\theta_1,\theta_2\in\mathbb{R}} -\theta_1 - \theta_2 + \mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*(\phi_1^*(\theta_2 - X) + \theta_1)\right]. \tag{2.38}$$

Hence, if $\mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*(\phi_1^*(\theta_2 - X) + \theta_1)\right] = \infty$ holds for all θ_1, θ_2, then the infimum on the right-hand side of (2.38) is indeed always ∞. If $\mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*(\phi_1^*(\theta_2 - X) + \theta_1)\right] < \infty$ for some $\theta_1, \theta_2 \in \mathbb{R}$, then this infimum is bounded away from $-\infty$. Similarly, (2.9) follows from (2.8). □

Proof of Proposition 2.8. By the assumptions on ψ, we have that $\phi^*(s)$ as defined in (2.12) is an increasing, continuous convex function on $\mathbb{R}$ with $\phi^*(0) = 0$ and $(\phi^*)'(0) = 1$. By definition, we have that $\phi(t) \triangleq \sup_{s\in\mathbb{R}} st - \phi^*(s)$. Therefore, ϕ is convex. We also have that $\phi(t) = +\infty$ for $t < 0$, since $\phi^*(s) < 0$ for $s < 0$ and thus $\phi(t) \geq \sup_{s\leq 0} st = \infty$, if $t < 0$. Moreover, ϕ is non-negative on $[0, \infty)$ with $\phi(1) = 0$. Indeed, $\phi^*(0) = 0$ ensures that $\phi(t) \geq 0 \cdot t - \phi^*(0) = 0$, and $(\phi^*)'(0) = 1$ implies that $\phi(1) = 0$ (Theorem 23.5, Rockafellar 1970). Therefore, $\phi \in \Phi_0$ and is thus a ϕ-divergence function.

If ψ is strictly convex and differentiable, then ϕ^* is, by construction, strictly convex and differentiable on $\mathbb{R}$. Theorem 26.3 of Rockafellar [1970] implies that its conjugate $\phi \triangleq (\phi^*)^*$ is essentially smooth. Therefore, ϕ is differentiable on the interior of its effective domain, denoted as $\mathrm{dom}(\phi)$. By Corollary 23.5.1 and Theorem 23.4 of Rockafellar [1970], we have that $\mathrm{dom}(\phi)$ contains the image of the derivative $(\phi^*)'$, which is the set $(0, \infty)$, since by assumption ψ' is continuous and strictly increasing with $\lim_{s\to\infty}\psi'(s) = \infty$, and that $(\phi^*)'(s) = \exp(s)$, for $s \leq 0$. Furthermore, $\phi(0) = 1 < \infty$, since $\phi(t) = t\log t - t + 1$ on $t \leq 1$. Hence, $\mathrm{dom}(\phi) = [0, \infty)$. Moreover, Corollary 23.5.1 of Rockafellar [1970] implies that $\phi' = ((\phi^*)')^{-1}$. □

The proof of Proposition 2.9 relies on the following Lemma.

Lemma 2.A.1. *Let f, g be two strictly increasing functions. If for some $y_0 \in \mathbb{R}$ that $f^{-1}(y) \leq g^{-1}(y)$ for all $y \geq y_0$, then $f(x) \geq g(x)$, for all $x \geq f^{-1}(y_0)$.*

Proof of Lemma 2.A.1. We prove by contradiction. Suppose there is a $x_0 \geq f^{-1}(y_0)$ such that $f(x_0) < g(x_0)$. Then, $f(x_0) \geq y_0$. Since f, g are increasing, we have that f^{-1}, g^{-1} are also increasing. Therefore,

$$x_0 = f^{-1}(f(x_0)) \leq g^{-1}(f(x_0)) < g^{-1}(g(x_0)) = x_0,$$

which is a contradiction. □

Proof of Proposition 2.9. Since ϕ_2' is strictly increasing and non-negative for $t \geq 1$, we have that for $t \geq 1$,

$$\phi_2(t) = \int_1^t \phi_2'(s)\mathrm{d}s \leq (t-1)\phi_2'(t) \leq t\phi_2'(t).$$

Since $(\psi_1')^{-1}(y) \leq (\phi_2')^{-1}(y)$ for all $y \geq y_0$, Lemma 2.A.1 implies that for all $t \geq t_0$, where $t_0 \triangleq \max\{(\psi_1')^{-1}(y_0), 1\}$, we have that $\psi_1'(t) \geq \phi_2'(t) \geq 0$. Therefore, we have

$$\begin{aligned}
&\int_{-\infty}^{\infty} \phi_2\left(\frac{g(x)}{f_0(x)}\right) f_0(x)\mathrm{d}x \\
&= \int_{x:\frac{g(x)}{f_0(x)}\leq 1} \phi_2\left(\frac{g(x)}{f_0(x)}\right) f_0(x)\mathrm{d}x + \int_{x:\frac{g(x)}{f_0(x)}>1} \phi_2\left(\frac{g(x)}{f_0(x)}\right) f_0(x)\mathrm{d}x \\
&\leq \int_{x:\frac{g(x)}{f_0(x)}\leq 1} \phi_2\left(\frac{g(x)}{f_0(x)}\right) f_0(x)\mathrm{d}x + \int_{x:\frac{g(x)}{f_0(x)}>1} g(x)\phi_2'\left(\frac{g(x)}{f_0(x)}\right)\mathrm{d}x.
\end{aligned}$$

Since ϕ_2 is bounded on $[0,1]$, we only need to bound the second integral. By the limit assumption (2.13), there exists a u_0, such that for all $x \in \mathbb{R}$ with $|x| \geq u_0$, we have $\psi_2'\left(\frac{1}{f_0(x)}\right) \leq M|x|^d$, for some constant $M > 0$. We have that $f_0(x) \to 0$ as $|x| \to \infty$. Hence, there exists a $x_0 \geq u_0$ such that for all $x \in \mathbb{R}$ with $|x| \geq x_0$, we have that $\frac{1}{f_0(x)} \geq t_0$. We then split the following integral by x_0:

$$\begin{aligned}
&\int_{x:\frac{g(x)}{f_0(x)}>1} g(x)\phi_2'\left(\frac{g(x)}{f_0(x)}\right)\mathrm{d}x \\
&= \int_{x:\frac{g(x)}{f_0(x)}>1\cap\{x:|x|\leq x_0\}} g(x)\phi_2'\left(\frac{g(x)}{f_0(x)}\right)\mathrm{d}x + \int_{x:\frac{g(x)}{f_0(x)}>1\cap\{x:|x|\geq x_0\}} g(x)\phi_2'\left(\frac{g(x)}{f_0(x)}\right)\mathrm{d}x.
\end{aligned}$$

The first term is bounded by the assumption that $\frac{g}{f}$ is bounded on any compact interval, and that ϕ_2' is bounded on any compact interval as well, due to its continuity. Therefore, we only need to examine the second term. Due to the non-negativity and monotonicity

of ϕ_2', ψ_2' on $[1, \infty)$, we have,

$$\begin{aligned}\int_{x:\frac{g(x)}{f_0(x)}>1\cap\{x:|x|\geq x_0\}} g(x)\phi_2'\left(\frac{g(x)}{f_0(x)}\right)\mathrm{d}x &\leq \int_{x:|x|\geq x_0} g(x)\phi_2'\left(\frac{g(x)}{f_0(x)}\right)\mathrm{d}x \\ &\leq \int_{x:|x|\geq x_0} g(x)\phi_2'\left(\frac{1}{f_0(x)}\right)\mathrm{d}x \\ &\leq \int_{x:|x|\geq x_0} g(x)\psi_1'\left(\frac{1}{f_0(x)}\right)\mathrm{d}x \\ &\leq \int_{x:|x|\geq x_0} g(x)|x|^d\mathrm{d}x < \infty.\end{aligned}$$

Conversely, if there exists a divergence ψ_2 such that $(\phi_2^*)'(y) \leq (\psi_2')^{-1}(y)$ for all $y \geq 0$, then we have that $\phi_2'(t) \geq \psi_2'(t)$ for all $t \geq 1$. Since $\phi_2(1) = \psi_2(1) = 0$, this also means that

$$\phi_2(t) = \int_1^t \phi_2'(s)\mathrm{d}s \geq \int_1^t \psi_2'(s)\mathrm{d}s = \psi_2(t), \ \forall t \geq 1.$$

Therefore, we have

$$\int_{x:\frac{g(x)}{f_0(x)}>1} \phi_2\left(\frac{g(x)}{f_0(x)}\right) f_0(x)\mathrm{d}x \geq \int_{x:\frac{g(x)}{f_0(x)}>1} \psi_2\left(\frac{g(x)}{f_0(x)}\right) f_0(x)\mathrm{d}x.$$

Since ψ_2 is bounded on $[0, 1]$, we have that $I_{\psi_2}(g, f) = \infty$ if and only if

$$\int_{x:\frac{g(x)}{f_0(x)}>1} \psi_2\left(\frac{g(x)}{f_0(x)}\right) f_0(x)\mathrm{d}x = \infty,$$

which implies that $I_{\phi_2}(g, f) = \infty$. □

Proof of Lemma 2.11. Following the construction in (2.12), we choose $\psi(s) = (s + e)\exp(a\log^p(s + e))$, where $a = 1/(p\,(\sigma d)^p)$. Clearly, $\psi(s)$ is increasing and twice continuously differentiable on $s \geq 0$. Moreover, it is strictly convex, since its second derivative is given by

$$\psi''(s) = \frac{ap\log^{p-2}(s + e)\exp(a\log^p(s + e))(ap\log^p(s + e) + p + \log(s + e) - 1)}{s + e},$$

which is strictly positive for all $s \geq 0$. Hence, ϕ_2^* as defined in (2.14) satisfies all properties given in Proposition 2.8.

We now show that the conjugate function ϕ_2^* satisfies the condition (2.10), which also implies (2.9). For any $\theta_1 \geq 0, \theta_2 \in \mathbb{R}$, we have that $\phi_1^*(\theta_2 - x) = \max\{(\theta_2 - x)/\alpha, 0\}$. Since $f_0(x)$ has support $(-\infty, 0]$, we have that $(\theta_2 - x)/\alpha \geq 0$, if $x \leq x_0$ for some x_0.

Therefore, for all $x \leq x_0$, we have

$$\phi_2^*\left(\phi_1^*(\theta_2 - x) + \theta_1\right) f_0(x) = \phi_2^*\left(\frac{1}{\alpha}(\theta_2 - x) + \theta_1\right) \frac{1}{C\sigma|x|} \exp\left\{-\frac{1}{p\sigma^p}(\log(|x|) - \mu)^p\right\}.$$

We analyze the asymptotic behavior of the above function. We have that

$$\begin{aligned}
&\phi_2^*\left(\frac{1}{\alpha}(\theta_2 - x) + \theta_1\right) \frac{1}{C\sigma|x|} \exp\left\{-\frac{1}{p\sigma^p}(\log|x| - \mu)^p\right\} \\
&= \frac{c_1\left(\frac{1}{\alpha}(\theta_2 + |x|) + \theta_1 + e\right)}{C\sigma|x|} \cdot \exp\left(\frac{1}{p(\sigma d)^p}\log^p\left(\frac{1}{\alpha}(\theta_2 + |x|) + \theta_1 + e\right) - \frac{1}{p\sigma^p}(\log|x| - \mu)^p\right) \\
&\quad + \left(\frac{c_2\left(\frac{1}{\alpha}(\theta_2 + |x|) + \theta_1\right) + c_3}{C\sigma|x|}\right) \cdot \exp\left(-\frac{1}{p\sigma^p}(\log|x| - \mu)^p\right) \\
&= O(1) \cdot \exp\left(\frac{1}{p(\sigma d)^p}\log^p\left(\frac{1}{\alpha}(\theta_2 + |x|) + \theta_1 + e\right) - \frac{1}{p\sigma^p}(\log|x| - \mu)^p\right) + \\
&\quad O(1) \cdot \exp\left(-\frac{1}{p\sigma^p}(\log(|x|) - \mu)^p\right).
\end{aligned}$$

We only need to examine that,

$$\begin{aligned}
&\exp\left(\frac{1}{p(\sigma d)^p}\log^p\left(\frac{1}{\alpha}(\theta_2 + |x|) + \theta_1 + e\right) - \frac{1}{p\sigma^p}(\log|x| - \mu)^p\right) \\
&= \exp\left(\frac{1}{p(\sigma d)^p}\log^p\left(|x|\left(\frac{1}{\alpha} + \frac{\frac{\theta_2}{\alpha} + \theta_1 + e}{|x|}\right)\right) - \frac{1}{p\sigma^p}\log^p|x|(1 + o(1))^p\right) \\
&= \exp\left(\frac{1}{p(\sigma d)^p}\log^p|x|\,(1 + o(1))^p - \frac{1}{p\sigma^p}\log^p|x|(1 + o(1))^p\right) \\
&= \exp\left(-\frac{1}{p\sigma^p}\log^p|x|\left(\frac{1}{d^p}(1 + o(1))^p - (1 + o(1))^p\right)\right).
\end{aligned}$$

Therefore, the above expression has the same asymptotic as $\exp\left(-\frac{1}{p\sigma^p}\log^p|x|\left(\frac{1}{d^p} - 1\right)\right)$. Since $d > 1, p \geq 2 \Rightarrow 1/d^p - 1 < 0$, we have that (2.10) is satisfied.

Finally, to show that the divergence ball of ϕ_2 contains distributions that have finite d-th moment. First, we examine the derivative of ϕ_2^*, which is also equal to $(\phi_2')^{-1}$:

$$(\phi_2^*)'(s) = \begin{cases} c_1 \exp\left\{\frac{1}{p(\sigma d)^p}\log^p(s + e)\right\}\left(\frac{1}{(\sigma d)^p}\log^{p-1}(s + e) + 1\right) + c_2 & s > 0 \\ \exp(s) & s \leq 0. \end{cases}$$

We define a function

$$(\psi_2')^{-1}(s) = \begin{cases} c_1 \exp\left\{\frac{1}{p(\sigma d)^p}\log^p(s + e)\right\}\left(\frac{1}{(\sigma d)^p} + 1\right) + c_2 & s > 0 \\ \exp(s) & s \leq 0. \end{cases}$$

We note that $(\phi_2^*)'(s) \geq (\psi_2')^{-1}(s)$, for all $s \in \mathbb{R}$. Moreover, ψ_2' is given by

$$\psi_2'(t) = \begin{cases} \exp\left\{\left(p\,(\sigma d)^p \log\left(\frac{t-c_2}{c_1\left(\frac{1}{(\sigma d)^p}+1\right)}\right)\right)^{\frac{1}{p}}\right\} - e & t \geq 1 \\ \log(t) & 0 < t \leq 1. \end{cases}$$

By Proposition 2.9, we only have to show that the limit (2.13) holds. We have that,

$$\begin{aligned} \psi_2'\left(\frac{1}{f(x)}\right) &= \psi_2'\left(C\sigma|x|\exp\left\{\frac{1}{p\sigma^p}(\log|x|-\mu)^p\right\}\right) \\ &= \exp\left\{\left(p(\sigma d)^p \cdot \log\left(|x|\exp\left\{\frac{1}{p\sigma^p}(\log|x|-\mu)^p\right\}\cdot O(1)\right)\right)^{\frac{1}{p}}\right\} \\ &= \exp\left\{\left(p(\sigma d)^p \cdot \left(\frac{1}{p\sigma^p}(\log|x|-\mu)^p + \log|x| + O(1)\right)\right)^{\frac{1}{p}}\right\} \\ &= \exp\left((d^p\,(\log|x|-\mu)^p \cdot O(1))^{\frac{1}{p}}\right) \\ &= O(|x|^d). \end{aligned}$$

Hence, ψ_2' satisfies the limit (2.13)

Similarly, we can also construct a function that bounds $(\phi_2^*)'$ from above. For any $\gamma > d$, we define

$$(\hat{\psi}_2^*)'(s) = \begin{cases} \max\left\{\exp\left\{\frac{1}{p(\gamma\sigma)^p}(\log^p(s+e)-1)\right\}, (\phi_2^*)'(s)\right\} & s > 0 \\ \exp(s) & s \leq 0. \end{cases}$$

Then, we have that $(\hat{\psi}_2^*)'(s) \geq (\phi_2^*)'(s)$ for all $s \in \mathbb{R}$. Moreover, $\hat{\psi}_2^*(s) \triangleq \int_0^s (\hat{\psi}_2^*)'(y)\mathrm{d}y$ is an increasing convex function with $\hat{\psi}_2^*(0) = 0, (\hat{\psi}_2^*)'(0) = 1$. Hence, $\hat{\psi}_2 \triangleq \hat{\psi}_2^{**}$ is also a divergence function, and $(\hat{\psi}_2^*)' = (\hat{\psi}_2')^{-1}$ by Proposition 2.8.

To show that the divergence $\hat{\psi}_2$ excludes all densities g that satisfy $\liminf_{x\to-\infty}|x|^{t+1}g(x) > 0$, for any $t \in (1,d)$. We let $\gamma \in (t,d)$ and use Proposition 3 of Kruse et al. [2019], for which we have to verify the limit conditions

$$\limsup_{u\to\infty} \frac{u\hat{\psi}_2''(u)}{\hat{\psi}_2'(u)} < \infty,$$

and

$$\liminf_{x\to-\infty} \frac{\hat{\psi}_2'\left(\frac{1}{f_0(x)}\cdot\frac{1}{|x|^{t+1}}\cdot c\right)}{|x|^t} > 0, \tag{2.39}$$

for all constants $c > 0$. We note that since $\frac{1}{p(\gamma\sigma)^p} > \frac{1}{p(d\sigma)^p}$ due to $\gamma < d$, we have that

$(\hat{\psi}_2')^{-1}(s) = (\hat{\psi}_2^*)'(s) = \exp\left\{\frac{1}{p(\gamma\sigma)^p}(\log^p(s+e)-1)\right\}$, if s is sufficiently large. This means that for u sufficiently large, we have $\hat{\psi}_2'(u) = \exp\left\{(p(\gamma\sigma)^p \log(u)+1)^{\frac{1}{p}}\right\} - e$. Hence, we have that

$$\begin{aligned}\limsup_{u\to\infty} \frac{u\hat{\psi}_2''(u)}{\hat{\psi}_2'(u)} &= \limsup_{u\to\infty} \frac{(\hat{\psi}_2'(u)+e)(\gamma\sigma)^p(p(\gamma\sigma)^p\log(u)+1)^{\frac{1}{p}-1}}{\hat{\psi}_2'(u)} \\ &= \lim_{u\to\infty}(\gamma\sigma)^p(p(\gamma\sigma)^p\log(u)+1)^{\frac{1}{p}-1} = 0 < \infty,\end{aligned}$$

since $1/p - 1 < 0$ due to $p \geq 2$. We also have that, for any constant $c > 0$:

$$\begin{aligned}\hat{\psi}_2'\left(\frac{1}{f_0(x)}\cdot\frac{1}{|x|^{t+1}}\cdot c\right) &= \exp\left((\gamma^p(\log|x|-\mu)^p + O(1)\cdot\log|x| + O(1))^{\frac{1}{p}}\right) \\ &= |x|^\gamma \cdot \exp(O(1)).\end{aligned}$$

Therefore, we have that $\lim_{x\to-\infty}|x|^{-t}\hat{\psi}_2'\left(\frac{1}{f_0(x)}\cdot\frac{1}{|x|^{t+1}}\cdot c\right) = \lim_{x\to-\infty}|x|^{\gamma-t}\exp(O(1)) > 0$, since $\gamma > t$. Therefore, we have that $I_{\hat{\psi}_2}(g, f_0) = \infty$, which implies that $I_{\phi_2}(g, f_0) = \infty$, by Proposition 2.9. □

Proof of Lemma 2.13. Following the construction in (2.12), we choose $\psi(s) = (s+1)\exp((s+1)^p)$, where $p = \frac{k}{d}$. The second derivative of ψ is given by:

$$\psi''(s) = p(s+1)^{p-1}\exp((s+1)^p)(p + p(s+1)^p + 1),$$

which is positive for all $s \geq 0$. Hence, ψ is an increasing, strictly convex, and twice continuously differentiable function on $s \geq 0$. By Proposition 2.8, ϕ_2^* defined in (2.15) induces a divergence function. We examine condition (2.10). Choose an arbitrary $\theta_2 \in \mathbb{R}, \theta_1 \geq 0$. For $x \to -\infty$, we have that

$$\begin{aligned}&\phi_2^*(\phi_1^*(\theta_2 - x) + \theta_1)f_0(x) \\ &= \phi_2^*\left(\frac{1}{\alpha}(\theta_2 - x) + \theta_1\right)\frac{k}{\lambda}\left(\frac{|x|}{\lambda}\right)^{k-1}\exp\left(-\left(\frac{|x|}{\lambda}\right)^k\right) \\ &= O(|x|^k)\cdot\exp\left(\frac{1}{\alpha}|x|^{\frac{k}{d}}\cdot O(1) - \left(\frac{|x|}{\lambda}\right)^k\right) + O\left(|x|^k\exp\left(-\left(\frac{|x|}{\lambda}\right)^k\right)\right).\end{aligned}$$

Since $\frac{k}{d} < k$, we have that the term $\exp\left(\frac{1}{\alpha}|x|^{\frac{k}{d}}\cdot O(1) - \left(\frac{|x|}{\lambda}\right)^k\right)$ is integrable. Hence, condition (2.10) is satisfied.

To examine the content of the divergence ball, we first examine the derivative of ϕ_2^*,

which is equal to

$$(\phi_2^*)'(s) = \begin{cases} c_1 \exp((s+1)^{\frac{k}{d}}) \left(1 + \frac{k}{d}(s+1)^{\frac{k}{d}}\right) + c_2 & s > 0 \\ \exp(s) & s \leq 0. \end{cases}$$

We then define the auxiliary function

$$(\psi_2')^{-1}(s) = \begin{cases} c_1 \exp((s+1)^{\frac{k}{d}}) \left(1 + \frac{k}{d}\right) + c_2 & s > 0 \\ \exp(s) & s \leq 0, \end{cases}$$

which has inverse

$$\psi_2'(t) = \begin{cases} \log^{\frac{d}{k}} \left(\frac{t-c_2}{c_1(1+k/d)}\right) - 1 & t > 1 \\ \log(t) & 0 < t \leq 1. \end{cases}$$

We examine the asymptotic behavior:

$$\begin{aligned} \psi_2' \left(\frac{1}{f_0(x)}\right) &= \log^{\frac{d}{k}} \left(|x|^{1-k} \exp\left(\left(\frac{|x|}{\lambda}\right)^k\right) \cdot O(1)\right) - 1 \\ &= \left(\left(\frac{|x|}{\lambda}\right)^k + (1-k)\log|x| + O(1)\right)^{\frac{d}{k}} - 1 \\ &= O(|x|^d). \end{aligned}$$

Hence, we have $\limsup_{x \to -\infty} |x|^{-d} \psi_2' \left(\frac{1}{f_0(x)}\right) < \infty$. Proposition 2.9 implies that $I_{\phi_2}(g, f_0) < \infty$ for any continuous density g with finite d-th moment.

On the other hand, for any $\gamma > 1$, we can define the function

$$(\hat{\psi}_2^*)'(s) = \begin{cases} \max\left\{\exp\left(\gamma(s+1)^{\frac{k}{d}} - \gamma\right), (\phi_2^*)'(s)\right\} & s > 0 \\ \exp(s) & s \leq 0. \end{cases}$$

Then, we have that $(\hat{\psi}_2^*)'(s) \geq (\phi_2^*)'(s)$ for all $s \in \mathbb{R}$, and that $\hat{\psi}_2^*(s) \triangleq \int_0^s (\hat{\psi}_2^*)'(y)\mathrm{d}y$ is an increasing convex function with $\hat{\psi}_2^*(0) = 0, (\hat{\psi}_2^*)'(0) = 1$. Hence, $\hat{\psi}_2 \triangleq \hat{\psi}_2^{**}$ is a divergence function by Proposition 2.8. Since the term $\exp\left(\gamma(s+1)^{\frac{k}{d}}\right)$ dominates the term $\exp\left((s+1)^{\frac{k}{d}}\right)$ in $(\phi_2^*)'$ as $s \to \infty$, we have that for s sufficiently large, that $(\hat{\psi}_2^*)'(s) = \exp\left(\gamma(s+1)^{\frac{k}{d}} - \gamma\right)$. Hence, for u sufficiently large, we have that

$$\hat{\psi}_2'(u) = \left(\frac{\log(u) + \gamma}{\gamma}\right)^{\frac{d}{k}} - 1.$$

Therefore, to apply Proposition 3 of Kruse et al. [2019], we examine the limit conditions

$$\limsup_{u\to\infty} \frac{u(\hat{\psi}_2)''(u)}{(\hat{\psi}_2)'(u)} = \limsup_{u\to\infty} \frac{d}{\gamma k} \left(\frac{\log(u)+\gamma}{\gamma}\right)^{-1} < \infty,$$

and for any $t \in (1, d)$, any constants $c > 0$, we have

$$\begin{aligned}\liminf_{x\to-\infty} \hat{\psi}_2' \left(\frac{1}{f_0(x)} \cdot \frac{1}{|x|^{t+1}} \cdot c\right) |x|^{-t} &= \liminf_{x\to-\infty} \left(\left(\left(\frac{|x|}{\lambda}\right)^k + O(\log|x|)\right) \cdot O(1)\right)^{\frac{d}{k}} |x|^{-t} \\ &= O(|x|^{d-t}) > 0.\end{aligned}$$

Hence, by Proposition (2.9), we have that $I_{\phi_2}(g, f_0) = \infty$, for any g such that $\liminf_{x\to-\infty} |x|^{t+1} g(x) > 0$. □

Proof of Lemma 2.15. It follows from the proof of Lemma 2.11 that ϕ_2^* as in (2.16) is convex and twice differentiable and indeed defines a ϕ-divergence. It remains to show that ϕ_2^* satisfies condition (2.10) and (2.9). To verify this, we choose $\theta_1 = \theta_2 = 0$. Then, we examine

$$\begin{aligned}&\phi_2^* \left(\gamma(\exp\{\gamma|x|\} - 1)\right) \frac{k}{\lambda} \left(\frac{|x|}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{|x|}{\lambda}\right)^k\right\} \\ &= c_1(\gamma(\exp\{\gamma|x|\} - 1) + e) \exp\left\{\frac{1}{(2\gamma\lambda)^k} \log^k\left(\gamma(\exp\{\gamma|x|\} - 1)\right)\right\} \\ &\quad \cdot \frac{k}{\lambda} \left(\frac{|x|}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{|x|}{\lambda}\right)^k\right\} + O\left(\exp\left(\gamma|x| - \left(\frac{|x|}{\lambda}\right)^k\right)\right).\end{aligned}$$

Since $k > 1$, $O\left(\exp\left(\gamma|x| - \left(\frac{|x|}{\lambda}\right)^k\right)\right)$ is integrable. Therefore, we only examine the dominant terms that determine the asymptotics:

$$\begin{aligned}&\exp\left\{\gamma|x| + \frac{1}{(2\gamma\lambda)^k} \log^k\left(\gamma(\exp\{\gamma|x|\} - 1)\right) - \left(\frac{|x|}{\lambda}\right)^k\right\} \\ &= \exp\left\{\gamma|x| + \left(\frac{|x|}{2\lambda}\right)^k (1 + o(1)) - \left(\frac{|x|}{\lambda}\right)^k\right\} \\ &= O\left(\exp\left\{\gamma|x| + \left(\frac{1}{(2\lambda)^k} - \frac{1}{\lambda^k}\right)|x|^k\right\}\right).\end{aligned}$$

Hence, this term is integrable and therefore (2.10) is satisfied.

We have that the derivative of ϕ_2^* is given by

$$(\phi_2^*)'(s) = \begin{cases} c_1 \exp\left\{\frac{1}{(2\gamma\lambda)^k}\log^k(s+e)\right\}\left(\frac{k}{(2\gamma\lambda)^k}\log^{k-1}(s+e)+1\right)+c_2 & s>0 \\ \exp(s) & s\leq 0. \end{cases}$$

We define a function

$$(\psi_2')^{-1}(s) = \begin{cases} c_1 \exp\left\{\frac{1}{(2\gamma\lambda)^k}\log^k(s+e)\right\}\left(\frac{k}{(2\gamma\lambda)^k}+1\right)+c_2 & s>0 \\ \exp(s) & s\leq 0. \end{cases}$$

We note that $(\phi_2^*)'(s) \geq (\psi_2')^{-1}(s)$, for all $s \in \mathbb{R}$. Moreover, ψ_2' is given by

$$\psi_2'(t) = \begin{cases} \exp\left\{\left((2\gamma\lambda)^k \log\left(\frac{t-c_2}{c_1\left(\frac{k}{(2\gamma\lambda)^k}+1\right)}\right)\right)^{\frac{1}{k}}\right\} - e & t \geq 1 \\ \log(t) & 0 < t \leq 1. \end{cases}$$

We consider a Weibull distribution g with slightly heavier tail:

$$g(x) = \frac{l}{\lambda}\left(\frac{|x|}{\lambda}\right)^{l-1}\exp\left(-\left(\frac{|x|}{\lambda}\right)^l\right),$$

where $1 < l < k$. Then, we show that $I_{\phi_2}(g, f_0) < \infty$. Following the proof of Proposition 2.9, it is sufficient to show that the following function is integrable on $\{x : |x| > x_0\}$ for some $x_0 \in \mathbb{R}$:

$$\begin{aligned} &g(x)\psi_2'\left(\frac{1}{f(x)}\right) \\ &= O\left(|x|^{l-1}\exp\left\{-\left(\frac{|x|}{\lambda}\right)^l + (2\gamma\lambda)\left(\left(\frac{|x|}{\lambda}\right)^k + (1-k)\log\left(\frac{|x|}{\lambda}\right)+O(1)\right)^{\frac{1}{k}}\right\}\right) \\ &= O\left(|x|^{l-1}\exp\left\{-\left(\frac{|x|}{\lambda}\right)^l + 2\gamma|x|\right\}\right). \end{aligned}$$

Since $l > 1$, we have that $g(x)\psi_2'\left(\frac{1}{f_0(x)}\right)$ is indeed integrable. Therefore, $I_{\phi_2}(g, f_0) < \infty$. □

Proof of Theorem 2.16. We express the following optimization problem in terms of

density functions:

$$\begin{aligned}
&\sup_{\lambda>0} \sup_{\{\mathbb{Q}: I_{\phi_2}(\mathbb{Q},\mathbb{P}_0)\le r\}} \sup_{\{\bar{\mathbb{Q}}:\bar{\mathbb{Q}}\ll\mathbb{Q}\}} \mathbb{E}_{\bar{\mathbb{Q}}}[-X] - I_{\phi_1}(\bar{\mathbb{Q}}, \lambda\mathbb{Q}).\\
&= - \inf_{\substack{\lambda>0, g, \bar{g}\ge 0,\\ \int_\Omega g(\omega)\mathrm{d}\mathbb{P}_0(\omega)=1, \int_\Omega \bar{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)=1\\ \int_\Omega \phi_2(g(\omega))\mathrm{d}\mathbb{P}_0(\omega)\le r}} \int_\Omega X(\omega)\bar{g}(\omega) + \lambda g(\omega)\phi_1\left(\frac{\bar{g}(\omega)}{\lambda g(\omega)}\right)\mathrm{d}\mathbb{P}_0(\omega).
\end{aligned}$$

Due to the product $\lambda g(\omega)$, this is a non-convex problem. However, using the substitution $\tilde{g}(\omega) = \lambda g(\omega)$, this is equivalent to the following convex problem

$$\mu_0 \triangleq \inf_{\substack{\lambda>0, \tilde{g}, \bar{g}\ge 0,\\ \int_\Omega \tilde{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)=\lambda, \int_\Omega \bar{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)=1\\ \int_\Omega \lambda\phi_2\left(\frac{\tilde{g}(\omega)}{\lambda}\right)\mathrm{d}\mathbb{P}_0(\omega)\le\lambda r}} \int_\Omega X(\omega)\bar{g}(\omega) + \tilde{g}(\omega)\phi_1\left(\frac{\bar{g}(\omega)}{\tilde{g}(\omega)}\right)\mathrm{d}\mathbb{P}_0(\omega).$$

Assume without loss of generality that μ_0 is finite. To apply Theorem 2.B.2, we choose $\mathcal{X} = (0,\infty)\times\mathcal{L}^1(\Omega)\times\mathcal{L}^1(\Omega)$, $\mathcal{C} = (0,\infty)\times\mathcal{L}^1_+(\Omega)\times\mathcal{L}^1_+(\Omega)$, $G(\lambda,\tilde{g},\bar{g}) = \int_\Omega \lambda\phi_2\left(\frac{\tilde{g}(\omega)}{\lambda}\right)\mathrm{d}\mathbb{P}_0(\omega) - \lambda r$, $H(\lambda,\tilde{g},\bar{g}) = \left(\int_\Omega \tilde{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega), \int_\Omega \bar{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)\right) - (\lambda, 1)$. Then, we have that for $f_1 \equiv 1$ the constant function, $(1, f_1, f_1)$ is a point such that $H(1, f_1, f_1) = (0,0)$ and $G(1, f_1, f_1) < 0$. Furthermore, $(0,0)$ is an interior point of the image set of H. Hence, strong duality implies that

$$\mu_0 = \sup_{\eta\ge 0} -\eta r\lambda + \inf_{\substack{\lambda>0,\tilde{g},\bar{g}\ge 0,\\ \int_\Omega \tilde{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)=\lambda\\ \int_\Omega \bar{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)=1}} \int_\Omega X(\omega)\bar{g}(\omega) + \tilde{g}(\omega)\phi_1\left(\frac{\bar{g}(\omega)}{\tilde{g}(\omega)}\right) + \eta\lambda\phi_2\left(\frac{\tilde{g}(\omega)}{\lambda}\right)\mathrm{d}\mathbb{P}_0(\omega).$$

We first study the case when the supremum is attained at $\eta = 0$, in which case we have that

$$\begin{aligned}
\mu_0 &= \inf_{\substack{\lambda>0,\tilde{g},\bar{g}\ge 0,\\ \int_\Omega \tilde{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)=\lambda\\ \int_\Omega \bar{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)=1}} \int_\Omega X(\omega)\bar{g}(\omega) + \tilde{g}(\omega)\phi_1\left(\frac{\bar{g}(\omega)}{\tilde{g}(\omega)}\right)\mathrm{d}\mathbb{P}_0(\omega)\\
&\ge \inf_{\substack{\lambda>0,\tilde{g},\bar{g}\ge 0,\\ \int_\Omega \tilde{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)=\lambda\\ \int_\Omega \bar{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)=1}} \int_\Omega X(\omega)\bar{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega) = \operatorname{ess\,inf}(X).
\end{aligned}$$

On the other hand, taking $\lambda = 1, \tilde{g} = \bar{g}$ shows that $\operatorname{ess\,inf}(X)$ is also an upper bound of μ_0. Hence, $\mu_0 = \operatorname{ess\,inf}(X)$. The dual formula (2.19) states that when $\eta = 0$, $\operatorname{ess\,inf}(X)$ is equal to

$$\sup_{\theta_1,\theta_2\in\mathbb{R}} \left\{ -\theta_2 \;\middle|\; \mathbb{E}_{\mathbb{P}_0}\left[0\phi_2^*\left(\frac{\phi_1^*(-\theta_2 - X) + \theta_1}{0}\right)\right] \le \theta_1 \right\},$$

which implies that $0 \le \theta_1 \le -\phi_1^*(-\theta_2 - \operatorname{ess\,inf}(X))$. This inequality is satisfied if we take

$-\theta_2 = \operatorname{ess\,inf}(X)$ and $\theta_1 = 0$, but violated if $-\theta_2 > \operatorname{ess\,inf}(X)$. Hence, the duality formula (2.19) holds for $\eta = 0$, and we only have to examine the case where the supremum is attained at $\eta > 0$, where strong duality implies

$$\mu_0 = \sup_{\substack{\eta>0\\ \theta_1,\theta_2\in\mathbb{R}}} \inf_{\substack{\lambda>0\\ \tilde{g},\bar{g}\geq 0}} (-\eta r + \theta_1)\lambda + \theta_2 + \int_\Omega \beta_{\eta,\lambda}(\omega, \tilde{g}(\omega), \bar{g}(\omega)) - \theta_1\tilde{g}(\omega) - \theta_2\bar{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega),$$

where

$$\beta_{\eta,\lambda}(\omega, s, t) = X(\omega)t + s\phi_1\left(\frac{t}{s}\right) + \lambda\eta\phi_2\left(\frac{s}{\lambda}\right).$$

Note that when $\eta > 0$, $\beta_{\eta,\lambda}(\omega, s, t)$ is only finite when $s, t \geq 0$. Analogous to the proof of and Theorem 2.B.3, we have

$$\begin{aligned}&\inf_{\tilde{g},\bar{g}\geq 0} \int_\Omega \beta_{\eta,\lambda}(\omega, \tilde{g}(\omega), \bar{g}(\omega)) - \theta_1\tilde{g}(\omega) - \theta_2\bar{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega)\\ &= -\int_\Omega \beta^*_{\eta,\lambda}(\omega, \theta_1, \theta_2)\mathrm{d}\mathbb{P}_0(\omega),\end{aligned}$$

where $\int_\Omega \beta^*_{\eta,\lambda}(\omega, \theta_1, \theta_2)\mathrm{d}\mathbb{P}_0(\omega) = +\infty$ if $\eta = 0$ and for $\eta > 0$, we have

$$\int_\Omega \beta^*_{\eta,\lambda}(\omega, \theta_1, \theta_2)\mathrm{d}\mathbb{P}_0(\omega) = \int_\Omega \lambda\eta\phi_2^*\left(\frac{\theta_1 + \phi_1^*(\theta_2 - X(\omega))}{\eta}\right)\mathrm{d}\mathbb{P}_0(\omega).$$

Hence, we have

$$\begin{aligned}\mu_0 &= \sup_{\substack{\eta>0\\ \theta_1,\theta_2\in\mathbb{R}}} \theta_2 + \inf_{\lambda>0}\left(\theta_1 - \eta r - \int_\Omega \eta\phi_2^*\left(\frac{\theta_1 + \phi_1^*(\theta_2 - X(\omega))}{\eta}\right)\mathrm{d}\mathbb{P}_0(\omega)\right)\lambda\\ &= \sup_{\substack{\eta>0\\ \theta_1,\theta_2\in\mathbb{R}}} \left\{\theta_2 \;\middle|\; \theta_1 - \eta r - \int_\Omega \eta\phi_2^*\left(\frac{\theta_1 + \phi_1^*(\theta_2 - X(\omega))}{\eta}\right)\mathrm{d}\mathbb{P}_0(\omega) \geq 0\right\}.\end{aligned}$$

□

Proof of Corollary 2.17. Clearly, (2.18) can only hold if there actually exists $\theta_1, \theta_2 \in \mathbb{R}$, $\eta \geq 0$, such that (2.20) is finite. Hence, suppose there exists some $\tilde{\theta}_1, \tilde{\theta}_2 \in \mathbb{R}$, $\tilde{\eta} \geq 0$ such that (2.20) is finite. If there also exists a $\hat{\theta}_2 \in \mathbb{R}$, such that

$$\int_{\mathbb{R}} \tilde{\eta}\phi_2^*\left(\frac{\phi_1^*(-\hat{\theta}_2 - x) + \tilde{\theta}_1}{\tilde{\eta}}\right) f_0(x)\mathrm{d}x \leq \tilde{\theta}_1 - \tilde{\eta}r, \tag{2.40}$$

then, we must have

$$\inf_{\theta_2\in\mathbb{R}} \int_{\mathbb{R}} \tilde{\eta}\phi_2^* \left(\frac{\phi_1^*(-\theta_2 - x) + \tilde{\theta}_1}{\tilde{\eta}} \right) f_0(x)\mathrm{d}x \leq \tilde{\theta}_1 - \tilde{\eta}r.$$

since both ϕ_2^* and ϕ_1^* are non-decreasing, monotone convergence theorem implies that

$$\begin{aligned}
&\inf_{\theta_2\in\mathbb{R}} \int_{\mathbb{R}} \tilde{\eta}\phi_2^* \left(\frac{\phi_1^*(-\theta_2 - x) + \tilde{\theta}_1}{\tilde{\eta}} \right) f_0(x)\mathrm{d}x \\
&= \int_{\mathbb{R}} \lim_{\theta_2\to-\infty} \tilde{\eta}\phi_2^* \left(\frac{\phi_1^*(-\theta_2 - x) + \tilde{\theta}_1}{\tilde{\eta}} \right) f_0(x)\mathrm{d}x \\
&= \int_{\mathbb{R}} \tilde{\eta}\phi_2^* \left(\frac{\phi_1^*(-\infty) + \tilde{\theta}_1}{\tilde{\eta}} \right) f_0(x)\mathrm{d}x = \tilde{\eta}\phi_2^* \left(\frac{\phi_1^*(-\infty) + \tilde{\theta}_1}{\tilde{\eta}} \right).
\end{aligned}$$

Therefore, if $\hat{\theta}_2$ satisfies (2.40), then one cannot have

$$\tilde{\eta}\phi_2^* \left(\frac{\phi_1^*(-\infty) + \tilde{\theta}_1}{\tilde{\eta}} \right) > \tilde{\theta}_1 - \tilde{\eta}r.$$

Hence, if (2.21) holds for all θ_1, θ_2, η, then (2.18) cannot hold.

On the other hand, if there exists some $\tilde{\theta}_1, \tilde{\theta}_2 \in \mathbb{R}$, $\tilde{\eta} \geq 0$ such that (2.20) is finite, and

$$\tilde{\eta}\phi_2^* \left(\frac{\phi_1^*(-\infty) + \tilde{\theta}_1}{\tilde{\eta}} \right) < \tilde{\theta}_1 - \tilde{\eta}r,$$

then it follows from the monotone convergence theorem that there exists a $\hat{\theta}_2 \in \mathbb{R}$ such that (2.40) is satisfied, and thus (2.18). □

Proof of Proposition 2.19. Let $x \in \mathbb{R}$, and denote $S(x,p) = \rho^l_{\mathrm{oce},\mathbb{P}}(X_p)$. We note that

$$\lim_{p\downarrow 0} \rho^l_{\mathrm{oce},\mathbb{P}}(X_p)/p = \frac{\partial S(x,p)}{\partial p}|_{p=0}.$$

We have that,

$$S(x,p) = \inf_{\theta_1,\theta_2\in\mathbb{R}} -\theta_1 - \theta_2 + p(\phi_2^*(\phi_1^*(\theta_2 - x) + \theta_1) + (1-p)\phi_2^*(\phi_1^*(\theta_2) + \theta_1),$$

and we denote

$$F(\theta_1,\theta_2,p) \triangleq -\theta_1 - \theta_2 + p(\phi_2^*(\phi_1^*(\theta_2 - x) + \theta_1) + (1-p)\phi_2^*(\phi_1^*(\theta_2) + \theta_1).$$

We note that for each p, the optimal value $\theta_1(x,p), \theta_2(x,p)$ satisfy the first-order conditions

$$\begin{aligned} p(\phi_2^*)'(\phi_1^*(\theta_2 - x) + \theta_1) + (1-p)(\phi_2^*)'(\phi_1^*(\theta_2) + \theta_1) &= 1 \\ p(\phi_2^*)'(\phi_1^*(\theta_2 - x) + \theta_1)(\phi_1^*)'(\theta_2 - x) + (1-p)(\phi_2^*)'(\phi_1^*(\theta_2) + \theta_1)(\phi_1^*)'(\theta_2) &= 1. \end{aligned}$$

Since $\phi_i^*(0) = 0, (\phi_i^*)'(0) = 1$ for $i = 1, 2$, we have that $(0,0)$ satisfies the above condition for $p = 0$. Hence, $\theta_1(x,0) = \theta_2(x,0) = 0$. Now, by the Envelope theorem (see e.g., Mas-Colell et al., 1995), we have that

$$\begin{aligned} \frac{\partial S(x,p)}{\partial p} &= \left.\frac{\partial F(\theta_1,\theta_2,p)}{\partial p}\right|_{(\theta_1(x,p),\theta_2(x,p),p)} \\ &= \phi_2^*(\phi_1^*(\theta_2(x,p) - x) + \theta_1(x,p)) - \phi_2^*(\phi_1^*(\theta_2(x,p)) + \theta_1(x,p)). \end{aligned}$$

Hence,

$$\begin{aligned} \left.\frac{\partial S(x,p)}{\partial p}\right|_{p=0} &= \phi_2^*(\phi_1^*(\theta_2(x,0) - x) + \theta_1(x,0)) - \phi_2^*(\phi_1^*(\theta_2(x,0)) + \theta_1(x,0)) \\ &= \phi_2^*(\phi_1^*(-x)). \end{aligned}$$

We now examine the robust shortfall risk measure. We first note that by weak duality, we have $\rho^l_{\mathrm{oce},\mathbb{P}}(X_p) \le \rho^l_{\mathrm{sf},\mathbb{P}}(X_p)$ for all $p > 0$. Therefore, we already have

$$\liminf_{p \downarrow 0} \frac{\rho^l_{\mathrm{sf},\mathbb{P}}(X_p)}{p} \ge \phi_2^*(\phi_1^*(-x)).$$

We will now show the reverse. By definition, we have that

$$\begin{aligned} \frac{1}{p}\rho^l_{\mathrm{sf},\mathbb{P}}(X_p) &= \inf_{\theta_1,\theta_2 \in \mathbb{R}} \left\{ \frac{1}{p}\theta_2 \;\middle|\; p\phi_2^*(\phi_1^*(-\theta_2 - x) + \theta_1) + (1-p)\phi_2^*(\phi_1^*(-\theta_2) + \theta_1) \le \theta_1 \right\} \\ &= \inf_{\theta_1,\theta_2 \in \mathbb{R}} \left\{ \theta_2 \mid p\phi_2^*(\phi_1^*(-p\theta_2 - x) + \theta_1) + (1-p)\phi_2^*(\phi_1^*(-p\theta_2) + \theta_1) \le \theta_1 \right\}. \end{aligned}$$

We bound the above infimum by considering only cases when $\theta_1 = 0$. Then, we show that as $p \downarrow 0$, the following inequality holds for any $\theta_2 > \phi_2^*(\phi_1^*(-x))$:

$$p\phi_2^*(\phi_1^*(-p\theta_2 - x)) + (1-p)\phi_2^*(\phi_1^*(-p\theta_2)) \le 0.$$

The above inequality holds if and only if

$$\phi_2^*(\phi_1^*(-p\theta_2 - x)) - \phi_2^*(\phi_1^*(-p\theta_2)) \le \theta_2 \cdot \frac{\phi_2^*(\phi_1^*(-p\theta_2))}{-p\theta_2}.$$

As $p \downarrow 0$, the left-hand side converges to $\phi_2^*(\phi_1^*(-x))$, and the right-hand side converges to $\theta_2 \cdot \lim_{t \to 0} \frac{\phi_2^*(\phi_1^*(t))}{t} = \theta_2 \cdot 1 = \theta_2$. Since $\theta_2 > \phi_2^*(\phi_1^*(-x))$, the above inequality is feasible

for p sufficiently small. Therefore, we also have $\limsup_{p\downarrow 0} \frac{\rho^l_{\mathrm{sf},\mathbb{P}}(X_p)}{p} \leq \phi_2^*(\phi_1^*(-x))$. □

Proof of Theorem 2.22. We have that (2.41) is equivalent (up to a sign convention) to the following optimization problem over density functions:

$$\begin{aligned} \text{Minimize } & \int_\Omega X(\omega)\bar{g}(\omega) + \tilde{g}(\omega)\phi_1\left(\frac{\bar{g}(\omega)}{\tilde{g}(\omega)}\right) + g(\omega)\phi_2\left(\frac{\tilde{g}(\omega)}{g(\omega)}\right) \mathrm{d}\mathbb{P}_0(\omega) \\ \text{subject to } & \int_\Omega g(\omega)\mathrm{d}\mathbb{P}_0(\omega) = \int_\Omega \tilde{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega) = \int_\Omega \bar{g}(\omega)\mathrm{d}\mathbb{P}_0(\omega) = 1 \\ & \int_\Omega \phi_3(g(\omega))\mathrm{d}\mathbb{P}_0(\omega) \leq r \\ & g(\omega), \tilde{g}(\omega), \bar{g}(\omega) \geq 0. \end{aligned} \tag{2.41}$$

We denote the optimal value of the primal problem (2.41) as J_0. Note that $J_0 < \infty$ since the constant unity function $g, \tilde{g}, \bar{g} \equiv 1$ is a feasible solution and we assumed $\int_\Omega X(\omega)\mathrm{d}\mathbb{P}_0(\omega) < \infty$. Strong duality then implies that

$$\begin{aligned} J_0 = \sup_{\substack{\theta_1,\theta_2,\theta_3\in\mathbb{R} \\ \lambda\geq 0}} -\lambda r + \inf_{\substack{g,\tilde{g},\bar{g}\geq 0 \\ \int g = \int \tilde{g} = \int \bar{g} = 1}} & \int_\Omega X(\omega)\bar{g}(\omega) + \tilde{g}(\omega)\phi_1\left(\frac{\bar{g}(\omega)}{\tilde{g}(\omega)}\right) + g(\omega)\phi_2\left(\frac{\tilde{g}(\omega)}{g(\omega)}\right) \mathrm{d} \\ & + \lambda\phi_3(g(\omega))\mathbb{P}_0(\omega). \end{aligned}$$

Suppose this supremum is attained at $\lambda = 0$, then we would have $J_0 = \operatorname{ess\,inf}(X)$. The dual formula in (2.32) would also yield

$$\begin{aligned} & \sup_{\theta_1,\theta_2,\theta_3\in\mathbb{R}} \theta_1 + \theta_2 + \theta_3 - \mathbb{E}_{\mathbb{P}_0}\left[0\phi_3^*\left(\frac{\phi_2^*(\phi_1^*(\theta_3 - X(\omega)) + \theta_2) + \theta_1}{0}\right)\right] \\ &= \sup_{\theta_2,\theta_3\in\mathbb{R}} \theta_2 + \theta_3 - \phi_2^*(\phi_1^*(\theta_2 + \theta_3 - \operatorname{ess\,inf}(X))) \\ &= \operatorname{ess\,inf}(X). \end{aligned}$$

Hence, the duality formula holds if the supremum is attained at $\lambda = 0$. We may now assume that the supremum is at $\lambda > 0$. Applying Theorem 2.B.2 and Theorem 14.60 of Rockafellar and Wets [2004], gives the duality

$$J_0 = \sup_{\substack{\theta_1,\theta_2,\theta_3\in\mathbb{R} \\ \lambda>0}} -\lambda r + \theta_1 + \theta_2 + \theta_3 - \int_\Omega \beta_\lambda^*(\omega, \theta_1, \theta_2, \theta_3)\mathrm{d}\mathbb{P}_0(\omega),$$

where

$$\beta_\lambda(\omega, t_1, t_2, t_3) = X(\omega)t_3 + t_2\phi_1\left(\frac{t_3}{t_2}\right) + t_1\phi_2\left(\frac{t_2}{t_1}\right) + \lambda\phi_3(t_1).$$

We note that $\beta_\lambda(\omega, t_1, t_2, t_3)$ is only finite if $t_1, t_2, t_3 \geq 0$, since $\lambda > 0$. Therefore, we have

$$\begin{aligned}
\beta_0^*(\omega, \theta_1, \theta_2, \theta_3) &= \sup_{t_1,t_2,t_3\geq 0} \theta_1 t_1 + \theta_2 t_2 + \theta_3 t_3 - X(\omega) t_3 - t_2 \phi_1\left(\frac{t_3}{t_2}\right) - t_1 \phi_2\left(\frac{t_2}{t_1}\right) - \lambda\phi_3(t_1) \\
&= \sup_{t_1\geq 0, t_2 \geq 0} \theta_1 t_1 + \theta_2 t_2 + t_2 \phi_1^*(\theta_3 - X(\omega)) - t_1 \phi_2\left(\frac{t_2}{t_1}\right) - \lambda\phi_3(t_1) \\
&= \sup_{t_1\geq 0} \theta_1 t_1 + t_1 \phi_2^*(\phi_1^*(\theta_3 - X(\omega)) + \theta_2) - \lambda\phi_3(t_1) \\
&= \lambda \phi_3^*\left(\frac{\phi_2^*(\phi_1^*(\theta_3 - X(\omega)) + \theta_2) + \theta_1}{\lambda}\right).
\end{aligned}$$

□

Proof of Proposition 2.20. Denote

$$K(\theta_1, \theta_2) \triangleq -\theta_1 - \theta_2 + \mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*(\phi_1^*(\theta_2 - X) + \theta_1)\right]. \tag{2.42}$$

Since $\phi_2(1) = 0$, we have that $\phi_2^*(s) \geq s, \forall s \in \mathbb{R}$. Therefore, for any $\theta_1, \theta_2 \in \mathbb{R}$:

$$K(\theta_1, \theta_2) \geq -\theta_2 + \mathbb{E}_{\mathbb{P}_0}\left[\phi_1^*(\theta_2 - X)\right].$$

Since $\phi_2 \in \Phi_0$, we have that $\mathrm{dom}(\phi_2)$ has a non-empty interior around 1. Therefore, let $x_0 < 1 < y_0$ be such that $\phi_2(x_0), \phi_2(y_0) < \infty$. Then, we have that

$$\begin{aligned}
\phi_1^*(\theta_2 - X) \geq \max\{&x_0(\theta_2 - X) - \phi_1(x_0), \\
&y_0(\theta_2 - X) - \phi_1(y_0)\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
K(\theta_1, \theta_2) \geq \max\{&(x_0 - 1)\theta_2 - \phi_1(x_0) - x_0\mathbb{E}_{\mathbb{P}_0}[X], \\
&(y_0 - 1)\theta_2 - \phi_1(y_0) - x_0\mathbb{E}_{\mathbb{P}_0}[X]\}.
\end{aligned}$$

By assumption, there exists a (θ_1^*, θ_2^*) such that $K(\theta_1^*, \theta_2^*) < \infty$. We now claim that θ_2 can be restricted to the set

$$\left[\frac{1}{x_0 - 1}(\phi_1(x_0) + x_0\mathbb{E}_{\mathbb{P}_0}[X] + K(\theta_1^*, \theta_2^*)),\ \frac{1}{y_0 - 1}(\phi_1(y_0) + y_0\mathbb{E}_{\mathbb{P}_0}[X] + K(\theta_1^*, \theta_2^*))\right]. \tag{2.43}$$

Indeed, if $\theta_2 < \frac{1}{x_0 - 1}(\phi_1(x_0) + x_0\mathbb{E}_{\mathbb{P}_0}[X] + K(\theta_1^*, \theta_2^*))$, then since $(x_0 - 1) < 0$, we have that

$$\begin{aligned}
K(\theta_1, \theta_2) &\geq (x_0 - 1)\theta_2 - \phi_1(x_0) - x_0\mathbb{E}_{\mathbb{P}_0}[X] \\
&> K(\theta_1^*, \theta_2^*).
\end{aligned}$$

Similarly, if $\theta_2 > \frac{1}{y_0-1}(\phi_1(y_0) + y_0\mathbb{E}_{\mathbb{P}_0}[X] + K(\theta_1^*, \theta_2^*))$, then $K(\theta_1, \theta_2) > K(\theta_1^*, \theta_2^*)$.

Now, let $L_2 > 0$ be a number such that $[-L_2, L_2]$ contains (2.43). We may thus restrict θ_2 on $[-L_2, L_2]$. Then, by the non-decreasing property of ϕ_1^*, ϕ_2^*, we have that

$$\begin{aligned} K(\theta_1, \theta_2) &\geq -\theta_1 - L_2 + \mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*(\phi_1^*(-L_2 - X) + \theta_1)\right] \\ &\geq -\theta_1 - L_2 + \mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*(-L_2 - X + \theta_1)\right]. \end{aligned} \tag{2.44}$$

Again, let $\tilde{x}_0 < 1 < \tilde{y}_0$ be such that $\phi_2^*(\tilde{x}_0), \phi_2^*(\tilde{y}_0) < \infty$. Then, with the same argument, we may restrict θ_1 on the interval

$$\left[\frac{1}{\tilde{x}_0 - 1}\left((\tilde{x}_0 + 1)L_2 + \tilde{x}_0\mathbb{E}_{\mathbb{P}_0}[X] + \phi_2(\tilde{x}_0) + K(\theta_1^*, \theta_2^*)\right)\right.$$
$$\left., \frac{1}{\tilde{y}_0 - 1}\left((\tilde{y}_0 + 1)L_2 + \tilde{y}_0\mathbb{E}_{\mathbb{P}_0}[X] + \phi_2(\tilde{y}_0) + K(\theta_1^*, \theta_2^*)\right)\right].$$

Hence, θ_1, θ_2 can be both restricted to compact sets, without changing the optimum. We now examine

$$V(\theta_1, \theta_2, \lambda) \triangleq -\theta_1 - \theta_2 + \lambda r + \mathbb{E}_{\mathbb{P}_0}\left[\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2 - X) + \theta_1}{\lambda}\right)\right].$$

We note that since $\lambda\phi_2^*\left(\frac{s}{\lambda}\right) = \sup_{t\geq 0}\{st - \lambda\phi_2(s)\} = (\lambda\phi_2)^*(s)$ and $\phi_2 \geq 0$, we have that $\lambda\phi_2^*\left(\frac{s}{\lambda}\right)$ is non-increasing in λ, and that $(\lambda\phi_2)^*(s) \geq s$, $\forall s \in \mathbb{R}$. Therefore, for any $\theta_1, \theta_2 \in \mathbb{R}$ and $\lambda > 0$, we have

$$\begin{aligned} V(\theta_1, \theta_2, \lambda) &= -\theta_1 - \theta_2 + \lambda r + \mathbb{E}_{\mathbb{P}_0}\left[(\lambda\phi_2)^*(\phi_1^*(\theta_2 - X) + \theta_1)\right] \\ &\geq -\theta_1 - \theta_2 + \lambda r + \mathbb{E}_{\mathbb{P}_0}\left[\phi_1^*(\theta_2 - X) + \theta_1\right] \\ &\geq \lambda r - \mathbb{E}_{\mathbb{P}_0}[X]. \end{aligned}$$

Since $V(\theta_1^*, \theta_2^*, 1) = K(\theta_1^*, \theta_2^*) < \infty$, we have that λ may be restricted to the set

$$\left[0, \frac{V(\theta_1^*, \theta_2^*, 1) + \mathbb{E}_{\mathbb{P}_0}[X]}{r}\right].$$

Let L_λ denotes this upper bound of λ. Then, we have that for all $\lambda \in [0, L_\lambda]$,

$$\begin{aligned} V(\theta_1, \theta_2, \lambda) &\geq -\theta_1 - \theta_2 + \mathbb{E}_{\mathbb{P}_0}\left[L_\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2 - X) + \theta_1}{L_\lambda}\right)\right] \\ &\geq -\theta_2 + \mathbb{E}_{\mathbb{P}_0}\left[\phi_1^*(\theta_2 - X)\right]. \end{aligned}$$

Hence, θ_2 can be bounded by the same interval as in (2.43), thus we may assume that

$\theta_2 \in [-L_2, L_2]$ for some $L_2 > 0$. Then, we also have

$$\begin{aligned} V(\theta_1, \theta_2, \lambda) &\geq -\theta_1 - L_2 + \mathbb{E}_{\mathbb{P}_0}\left[L_\lambda \phi_2^*\left(\frac{\phi_1^*(-L_2 - X) + \theta_1}{L_\lambda}\right)\right] \\ &\geq -\theta_1 - L_2 + \mathbb{E}_{\mathbb{P}_0}\left[L_\lambda \phi_2^*\left(\frac{-L_2 - X + \theta_1}{L_\lambda}\right)\right] \\ &= -\theta_1 - L_2 + \mathbb{E}_{\mathbb{P}_0}\left[(L_\lambda \phi_2)^*\left(-L_2 - X + \theta_1\right)\right]. \end{aligned}$$

Since $L_\lambda \phi_2$ is also a divergence function belonging to Φ_0. It follows from the analysis of (2.44) that θ_1 can also be bounded in the interval

$$\begin{aligned} \Bigg[&\frac{1}{\tilde{x}_0 - 1}\left((\tilde{x}_0 + 1)L_2 + \tilde{x}_0 \mathbb{E}_{\mathbb{P}_0}[X] + L_\lambda \phi_2(\tilde{x}_0) + K(\theta_1^*, \theta_2^*)\right) \\ &, \frac{1}{\tilde{y}_0 - 1}\left((\tilde{y}_0 + 1)L_2 + \tilde{y}_0 \mathbb{E}_{\mathbb{P}_0}[X] + L_\lambda \phi_2(\tilde{y}_0) + K(\theta_1^*, \theta_2^*)\right)\Bigg], \end{aligned}$$

where $\tilde{x}_0 < 1 < \tilde{y}_0$ are such that $\tilde{x}_0, \tilde{y}_0 \in \mathrm{dom}(\phi_2)$. □

Proof of Theorem 2.21. We verify the six conditions of Theorem 5.4 in Shapiro et al. [2009]. We have that (2.7) is equal to

$$\inf_{\theta_1, \theta_2 \in \mathbb{R}} \mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*(\phi_1^*(\theta_2 - X) + \theta_1) - \theta_1 - \theta_2\right] \triangleq \inf_{\boldsymbol{\theta} \in \mathbb{R}^2} \mathbb{E}_{\mathbb{P}_0}[F(\boldsymbol{\theta}, X)],$$

and (2.8) is equal to

$$\inf_{\lambda \geq 0, \theta_1, \theta_2 \in \mathbb{R}} \mathbb{E}_{\mathbb{P}_0}\left[\lambda \phi_2^*\left(\frac{\phi_1^*(\theta_2 - X) + \theta_1}{\lambda}\right) - \theta_1 - \theta_2 + \lambda r\right] \triangleq \inf_{\lambda \geq 0, \boldsymbol{\theta} \in \mathbb{R}^2} \mathbb{E}_{\mathbb{P}_0}[G(\boldsymbol{\theta}, \lambda, X)],$$

where we replaced $\lambda > 0$ with $\lambda \geq 0$, which does not change the optimization problem.

By our assumptions $\mathrm{dom}(\phi_2^*) = \mathrm{dom}(\phi_1^*) = \mathbb{R}$, it follows from convexity that ϕ_1^*, ϕ_2^* are continuous on $\mathbb{R}$. Therefore, the functions $F(\boldsymbol{\theta}, X(\omega))$ and $G((\boldsymbol{\theta}, \lambda, X(\omega))$ are convex, lower-semicontinuous functions in $\boldsymbol{\theta}$ or $(\boldsymbol{\theta}, \lambda)$ with non-empty interior effective domain, for each $\omega \in \Omega$. Clearly, for each $\boldsymbol{\theta}$ or $(\boldsymbol{\theta}, \lambda)$, they are also measurable functions in ω. Hence, it follows from Proposition 14.39 of Rockafellar and Wets [2004] that they are normal convex integrand. Therefore, conditions (i) and (ii) of Theorem 5.4 are satisfied. The sets $\mathbb{R}^2$ and $\mathbb{R}^2 \times [0, \infty)$ are also closed and convex. Thus, (iii) of Theorem 5.4 is satisfied. We also have that for any $\boldsymbol{\theta} \in \mathbb{R}^2$ and $\omega \in \Omega$, that $F(\boldsymbol{\theta}, X(\omega)) \geq -X(\omega)$ and for any $\boldsymbol{\theta} \in \mathbb{R}^2, \lambda \geq 0$ and $\omega \in \Omega$, that $G(\boldsymbol{\theta}, \lambda, X(\omega)) \geq -X(\omega) + \lambda r \geq -X(\omega)$. Since X is assumed to be $\mathbb{P}_0$-integrable, it follows from Theorem 7.42 of Shapiro et al. [2009] that condition (iv) of Theorem 5.4 holds. Condition (v), which requires boundedness and non-emptiness of the set of solutions follows from Proposition 2.20 and the assumption. Finally, condition (vi), requires that the law of large number must hold pointwise for $F(\boldsymbol{\theta}, X)$ and $G(\boldsymbol{\theta}, \lambda, X)$. For cases where $\mathbb{E}_{\mathbb{P}_0}[F(\boldsymbol{\theta}, X)] < \infty$ or $\mathbb{E}_{\mathbb{P}_0}[G(\boldsymbol{\theta}, \lambda, X)] < \infty$, this

is trivial. For other cases, this holds by Theorem 2.4.5 of Durrett [2019], due to the fact that $F(\boldsymbol{\theta}, X) \geq -X$ and $G(\boldsymbol{\theta}, \lambda, X) \geq -X$, with $\mathbb{E}_{\mathbb{P}_0}[-X] < \infty$. □

2.B Minimizing Convex Integral

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Consider the convex integral minimization problem over integrable functions.

$$J_\beta(\mathbf{a}) \triangleq \inf_{\mathbf{g} \in \mathcal{L}^1_+(\mathbf{a})} \int_\Omega \beta(\omega, \mathbf{g}(\omega)) \mathrm{d}\mathbb{P}(\omega), \tag{2.45}$$

where $\mathbf{a} \in \mathbb{R}^d$ and the set of d-dimensional non-negative $\mathbb{P}$-integrable functions with value $\mathbf{a}$, denoted as:

$$\mathcal{L}^1_+(\mathbf{a}) = \left\{ \mathbf{g} = (g_1, \ldots, g_d) \;\middle|\; g_i(\omega) \geq 0, \; \int_\Omega g_i(\omega) \mathrm{d}\mathbb{P}(\omega) = a_i, \forall i = 1, \ldots, d \right\}.$$

Here, $\beta(\omega, \mathbf{y}) : \mathbb{R}^d \to \mathbb{R}$ is a convex function in $\mathbf{y}$ for $\mathbb{P}$-almost every ω and a normal convex integrand in the sense of Rockafellar and Wets [2004]. Furthermore, we assume that $\mathrm{dom}(\beta(\omega, \mathbf{y})) \subset [0, \infty)^d$ for all $\omega \in \Omega$. Therefore, (2.45) is also equivalent to

$$J_\beta(\mathbf{a}) = \inf_{\mathbf{g} \in \mathcal{L}^1(\mathbf{a})} \int_\Omega \beta(\omega, \mathbf{g}(\omega)) \mathrm{d}\mathbb{P}(\omega), \tag{2.46}$$

where $\mathcal{L}^1(\mathbf{a})$ is the set of all $\mathbb{P}$-integrable functions with integral equals to $\mathbf{a}$:

$$\mathcal{L}^1(\mathbf{a}) = \left\{ \mathbf{g} = (g_1, \ldots, g_d) \;\middle|\; \int_\Omega g_i(\omega) \mathrm{d}\mathbb{P}(\omega) = a_i, \forall i = 1, \ldots, d \right\}.$$

In the sequel, we let $\mathcal{L}^1 \triangleq \bigcup_{\mathbf{a} \in \mathbb{R}^d} \mathcal{L}^1(\mathbf{a})$ denote the set of all $\mathbb{P}$-integrable functions. Problem (2.46) can be reformulated using Lagrangian duality. Since we are optimizing over density functions, which belong to an infinite-dimensional vector space of integrable functions, we invoke the following Lagrangian duality theorem given in Luenberger [1969], which we state here for completeness.

Theorem 2.B.2 (Luenberger 1969, exercise 8.8.7)**.** *Let $\mathcal{X}$ be a real vector space, $\mathcal{C}$ a convex subset of $\mathcal{X}$. Let f be a real-valued convex functional on $\mathcal{C}$ and $G : \mathcal{C} \to \mathbb{R}^d$ a convex mapping, i.e., $G(\lambda x_1 + (1-\lambda)x_2) \leq_{\mathbb{R}^d} \lambda G(x_1) + (1-\lambda)G(x_2), \forall x_1, x_2 \in \mathcal{C}, \lambda \in (0,1)$ ($x \leq_{\mathbb{R}^d} y \Leftrightarrow x_i \leq y_i, \forall i$). Let $H : \mathcal{X} \to \mathbb{R}^m$ be an affine mapping, i.e., $H(x) = Ax + b$ for some linear mapping A. Consider the optimization problem,*

$$\mu_0 = \inf_{x \in \mathcal{X}} \{ f(x) \mid x \in \mathcal{C}, G(x) \leq \mathbf{0}, H(x) = \mathbf{0} \},$$

where $\mathbf{0}$ denotes the zero-vector. Assume the existence of a $x_1 \in \mathcal{C}$ such that $G(x_1) <_{\mathbb{R}^d} \mathbf{0}$

and $H(x_1) = \mathbf{0}$. *Suppose also that* $\mathbf{0}$ *is an interior point of the image set* $\{y \in \mathbb{R}^m \mid H(x) = y \text{ for some } x \in \mathcal{C}\}$. *If* μ_0 *is finite, then there exists a* $\boldsymbol{\mu}^* \in \mathbb{R}^d$ *with* $\mu \geq_{\mathbb{R}^d} \mathbf{0}$ *and* $\boldsymbol{\nu}^* \in \mathbb{R}^m$, *such that*

$$\begin{aligned}\mu_0 &= \inf_{x\in\mathcal{C}}\{f(x) + \boldsymbol{\mu}^{*T}G(x) + \boldsymbol{\nu}^{*T}H(x)\} \\ &= \sup_{\substack{\boldsymbol{\mu}\geq_{\mathbb{R}^d}\mathbf{0} \\ \boldsymbol{\nu}\in\mathbb{R}^m}} \inf_{x\in\mathcal{C}}\{f(x) + \boldsymbol{\mu}^{T}G(x) + \boldsymbol{\nu}^{T}H(x)\}.\end{aligned}$$

To apply Theorem 2.B.2, we set $\mathcal{X} = (\mathcal{L}^1)^d$ and $H(\mathbf{g}) = \int_\Omega \mathbf{g}(\omega)\mathrm{d}\mathbb{P}(\omega) - \mathbf{a}$. Note that the zero vector $\mathbf{0} \in \mathbb{R}^d$ is an interior point of the image set of H, since any vector $\boldsymbol{\epsilon} \in \mathbb{R}^d$ sufficiently close to $\mathbf{0}$ is the image of $H(\mathbf{a} + \boldsymbol{\epsilon})$ of the constant function $\mathbf{g}(\omega) = \mathbf{a} + \boldsymbol{\epsilon}$. Hence, Theorem 2.B.2 can be applied to $J_\beta(\mathbf{a})$ if $J_\beta(\mathbf{a}) < +\infty$.

Theorem 2.B.3. *Assume* $J_\beta(\mathbf{a}) < +\infty$. *Then, we have the duality*

$$J_\beta(\mathbf{a}) = \sup_{\boldsymbol{\theta}\in\mathbb{R}^d} \boldsymbol{\theta}^T\mathbf{a} - K_\beta(\boldsymbol{\theta}), \tag{2.47}$$

where $K_\beta(\boldsymbol{\theta}) = \int_\Omega \beta^*(\omega, \boldsymbol{\theta})\mathrm{d}\mathbb{P}(\omega)$ *and* $\beta^*(\omega, \boldsymbol{\theta})$ *is the convex conjugate of* $\beta(\omega, \mathbf{y})$. *Moreover, if* $J_\beta(\mathbf{a}) > -\infty$, *then there exists* $\boldsymbol{\theta}^* \in \mathbb{R}^d$ *such that the supremum in* (2.47) *is attained.*

Proof. Applying Theorem 2.B.2 to (2.46) yields that

$$J_\beta(\mathbf{a}) = \sup_{\boldsymbol{\theta}\in\mathbb{R}^d} \boldsymbol{\theta}^T\mathbf{a} + \inf_{\mathbf{g}\in\mathcal{L}^1} \int_\Omega \beta(\omega, \mathbf{g}(\omega)) - \boldsymbol{\theta}^T\mathbf{g}(\omega)\mathrm{d}\mathbb{P}(\omega).$$

We have,

$$\begin{aligned}&\inf_{\mathbf{g}\in\mathcal{L}^1} \int_\Omega \beta(\omega, \mathbf{g}(\omega)) - \boldsymbol{\theta}^T\mathbf{g}(\omega)\mathrm{d}\mathbb{P}(\omega) \\ &\overset{(*)}{=} \int_\Omega \inf_{\mathbf{y}\in\mathbb{R}^d} (\beta(\omega, \mathbf{y}) - \boldsymbol{\theta}^T\mathbf{y})\mathrm{d}\mathbb{P}(\omega) \\ &\overset{(**)}{=} \int_\Omega \inf_{\mathbf{y}\geq\mathbf{0}} (\beta(\omega, \mathbf{y}) - \boldsymbol{\theta}^T\mathbf{y})\mathrm{d}\mathbb{P}(\omega) \\ &= -\int_\Omega \beta^*(\omega, \boldsymbol{\theta})\mathrm{d}\mathbb{P}(\omega),\end{aligned}$$

where the interchange of infimum and integral in $(*)$ follows from [Rockafellar and Wets [2004], Theorem 14.60], for which we may apply the theorem since $J_\beta(\mathbf{a}) < \infty$ and $\mathcal{L}^1$ is decomposable (see Rockafellar and Wets [2004], Definition 14.59). In $(**)$ we used that $\beta(\omega, \mathbf{y}) = +\infty$ if $\mathbf{y}$ has negative component. Finally, the existence of $\boldsymbol{\theta}^*$ is guaranteed by Theorem 2.B.2. □

Hence, the *primal problem* $J_\beta(\mathbf{a})$ can be reformulated into a finite-dimensional *dual*

problem as in (2.47). We call the solutions of the corresponding problems the *primal* and *dual solution* respectively. The following Theorem establishes the condition in which a convex function of the form $\boldsymbol{\theta} \mapsto K_\beta(\boldsymbol{\theta}) = \int_\Omega \beta^*(\omega, \boldsymbol{\theta}) \mathrm{d}\mathbb{P}(\omega)$ is differentiable.

Theorem 2.B.4. *Let $\boldsymbol{\theta} \in \mathbb{R}^d$. If $\mathrm{int}(\mathrm{dom}(K_\beta)) \neq \emptyset$ and $\beta^*(\omega, \cdot)$ is differentiable on its interior effective domain for $\mathbb{P}$-a.e. $\omega \in \Omega$. Then, K_β is differentiable on $\mathrm{int}(\mathrm{dom}(K_\beta))$ and we have*

$$\nabla K_\beta(\boldsymbol{\theta}) = \int_\Omega \nabla \beta^*(\omega, \boldsymbol{\theta}) \mathrm{d}\mathbb{P}(\omega).$$

Proof. Since K_β is a convex function, we can show its differentiability by showing its directional derivative function is linear (Rockafellar [1970], Theorem 25.2). Let $\boldsymbol{\theta}_0 \in \mathrm{int}(\mathrm{dom}(K_\beta))$. For any other $\boldsymbol{\theta} \in \mathrm{int}(\mathrm{dom}(K_\beta))$, we examine its directional derivative

$$\lim_{t \downarrow 0} \frac{1}{t} \left(K_\beta(\boldsymbol{\theta}_0 + t(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) - K_\beta(\boldsymbol{\theta}_0) \right) = \lim_{t \downarrow 0} \int_\Omega \frac{1}{t} \left(\beta^*(\omega, \boldsymbol{\theta}_0 + t(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) - \beta^*(\omega, \boldsymbol{\theta}_0) \right) \mathrm{d}\mathbb{P}(\omega).$$

Since, $\boldsymbol{\theta}_0 \in \mathrm{int}(\mathrm{dom}(K_\beta))$, we have that $\boldsymbol{\theta}_0 \in \mathrm{int}(\mathrm{dom}(\beta^*(\omega, \cdot)))$ for $\mathbb{P}$-a.e. $\omega \in \Omega$. For these ω, the difference quotient

$$q_\omega(t) \triangleq \frac{1}{t} (\beta^*(\omega, \boldsymbol{\theta}_0 + t(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) - \beta^*(\omega, \boldsymbol{\theta}_0))$$

is non-increasing as $t \downarrow 0$ (Theorem 23.1, Rockafellar [1970]). Therefore, the quotient $q_\omega(t) - q_\omega(\frac{1}{2})$ is negative and non-increasing for $t \leq 1/2$. Hence, applying the monotone convergence theorem and removing the term $q_\omega(\frac{1}{2})$ afterward (note that $q_\omega(\frac{1}{2})$ is a finite number since $\boldsymbol{\theta}, \boldsymbol{\theta}_0 \in \mathrm{dom}(\beta^*(\omega, \cdot))$, for $\mathbb{P}$-a.e. $\omega \in \Omega$) shows that K_β is differentiable at $\boldsymbol{\theta}_0$. Indeed,

$$\begin{aligned}
\lim_{t \downarrow 0} \frac{1}{t} \left(K_\beta(\boldsymbol{\theta}_0 + t(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) - K_\beta(\boldsymbol{\theta}_0) \right) &= \int_\Omega \lim_{t \downarrow 0} \frac{1}{t} (\beta^*(\omega, \boldsymbol{\theta}_0 + t(\boldsymbol{\theta} - \boldsymbol{\theta}_0)) - \beta^*(\omega, \boldsymbol{\theta}_0)) \mathrm{d}\mathbb{P}(\omega) \\
&\overset{(*)}{=} \int_\Omega \langle \boldsymbol{\theta} - \boldsymbol{\theta}_0, \nabla \beta^*(\omega, \boldsymbol{\theta}_0) \rangle \mathrm{d}\mathbb{P}(\omega) \\
&\overset{(**)}{=} \left\langle \boldsymbol{\theta} - \boldsymbol{\theta}_0, \int_\Omega \nabla \beta^*(\omega, \boldsymbol{\theta}_0) \mathrm{d}\mathbb{P}(\omega) \right\rangle,
\end{aligned}$$

where in $(*)$ we used the differentiability of $\beta^*(\omega, \cdot)$ and in $(**)$ we used the linearity of inner product. □

Theorem 2.B.5. *Assume $J_\beta(\mathbf{a}) < +\infty$. Let $\boldsymbol{\theta}^*$ be such that $K_\beta(\boldsymbol{\theta}^*) < \infty$. If the partial derivatives $\mathbf{g}_{\boldsymbol{\theta}^*}(\omega) = \nabla \beta^*(\omega, \boldsymbol{\theta}^*)$ exist for $\mathbb{P}$-almost every ω and satisfy:*

$$\int_\Omega \mathbf{g}_{\boldsymbol{\theta}^*}(\omega) \mathrm{d}\mathbb{P}(\omega) = \mathbf{a}. \tag{2.48}$$

Then, $\boldsymbol{\theta}^*$ is a dual solution and $\mathbf{g}_{\boldsymbol{\theta}^*}$ is a primal solution. On the other hand, if a primal

solution $\mathbf{g}^*$ *exists, then* $\mathbf{g}^* = \mathbf{g}_{\boldsymbol{\theta}^*}$ *for a particular dual solution* $\boldsymbol{\theta}^*$*, if the partial derivatives* $\nabla\beta^*(\omega, \boldsymbol{\theta}^*)$ *at this particular* $\boldsymbol{\theta}^*$ *exists, for* $\mathbb{P}$*-almost every* $\omega \in \Omega$*.*

Proof. By the definition of the conjugate, we have for any $\boldsymbol{\theta} \in \mathbb{R}^d$ and functions $\mathbf{g}$, the inequality

$$\beta(\omega, \mathbf{g}(\omega)) + \beta^*(\omega, \boldsymbol{\theta}) \geq \boldsymbol{\theta}^T \mathbf{g}(\omega). \tag{2.49}$$

Inequality (2.49) is tight if and only if the gradient of $\beta^*(\omega, \cdot)$ at $\boldsymbol{\theta}$ exists and is equal to $\mathbf{g}(\omega)$ ([Rockafellar, 1970], Theorem 23.5). Integrating both sides of (2.49) with respect to ω, for any function $\mathbf{g}$ such that $\int_\Omega \mathbf{g}(\omega)\mathrm{d}\mathbb{P}(\omega) = \mathbf{a}$ yields

$$\int_\Omega \beta(\omega, \mathbf{g}(\omega))\mathrm{d}\mathbb{P}(\omega) + \int_\Omega \beta^*(\omega, \boldsymbol{\theta})\mathrm{d}\mathbb{P}(\omega) \geq \boldsymbol{\theta}^T \mathbf{a}. \tag{2.50}$$

In particular, inequality (2.49) is an equality for $\mathbf{g}_{\boldsymbol{\theta}^*}(\omega) = \nabla\beta^*(\omega, \boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^*$ that satisfies condition (2.48). Hence, we have that (2.50) is also an equality and thus

$$\int_\Omega \beta(\omega, \mathbf{g}_{\boldsymbol{\theta}^*}(\omega))\mathrm{d}\mathbb{P}(\omega) = \mathbf{a}^T\boldsymbol{\theta}^* - \int_\Omega \beta^*(\omega, \boldsymbol{\theta}^*)\mathrm{d}\mathbb{P}(\omega).$$

Therefore, by Theorem 2.B.3 we conclude that $\boldsymbol{\theta}^*$ is a dual solution and $\mathbf{g}_{\boldsymbol{\theta}^*}$ is a primal solution.

Conversely, let $\mathbf{g}^*$ be a primal solution. Then, $J_\beta(\mathbf{a})$ is finite, and thus a dual solution $\boldsymbol{\theta}^*$ exists and we must have $K_\beta(\boldsymbol{\theta}^*) < \infty$. Furthermore, (2.50) holds for $\mathbf{g}^*$ and $\boldsymbol{\theta}^*$ and is an equality. Therefore, we have that the following integral with a positive integrand is zero:

$$\int_\Omega \beta(\omega, \mathbf{g}^*(\omega)) + \beta^*(\omega, \boldsymbol{\theta}^*) - (\mathbf{g}^*(\omega))^T\boldsymbol{\theta}^* \ \mathrm{d}\mathbb{P}(\omega) = 0.$$

Hence, (2.49) is an equality for $\mathbf{g}^*$ and $\boldsymbol{\theta}^*$ for $\mathbb{P}$-a.e. $\omega \in \Omega$. Therefore, $\mathbf{g}^* = \mathbf{g}_{\boldsymbol{\theta}^*}$, by Rockafellar [1970], Theorem 23.5. □

2.C Details for Remark 2.3

In practice, X might depend on many underlying risk factors $Z_1, \ldots, Z_I$, where only the nominal distributions of the marginals are specified and have an explicit form. The following proposition, adapted from Kruse et al. [2019], allows us to choose the divergence based on the nominal distributions of the marginals.

Proposition 2.C.6. *Let $X \in L^1(\mathbb{P}_0)$. Suppose there exists a constant $C > 0$, such that*

$$|X| \leq C\left(1 + \sum_{i=1}^{m} |Z_i|\right), \tag{2.51}$$

holds $\mathbb{P}_0$-almost surely. If there exist $\theta_1, \theta_2 \in \mathbb{R}$ such that, for all $i = 1, \ldots, I$,

$$\mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*\left(\theta_1 + \phi_1^*(\theta_2 + C(1 + m \cdot |Z_i|))\right)\right] < \infty,$$

then the integral conditions (2.9) *and* (2.10) *are also satisfied.*

Proof. Apply Corollary 2.2 and note that

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*\left(\theta_1 + \phi_1^*(\theta_2 - X)\right)\right] &\leq \mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*\left(\theta_1 + \phi_1^*\left(\theta_2 + C(1 + \sum_{i=1}^{m} |Z_i|)\right)\right)\right] \\
&= \mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*\left(\theta_1 + \phi_1^*\left(\frac{1}{m}\sum_{i=1}^{m} \theta_2 + C(1 + m|Z_i|)\right)\right)\right] \\
&\leq \mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*\left(\theta_1 + \frac{1}{m}\sum_{i=1}^{m} \phi_1^*\left(\theta_2 + C(1 + m|Z_i|)\right)\right)\right] \\
&= \mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*\left(\frac{1}{m}\sum_{i=1}^{m} \theta_1 + \phi_1^*\left(\theta_2 + C(1 + m|Z_i|)\right)\right)\right] \\
&\leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbb{P}_0}\left[\phi_2^*\left(\theta_1 + \phi_1^*\left(\theta_2 + C(1 + m|Z_i|)\right)\right)\right],
\end{aligned}$$

where we used Jensen's inequality and the monotonicity of the conjugate functions ϕ_1^*, ϕ_2^*. □

2.D Characterization of Risk Aversion of Robust OCE and its Preservation of Convex Order

The formulation of the robust OCE and robust shortfall risk measures in terms of (ϕ_1^*, ϕ_2^*) also allows us to characterize risk aversion by simple convexity conditions imposed on ϕ_1^*, ϕ_2^*. We say that a robust risk measure $\rho \in \{\rho^l_{\mathrm{oce},\mathbb{P}}, \rho^l_{\mathrm{sf},\mathbb{P}}\}$ is risk-averse, if and only if

$$\rho(X) \geq \mathbb{E}_{\mathbb{P}}[-X], \forall X. \tag{2.52}$$

We show that for the robust OCE and shortfall risk measures, the risk aversion property can be easily characterized by some mild conditions on ϕ_1^*, ϕ_2^*.

Proposition 2.D.7. *For any non-decreasing convex function ϕ_1^*, ϕ_2^* with $\phi_1^*(0) = \phi_2^*(0) =$*

0, *we have that* $\rho^l_{\text{oce},\mathbb{P}}$ *satisfies* (2.52)*, if and only if*

$$\phi_2^*(x) \geq x, \phi_1^*(x) \geq x, \forall x \in \mathbb{R}.$$

If we further assume that $\phi_1^*(x) > 0, \forall x > 0$*, then we have that* $\rho^l_{\text{sf},\mathbb{P}}$ *satisfies* (2.52)*, if and only if*

$$\phi_2^*(x) \geq x, \forall x \in \mathbb{R}.$$

Proof of Proposition 2.D.7. For $\rho^l_{\text{oce},\mathbb{P}}$, we have that (2.52) implies that for all $x \in \mathbb{R}$, that

$$\inf_{\theta_1,\theta_2 \in \mathbb{R}} \{-\theta_1 - \theta_2 + \phi_2^*(\phi_1^*(\theta_2 - x) + \theta_1)\} \geq -x,$$

which implies in particular that for $\theta_2 = x$:

$$\phi_2^*(\theta_1) \geq \theta_1, \forall \theta_1 \in \mathbb{R}.$$

It also implies that for $\theta_1 = -\phi_1^*(x), \theta_2 = 2x$, that

$$\phi_1^*(x) \geq x, \forall x \in \mathbb{R}.$$

Conversely, if $\phi_2^*(x) \geq x, \phi_1^*(x) \geq x, \forall x \in \mathbb{R}$, then by the monotonicity of ϕ_1^*, ϕ_2^*, we have that for any $\theta_1, \theta_2 \in \mathbb{R}$ and any X:

$$-\theta_1 - \theta_2 + \mathbb{E}_{\mathbb{P}}[\phi_2^*(\phi_1^*(\theta_2 - X) + \theta_1)] \geq -\theta_1 - \theta_2 + \mathbb{E}_{\mathbb{P}}[\theta_2 - X + \theta_1] \geq \mathbb{E}_{\mathbb{P}}[-X].$$

For $\rho^l_{\text{sf},\mathbb{P}}$, we have that (2.52) implies that for all $x \in \mathbb{R}$,

$$\inf_{\theta_1,\theta_2 \in \mathbb{R}} \{\theta_2 \mid \phi_2^*(\phi_1^*(-\theta_2 - x) + \theta_1) \leq \theta_1\} \geq -x.$$

This implies that for all $x, \theta_1 \in \mathbb{R}$,

$$\phi_2^*(\phi_1^*(-(-x) - x) + \theta_1) = \phi_2^*(\theta_1) \geq \theta_1.$$

On the other hand, if $\phi_2^*(x) \geq x$ for all $x \in \mathbb{R}$, then for any $\theta_1 \in \mathbb{R}$ and any random variable X,

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[\phi_2^*(\phi_1^*(-\theta_2 - X) + \theta_1)] \leq \theta_1 &\Rightarrow \mathbb{E}_{\mathbb{P}}[\phi_1^*(-\theta_2 - X)] \leq 0 \\ &\Rightarrow \phi_1^*(-\theta_2 + \mathbb{E}[-X]) \leq 0.\end{aligned}$$

By the non-decreasing property of ϕ_1^* and the assumption that $\phi_1^*(x) > 0, \forall x > 0$, we can then conclude that the above inequality implies that $\theta_2 \geq \mathbb{E}[-X]$. Therefore, $\rho^l_{\text{sf},\mathbb{P}}(X) \geq$

$\mathbb{E}_{\mathbb{P}}[-X]$. □

Furthermore, it is also easy to show that the robust OCE and shortfall risk measure preserve convex order. We say that X is convex less order than Y, i.e. $X \leq_{\text{cv}} Y$ if for all convex function f, we have $\mathbb{E}_{\mathbb{P}}[f(X)] \leq \mathbb{E}_{\mathbb{P}}[f(Y)]$.

Proposition 2.D.8. *Let ϕ_1^*, ϕ_2^* be a non-decreasing convex function. Then*

$$X \leq_{\text{cv}} Y \Rightarrow \rho^l_{\text{oce},\mathbb{P}}(X) \leq \rho^l_{\text{oce},\mathbb{P}}(Y) \\ \text{and } \rho^l_{\text{sf},\mathbb{P}}(X) \leq \rho^l_{\text{sf},\mathbb{P}}(Y).$$

Proof of Proposition 2.D.8. Suppose $X \leq_{\text{cv}} Y$, then, we have that for all $\theta_1, \theta_2 \in \mathbb{R}$,

$$\mathbb{E}_{\mathbb{P}}[\phi_2^*(\phi_1^*(\theta_2 - X) + \theta_1)] \geq \mathbb{E}_{\mathbb{P}}[\phi_2^*(\phi_1^*(\theta_2 - Y + \theta_1)],$$

since the function $\phi_2^*(\phi_1^*(\theta_2 - x) + \theta_1)$ is convex in x. Hence, $\rho^l_{\text{oce},\mathbb{P}}(X) \leq \rho^l_{\text{oce},\mathbb{P}}(Y)$.

Similarly, the function $\phi_2^*(\phi_1^*(-\theta_2 - x) + \theta_1)$ is also convex in x. Therefore, for any $\theta_1, \theta_2 \in \mathbb{R}$, if $X \leq_{\text{cv}} Y$, then

$$\mathbb{E}[\phi_2^*(\phi_1^*(-\theta_2 - Y) + \theta_1)] \leq \theta_1 \Rightarrow \mathbb{E}[\phi_2^*(\phi_1^*(-\theta_2 - X) + \theta_1)] \leq \theta_1.$$

Hence, $\rho^l_{\text{sf},\mathbb{P}}(X) \leq \rho^l_{\text{sf},\mathbb{P}}(Y)$. □

2.E Additional Details on Tail Analysis

In this section, we examine conditions (2.9) and (2.10) for the combinations of ϕ_2-divergences and nominal distributions f_0 (distribution of X under $\mathbb{P}_0$) that are displayed in Table 2.1 and 2.2, for respectively the Conditional-Value-at-Risk and the entropic risk measure. We examine the following nominal distributions $f_0(x)$:

- Gaussian $N(\mu, \sigma)$: $f_0(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$
- Lognormal $\ln(\mu, \sigma^2)$: $f_0(x) = \frac{1}{|x|\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log|x|-\mu)^2}{2\sigma^2}\right), \quad x < 0$
- Weibull $W(\lambda_0, k)$: $f_0(x) = \frac{k}{\lambda_0}\left(\frac{|x|}{\lambda_0}\right)^{k-1} \exp\left(-\left(\frac{|x|}{\lambda_0}\right)^k\right), \quad x < 0$
- Pareto $P(x_m, p_0)$: $f_0(x) = \frac{p|x_m|^{p_0}}{|x|^{p_0+1}}, \quad x \leq x_m$
- Student-t $t(\nu)$: $f_0(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}$

For the divergences, we have the following conjugate functions (see Ben-Tal et al. 2013):

- Kullback-Leibler: $\phi_2^*(s) = \exp(s) - 1$

- Polynomial $p > 1$: $\phi_2^*(s) = \frac{1}{p}\max\{1 + s(p-1), 0\}^{\frac{p}{p-1}} - \frac{1}{p}$
- Modified χ^2-divergence: $\phi_2^*(s) = \begin{cases} -1 & s < -2 \\ s + s^2/4 & s \geq -2. \end{cases}$

2.E.1 Details for Table 2.1

For CVaR_α, we have that $\phi_1^*(s) = \max\{s/(1-\alpha), 0\}$. We also note that all nominal distributions considered in Table 2.1 have $f_0(x) > 0$ for all $x \in (-\infty, c)$ for some $c \in \mathbb{R}$. Hence, for all divergences such that $\phi_2^*(s) = \infty$ if s sufficiently large (such as Burg entropy, variation distance), the corresponding robust CVaR_α risk measure is not finite.

- Gaussian vs KL-divergence: as $x \to -\infty$, we have

$$\begin{aligned} &O\left(\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2 - x) + \theta_1}{\lambda}\right) f_0(x)\right) \\ &= O\left(\lambda\left(\exp\left(\frac{\frac{\theta_2+|x|}{1-\alpha} + \theta_1}{\lambda}\right) - 1\right)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) \\ &= O\left(\lambda\exp\left(\frac{|x|}{\lambda(1-\alpha)} - \frac{(x-\mu)^2}{2\sigma^2}\right)\right), \end{aligned}$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$. As $x \to \infty$, we have $\phi_1^*(\theta_2 - x) = 0$ for all $x > \theta_2$. Then, the above integrand is also integrable.

- Pareto vs KL-divergence: as $x \to -\infty$, we have

$$\begin{aligned} &O\left(\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2 - x) + \theta_1}{\lambda}\right) f_0(x)\right) \\ &= O\left(\lambda\left(\exp\left(\frac{\frac{\theta_2+|x|}{1-\alpha} + \theta_1}{\lambda}\right) - 1\right)\frac{p_0|x_m|^{p_0}}{|x|^{p_0+1}}\right) \\ &= O\left(\lambda\exp\left(\frac{|x|}{\lambda(1-\alpha)}\right)|x|^{-p_0-1}\right), \end{aligned}$$

  which is not integrable for any $\theta_1, \theta_2 \in \mathbb{R}$, $\lambda \geq 0$.

- Log-normal vs KL-divergence: for $x \to -\infty$,

$$\begin{aligned} &O\left(\lambda\left(\exp\left(\frac{\frac{\theta_2+|x|}{1-\alpha} + \theta_1}{\lambda}\right) - 1\right)\frac{1}{|x|\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log|x| - \mu)^2}{2\sigma^2}\right)\right) \\ &= O\left(\lambda\exp\left(\frac{|x|}{\lambda(1-\alpha)}\right) - \frac{(\log|x| - \mu)^2}{2\sigma^2} - \log|x|\right), \end{aligned}$$

  which is not integrable for any $\theta_1, \theta_2 \in \mathbb{R}$, $\lambda \geq 0$.

- Weibull vs KL-divergence: we have the tail behavior

$$O\left(\lambda\left(\exp\left(\frac{\frac{\theta_2+|x|}{1-\alpha}+\theta_1}{\lambda}\right)-1\right)\frac{k}{\lambda_0}\left(\frac{|x|}{\lambda_0}\right)^{k-1}\exp\left(-\left(\frac{|x|}{\lambda_0}\right)^{k}\right)\right).$$

  For $\lambda > 0$, this is integrable if $k > 1$, and not integrable if $k < 1$. For $k = 1$, we note that if we choose any $\lambda > \frac{\lambda_0}{1-\alpha}$, then $\frac{1}{\lambda(1-\alpha)} - \frac{1}{\lambda_0} < 0$. Hence, condition (2.9) is satisfied.

- Student's t vs KL-divergence: As $x \to -\infty$, we have

$$O\left(\lambda\left(\exp\left(\frac{\frac{\theta_2+|x|}{1-\alpha}+\theta_1}{\lambda}\right)-1\right)\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right).$$

  Clearly, the exponential term $\exp\left(\frac{|x|}{\lambda(1-\alpha)}\right)$ dominates the polynomial term $\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$. Hence, we have no integrability for any $\theta_1, \theta_2 \in \mathbb{R},\ \lambda \geq 0$.

- Gaussian vs Polynomial $p > 1$, we examine the tail behavior for $x \to -\infty$ (the case for $x \to \infty$ follows similarly, where $\phi_1^*(\theta_2 - x) = 0$ for all $x > \theta_2$), we have

$$O\left(\lambda\left(1+\left(\frac{\theta_2+|x|}{\lambda(1-\alpha)}+\frac{\theta_1}{\lambda}\right)(p-1)\right)^{\frac{p}{p-1}}\exp\left(-(x-\mu)^2/(2\sigma^2)\right)\right),$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}$, $\lambda > 0$.

- Log-normal vs Polynomial $p > 1$, we study the tail behavior

$$\begin{aligned}
&O\left(\lambda\left(1+\left(\frac{\theta_2+|x|}{\lambda(1-\alpha)}+\frac{\theta_1}{\lambda}\right)(p-1)\right)^{\frac{p}{p-1}}\frac{1}{|x|\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log|x|-\mu)^2}{2\sigma^2}\right)\right)\\
&= O\left(|x|^{p/(p-1)-1}\exp\left(-\frac{(\log|x|-\mu)^2}{2\sigma^2}\right)\right)\\
&= O\left(\exp\left(-\frac{\log^2|x|}{2\sigma^2}+\frac{1}{p-1}\log(x)\right)\right),
\end{aligned}$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}$, $\lambda > 0$.

- Weibull vs Polynomial $p > 1$, the tail behavior is

$$O\left(\lambda\left(1+\left(\frac{\theta_2+|x|}{\lambda(1-\alpha)}+\frac{\theta_1}{\lambda}\right)(p-1)\right)^{\frac{p}{p-1}}\frac{k}{\lambda_0}\left(\frac{|x|}{\lambda_0}\right)^{k-1}\exp\left(-\left(\frac{|x|}{\lambda_0}\right)^{k}\right)\right).$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}$, $\lambda > 0$, since $k > 0$.

- Pareto vs Polynomial $p > 1$, the tail behavior is

$$O\left(\lambda\left(1+\left(\frac{\theta_2+|x|}{\lambda(1-\alpha)}+\frac{\theta_1}{\lambda}\right)(p-1)\right)^{\frac{p}{p-1}}\frac{p_0|x_m|^{p_0}}{|x|^{p_0+1}}\right)$$
$$= O\left(|x|^{p/(p-1)-p_0-1}\right),$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}$, $\lambda > 0$, if and only if $\frac{p}{p-1} < p_0$.

- Student's t vs Polynomial $p > 1$: As $x \to -\infty$, we have

$$O\left(\lambda\left(1+\left(\frac{\theta_2+|x|}{\lambda(1-\alpha)}+\frac{\theta_1}{\lambda}\right)(p-1)\right)^{\frac{p}{p-1}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right)$$
$$= O\left(|x|^{p/(p-1)-\nu-1}\right),$$

  which is integrable if and only if $p/(p-1) < \nu$.

- Modified χ^2 vs Gaussian: we have

$$O\left(\lambda\left(\frac{\frac{\theta_2+|x|}{1-\alpha}+\theta_1}{\lambda}\right)^2\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right),$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$.

- Log-normal vs Modified χ^2: we have

$$O\left(\lambda\left(\frac{\frac{\theta_2+|x|}{1-\alpha}+\theta_1}{\lambda}\right)^2\frac{1}{|x|}\exp\left(-\frac{(\log|x|-\mu)^2}{2\sigma^2}\right)\right),$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$.

- Pareto vs Modified χ^2: as $x \to -\infty$, we have

$$O\left(\lambda\left(\frac{\frac{\theta_2+|x|}{1-\alpha}+\theta_1}{\lambda}\right)^2|x|^{-p_0-1}\right),$$

  which is integrable if and only if $2-1-p_0 < -1$, which is equivalent to $p_0 > 2$.

- Modified χ^2 vs Weibull: as $x \to -\infty$, we have

$$O\left(\lambda\left(\frac{\frac{\theta_2+|x|}{1-\alpha}+\theta_1}{\lambda}\right)^2\left(\frac{|x|}{\lambda_0}\right)^{k-1}\exp\left(-\left(\frac{|x|}{\lambda_0}\right)^k\right)\right),$$

  which is integrable for $k > 0$ and any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$.

- Student's t vs Modified χ^2: as $x \to -\infty$, we have

$$O\left(\lambda\left(\frac{\frac{\theta_2+|x|}{1-\alpha}+\theta_1}{\lambda}\right)^2\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right),$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$ if and only if $\nu > 2$.

2.E.2 Details for Table 2.2

For the entropic risk measure $\rho_{e,\gamma}(X) = \log\left(\mathbb{E}[\exp(-\gamma X)]\right)/\gamma$, $\gamma > 0$, we have $\phi_2^*(s) = (\exp(s/\gamma) - 1)/\gamma$. We note that Log-normal, Pareto, Weibull with $k < 1$, and Student's t distribution do not have a finite moment generating function. Hence, the robust entropic risk measure for those nominal distribution is not finite for any divergences. We also note that for all divergences such that $\phi_2^*(s) = \infty$ if s is sufficiently large (such as Burg entropy, variation distance), the corresponding robust entropic risk measure is not finite, for the nominal distributions that we consider in Table 2.2.

- Weibull vs KL-divergence: we have

$$\begin{aligned}
&O\left(\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2 - x) + \theta_1}{\lambda}\right) f_0(x)\right) \\
&= O\left(\lambda \exp\left(\frac{\frac{1}{\gamma}(\exp((\theta_2 - x)/\gamma - 1)) + \theta_1}{\lambda}\right)\left(\frac{|x|}{\lambda_0}\right)^{k-1} \exp\left(-\left(\frac{|x|}{\lambda_0}\right)^k\right)\right),
\end{aligned}$$

  which is not integrable for any $k > 0$, and $\theta_1, \theta_2 \in \mathbb{R}, \lambda \geq 0$.

- Gaussian vs KL-divergence: we have

$$O\left(\lambda \exp\left(\frac{\frac{1}{\gamma}(\exp((\theta_2 - x)/\gamma - 1)) + \theta_1}{\lambda}\right)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right),$$

  which is not integrable for any $\theta_1, \theta_2 \in \mathbb{R}, \lambda \geq 0$.

- Weibull vs Polynomial $p > 1$: we have

$$O\left(\left(\frac{1}{\gamma}(\exp((\theta_2 - x)/\gamma - 1))\right)^{\frac{p}{p-1}}\left(\frac{|x|}{\lambda_0}\right)^{k-1}\exp\left(-\left(\frac{|x|}{\lambda_0}\right)^k\right)\right).$$

  We note that the integrability of the integrand above is mainly determined by the growth of

$$\exp\left(\frac{|x|}{\gamma}\frac{p}{p-1} - \left(\frac{|x|}{\lambda_0}\right)^k\right)$$

as $x \to -\infty$.

If $k > 1$, then we clearly have integrability. If $k = 1$, we have integrability if and only if $\frac{p}{p-1} < \frac{\gamma}{\lambda_0}$.

- Gaussian vs Polynomial $p > 1$: we have

$$O\left(\left(\frac{\frac{1}{\gamma}\left(\exp\left(\frac{\theta_2 - x}{\gamma}\right) - 1\right) + \theta_1}{\lambda}(p-1)\right)^{\frac{p}{p-1}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right),$$

the tail behavior of this integrand is determined by the growth of

$$\exp\left(\frac{|x|}{\gamma}\frac{p}{p-1} - \frac{(x-\mu)^2}{2\sigma^2}\right),$$

which is integrable.

- Weibull vs Modified χ^2: we have

$$O\left(\lambda\left(\frac{\frac{1}{\gamma}(\exp((\theta_2 - x)/\gamma - 1)) + \theta_1}{\lambda}\right)^2 \left(\frac{|x|}{\lambda_0}\right)^{k-1} \exp\left(-\left(\frac{|x|}{\lambda_0}\right)^k\right)\right),$$

for which the integrability depends on

$$\exp\left(\frac{|x|^2}{\gamma^2} - \frac{|x|^k}{\lambda_0^k}\right)\left(\frac{|x|}{\lambda_0}\right)^{k-1}.$$

For $k > 2$, we have integrability, whereas for $k < 2$ we do not. If $k = 2$, then this is integrable if and only if $\gamma > \lambda_0$.

- Gaussian vs Modified χ^2: we have

$$O\left(\lambda\left(\frac{\frac{1}{\gamma}(\exp((\theta_2 - x)/\gamma - 1)) + \theta_1}{\lambda}\right)^2 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right),$$

for which the integrability depends on

$$\exp\left(\frac{x^2}{\gamma^2} - \frac{x^2}{2\sigma^2}\right),$$

which is integrable if and only if $\gamma > 2\sigma^2$.

2.E.3 Details for Table 2.3

We consider the OCE risk measure with the following quadratic utility function as in Example 2.2 of Ben-Tal and Teboulle [2007]:

$$u(x) = \begin{cases} x - \frac{1}{2}x^2 & x < 1 \\ \frac{1}{2} & x \geq 1. \end{cases}$$

By the relation $\phi_1^*(s) = -u(-s)$, we have that

$$\phi_1^*(s) = \begin{cases} s + \frac{1}{2}s^2 & s > -1 \\ -\frac{1}{2} & s \leq -1. \end{cases}$$

- Gaussian vs KL-divergence: as $x \to -\infty$, we have

$$\begin{aligned} & O\left(\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2 - x) + \theta_1}{\lambda}\right) f_0(x)\right) \\ &= O\left(\lambda\left(\exp\left(\frac{(\theta_2 - x)^2/2 + (\theta_2 - x) + \theta_1}{\lambda}\right) - 1\right)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) \\ &= O\left(\lambda\exp\left(\frac{(x-\theta_2)^2}{2\lambda} - \frac{(x-\mu)^2}{2\sigma^2}\right)\right), \end{aligned}$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}$ by choosing $\lambda > \sigma^2$. As $x \to \infty$, we have $\phi_1^*(\theta_2 - x) = -1/2$ for all $x > \theta_2 + 1$. Then, the above integrand is also integrable.

- Pareto vs KL-divergence: as $x \to -\infty$, we have

$$\begin{aligned} & O\left(\lambda\phi_2^*\left(\frac{\phi_1^*(\theta_2 - x) + \theta_1}{\lambda}\right) f_0(x)\right) \\ &= O\left(\lambda\left(\exp\left(\frac{(\theta_2 - x)^2/2 + (\theta_2 - x) + \theta_1}{\lambda}\right) - 1\right)\frac{p_0|x_m|^{p_0}}{|x|^{p_0+1}}\right) \\ &= O\left(\lambda\exp\left(\frac{x^2}{2\lambda}\right)|x|^{-p_0-1}\right), \end{aligned}$$

  which is not integrable for any $\theta_1, \theta_2 \in \mathbb{R}$, $\lambda \geq 0$.

- Log-normal vs KL-divergence: for $x \to -\infty$,

$$\begin{aligned} & O\left(\lambda\left(\exp\left(\frac{(\theta_2 - x)^2/2 + (\theta_2 - x) + \theta_1}{\lambda}\right) - 1\right)\frac{1}{|x|\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log|x| - \mu)^2}{2\sigma^2}\right)\right) \\ &= O\left(\lambda\exp\left(\frac{(\theta_2 - x)^2}{2\lambda} - \frac{(\log|x| - \mu)^2}{2\sigma^2} - \log|x|\right)\right), \end{aligned}$$

  which is not integrable for any $\theta_1, \theta_2 \in \mathbb{R}$, $\lambda \geq 0$.

- Weibull vs KL-divergence: we have the tail behavior

$$O\left(\lambda\left(\exp\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)+\theta_1}{\lambda}\right)-1\right)\frac{k}{\lambda_0}\left(\frac{|x|}{\lambda_0}\right)^{k-1}\exp\left(-\left(\frac{|x|}{\lambda_0}\right)^k\right)\right).$$

  This is integrable if $k>2$, and not integrable if $k<2$. For $k=2$, it is integrable if we choose any $\lambda>\lambda_0/2$.

- Student's t vs KL-divergence: As $x\to-\infty$, we have

$$O\left(\lambda\left(\exp\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)+\theta_1}{\lambda}\right)-1\right)\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right),$$

  which is not integrable for any $\theta_1,\theta_2\in\mathbb{R},\ \lambda\geq 0$.

- Gaussian vs Polynomial $p>1$, we examine the tail behavior for $x\to-\infty$ (the case for $x\to\infty$ follows similarly, where $\phi_1^*(\theta_2-x)=0$ for all $x>\theta_2$), we have

$$O\left(\lambda\left(1+\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)}{\lambda}+\frac{\theta_1}{\lambda}\right)(p-1)\right)^{\frac{p}{p-1}}\exp\left(-(x-\mu)^2/(2\sigma^2)\right)\right),$$

  which is integrable for any $\theta_1,\theta_2\in\mathbb{R}$, $\lambda>0$.

- Log-normal vs Polynomial $p>1$, we study the tail behavior

$$O\left(\lambda\left(1+\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)}{\lambda}+\frac{\theta_1}{\lambda}\right)(p-1)\right)^{\frac{p}{p-1}}\frac{1}{|x|\sigma\sqrt{2\pi}}\exp\left(-\frac{(\log|x|-\mu)^2}{2\sigma^2}\right)\right),$$

  which is integrable for any $\theta_1,\theta_2\in\mathbb{R}$, $\lambda>0$.

- Weibull vs Polynomial $p>1$, the tail behavior is

$$O\left(\lambda\left(1+\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)}{\lambda}+\frac{\theta_1}{\lambda}\right)(p-1)\right)^{\frac{p}{p-1}}\frac{k}{\lambda_0}\left(\frac{|x|}{\lambda_0}\right)^{k-1}\exp\left(-\left(\frac{|x|}{\lambda_0}\right)^k\right)\right),$$

  which is integrable for any $\theta_1,\theta_2\in\mathbb{R}$, $\lambda>0$, since $k>0$.

- Pareto vs Polynomial $p>1$, the tail behavior is

$$\begin{aligned}&O\left(\lambda\left(1+\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)}{\lambda}+\frac{\theta_1}{\lambda}\right)(p-1)\right)^{\frac{p}{p-1}}\frac{p_0|x_m|^{p_0}}{|x|^{p_0+1}}\right)\\&=O\left(|x|^{2p/(p-1)-p_0-1}\right),\end{aligned}$$

  which is integrable for any $\theta_1,\theta_2\in\mathbb{R}$, $\lambda>0$, if and only if $\frac{2p}{p-1}<p_0$.

- Student's t vs Polynomial $p > 1$: As $x \to -\infty$, we have,

$$O\left(\lambda\left(1+\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)}{\lambda}+\frac{\theta_1}{\lambda}\right)(p-1)\right)^{\frac{p}{p-1}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right)$$
$$= O\left(|x|^{2p/(p-1)-\nu-1}\right),$$

  which is integrable if and only if $2p/(p-1) < \nu$.

- Modified χ^2 vs Gaussian: we have

$$O\left(\lambda\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)+\theta_1}{\lambda}\right)^2\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right),$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$.

- Log-normal vs Modified χ^2: we have

$$O\left(\lambda\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)+\theta_1}{\lambda}\right)^2\frac{1}{|x|}\exp\left(-\frac{(\log|x|-\mu)^2}{2\sigma^2}\right)\right),$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$.

- Pareto vs Modified χ^2: as $x \to -\infty$, we have

$$O\left(\lambda\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)+\theta_1}{\lambda}\right)^2|x|^{-p_0-1}\right),$$

  which is integrable if and only if $4-1-p_0 < -1$, which is equivalent to $p_0 > 4$.

- Modified χ^2 vs Weibull: as $x \to -\infty$, we have

$$O\left(\lambda\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)+\theta_1}{\lambda}\right)^2\left(\frac{|x|}{\lambda_0}\right)^{k-1}\exp\left(-\left(\frac{|x|}{\lambda_0}\right)^k\right)\right),$$

  which is integrable for $k > 0$ and any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$.

- Student's t vs Modified χ^2: as $x \to -\infty$, we have

$$O\left(\lambda\left(\frac{(\theta_2-x)^2/2+(\theta_2-x)+\theta_1}{\lambda}\right)^2\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}\right),$$

  which is integrable for any $\theta_1, \theta_2 \in \mathbb{R}, \lambda > 0$ if and only if $\nu > 4$.

Chapter 3

Sample Average Approximation of Risk Functionals: Non-Asymptotic Error Bounds

3.1 Introduction

Many decision problems in finance, microeconomics and operations research suppose that the decision maker has preferences that are represented by a real-valued *risk measure* ρ, in the sense that a random loss X is preferred to Y, if and only if $\rho(X) \leq \rho(Y)$. An optimal decision can therefore be formulated as a solution of the following risk minimization problem:

$$\inf_{\theta \in \Theta} \rho\big(G(\theta, Z)\big), \tag{3.1}$$

where $\Theta \subset \mathbb{R}^m$ is a compact space of decisions, Z is a d-dimensional random vector with distribution $\mathbb{P}^Z$, and $G(\theta, Z)$ is a goal function. Depending on the choice of ρ, Θ and G, problem (3.1) encompasses a wide range of decision problems, including portfolio optimization, inventory management, as well as learning problems such as (generalized) least squares regressions.

A simple example of ρ is the expected value, in which case the decision maker is supposed to be *risk neutral*. This assumption is, however, too idealistic, and one is often more inclined to work with a ρ that accounts for *risk aversion*. Besides the expected value, other popular choices of ρ include expected utility evaluations, optimized certainty equivalents, and distortion risk measures. In this paper, we assume that ρ is a member of the class of *law-invariant, convex* risk measures, which encompasses all aforementioned examples and more (see e.g., Föllmer and Schied [2016], Föllmer and Knispel [2013]).

In general, the parameterized distribution of the goal function G is unknown, and the only available information must be inferred from an i.i.d. sample, formally expressed by a sequence $(Z_j)_{j \in \mathbb{N}}$ of independent d-dimensional random vectors on a probability space

$(\Omega, \mathcal{F}, \mathbb{P})$ that are identically distributed as the d-dimensional random vector Z. By law-invariance, we may associate ρ with a functional $\mathcal{R}_\rho$ on sets of distribution functions. In this case, (3.1) reads as follows:

$$\inf_{\theta\in\Theta} \mathcal{R}_\rho(F_\theta),$$

where F_θ is the distribution function of $G(\theta, Z)$. Based on the i.i.d. sample $(Z_1, \ldots, Z_n)$, we may replace any distribution function F_θ with its empirical counterpart $\widehat{F}_{n,\theta}$, defined by

$$\widehat{F}_{n,\theta}(t) \triangleq \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{(-\infty,t]}\big(G(\theta, Z_j)\big).$$

Then the original optimization problem (3.1) may be approximated by the following one, which is generally known as the Sample Average Approximation (SAA):

$$\inf_{\theta\in\Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta}) \quad (n \in \mathbb{N}). \tag{3.2}$$

An interesting and highly relevant question is the following: How good is the SAA approximation? Since (3.2) depends on each realization of the i.i.d. samples, one way to answer this question is to examine the following deviation probability:

$$\mathbb{P}\Big(\Big\{\big|\inf_{\theta\in\Theta} \mathcal{R}_\rho\big(\widehat{F}_{n,\theta}\big) - \inf_{\theta\in\Theta} \mathcal{R}_\rho\big(F_\theta\big)\big| \geq \epsilon\Big\}\Big) \quad (n \in \mathbb{N}, \epsilon > 0). \tag{3.3}$$

The deviation probability provides a means to quantify the approximation error of SAA by measuring the likelihood that the true optimal value $\inf_{\theta\in\Theta} \mathcal{R}_\rho\big(F_\theta\big)$ lies within the interval $\inf_{\theta\in\Theta} \mathcal{R}_\rho\big(\widehat{F}_{n,\theta}\big) \pm \epsilon$. It is therefore crucial to have an upper estimate on (3.3) that depends explicitly on the sample size n (which we refer to as a non-asymptotic upper bound), since this would shed light on the sample size n that is required to achieve a given confidence level $\beta > 0$ for the interval $\inf_{\theta\in\Theta} \mathcal{R}_\rho\big(\widehat{F}_{n,\theta}\big) \pm \epsilon$. A simple way to obtain an upper bound on the deviation probability is through Markov's inequality, which estimates (3.3) by its expected deviation:

$$\mathbb{P}\Big(\Big\{\big|\inf_{\theta\in\Theta} \mathcal{R}_\rho\big(\widehat{F}_{n,\theta}\big) - \inf_{\theta\in\Theta} \mathcal{R}_\rho\big(F_\theta\big)\big| \geq \epsilon\Big\}\Big) \leq \mathbb{E}\Big|\inf_{\theta\in\Theta} \mathcal{R}_\rho\big(\widehat{F}_{n,\theta}\big) - \inf_{\theta\in\Theta} \mathcal{R}_\rho\big(F_\theta\big)\Big|\Big/\epsilon.$$

Then, the expected deviation can be further estimated using upper bounds that are obtained, for example, in Bartl and Tangpi [2022]. However, the best possible convergence rate that can be achieved for the expected deviation is known to be $1/\sqrt{n}$, since the expected value itself is a law-invariant risk measure. Therefore, an estimation of (3.3) based on Markov's inequality and expected deviation can be too conservative for practical purposes, as it implies that the sample size n must be at the order of $1/(\beta\epsilon)^2$ to achieve a confidence level β for the deviation probability, which can be very large even for modestly small ϵ and β.

In the risk neutral case, where ρ is the expected value, it is widely known that tighter estimation than Markov's inequality on the deviation probability can be obtained. For example, the classical Hoeffding's inequality Hoeffding [1963] asserts that for any distributions with bounded support, one has

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right| \geq \epsilon\right) \leq 2\exp\left(-n\epsilon^2/B^2\right), \tag{3.4}$$

where $B > 0$ is such that $|X| \leq B$ almost-surely. The bound in (3.4) is more efficient than Markov's inequality, since it reduces the sample size n to $\log(2/\beta)B^2/\epsilon^2$ for a given confidence level β. Similar bounds for stochastic optimization problems involving expected values have also been established in Krätschmer [2024a] using tools from empirical process theory and Talagrand's inequality (see Talagrand 1994, 1996). However, to the best of our knowledge, there are currently no universal approaches on how to obtain explicit non-asymptotic upper bounds for the deviation probability (3.3) that are analogous to (3.4), for general law-invariant convex risk measures and non-linear goal functions G. Indeed, past works such as Bartl and Tangpi [2022] and Krätschmer [2024a] have obtained bounds on (3.3) for only subclasses of ρ and G (see Section 3.1.1 for a more detailed discussion on related literature). Moreover, the literature on SAA for risk-averse stochastic optimization has primarily focused on risk measures that are closely related to expected utility theory (such as the optimized certainty equivalent and the robust expected utility, see Ben-Tal and Teboulle [2007], Duchi and Namkoong [2021], Föllmer and Schied [2016]), while little attention has been devoted to the important class of distortion risk measures that are motivated by rank-dependent utility theory (see Quiggin 1982, Schmeidler 1986, 1989). A major reason for this is that distortion risk measures are, in general, more difficult to handle due to their non-linearity in probability.

The main goal of this paper is to address the aforementioned gaps in the literature on non-asymptotic upper bounds for (3.3). In particular, we pay special attention to distortion risk measures, for which the conditions of our main results can be simplified significantly. For a clear presentation, we summarize our non-asymptotic upper bound on the deviation probability (3.3) in the following informal representation of our main theorem:

Main Theorem (Informal). Under certain regularity conditions imposed on the risk measure ρ, and the assumption that the goal function $G(\theta, z)$ is uniformly bounded by some $B > 0$ for all $\theta \in \Theta$ and $z \in \mathbb{R}^d$, if the function class $\{G(\theta, \cdot \mid \theta \in \Theta\}$ has a finite complexity measure (defined by a uniform entropy integral), then there exists a $q \geq 1$ that is determined by the risk measure ρ, and constants $C_1, C_2 > 0$ that can be calculated

explicitly, such that

$$\mathbb{P}\Big(\Big\{\big|\inf_{\theta\in\Theta}\mathcal{R}_\rho\big(\widehat{F}_{n,\theta}\big)-\inf_{\theta\in\Theta}\mathcal{R}_\rho\big(F_\theta\big)\big|\geq\epsilon\Big\}\Big)\leq\exp\left(-C_1n\epsilon^{2q}/B^{2q}\right), \tag{3.5}$$

for all $\epsilon > C_2/n^{1/(2q)}$.

A clear distinction between our bound (3.5) for general law-invariant risk measures and Hoeffding's bound (3.4) for expected values is the dependence on a parameter $q \geq 1$, which measures how much a risk measure ρ overweights the right-tail probability for large loss realizations. This indicates that the SAA convergence rate for risk measures, *uniformly* over all distributions, is at least $1/n^{1/(2q)}$, which is worse than the usual rate $1/\sqrt{n}$ unless $q = 1$ holds for the risk measure. The dependence on q is not due to a deficiency in our estimation method, but rather the existence of distortion risk measures that overweight the right-tail probability by a power of $1/q$. This is stated formally in Subsection 3.3.4, where we provide a lower bound on the supremum of the deviation probability (3.3) over all distributions.

In summary, our contributions are as follows.

- We provide non-asymptotic upper estimation (3.5) for the deviation probability (3.3) that (i) holds for a general class of convex law-invariant risk measures, (ii) allows for non-linear goal functions, and (iii) features explicit constants, enabling the construction of confidence regions. In addition, we provide an upper bound that is similar to (3.5) for goal functions with unbounded support.

- As compared to previous work by Bartl and Tangpi [2022] where a similar bound as in (3.5) with a q-dependence is obtained for linear goal functions, our bound on (3.3) requires less strict assumptions, achieves a tighter convergence rate with a smaller q, and is derived using different proof techniques, based on an empirical process "peeling device" and McDiarmid's concentration inequality McDiarmid [1989].

- In particular, we show that the conditions for our main theorems are significantly simplified for the class of distortion risk measures and expectiles, providing explicit upper bounds on (3.3) that are easily obtainable for these risk measures. Moreover, our estimation methodology can also be adapted to provide an upper bound on (3.3) for rank-dependent utility risk measures, where the distortion function has an inverse S-shape.

In the following subsection, we review some of the existing literature.

3.1.1 Related literature

Sample average approximation for risk-neutral stochastic programs has been well studied in the literature, where a general survey can be found in Kim et al. [2015]. Consistency and

asymptotic normality for both the optimal values and solutions are established in works such as Shapiro [1993, 2013], Shapiro et al. [2009]. Non-asymptotic lower and upper confidence bounds on the deviation probability (3.3) for risk-neutral stochastic programs have been provided by Guigues et al. [2017], under the assumption that the goal function G is convex in the decision variable, and the additional sub-Gaussian conditions that are imposed on the distributions of $G(\theta, Z_1) - \mathbb{E}_{\mathbb{P}}[G(\theta, Z_1)]$. On the other hand, Krätschmer [2024a] was able to provide non-asymptotic upper estimates on (3.3) for more general goal functions G, but not for a lower estimate.

The most related works on non-asymptotic upper bounds for SAA of risk-averse stochastic programs are Bartl and Tangpi [2022] and Krätschmer [2024a]. Bartl and Tangpi [2022] provides non-asymptotic convergence rates for SAA of stochastic programs with general convex risk measures, both for the deviation probability and the expected deviation. However, their bounds do not have explicit constants, and the goal function G in the stochastic program is restricted to linear cases, and is uniformly bounded in $\theta \in \Theta$ when estimating the deviation probability. By contrast, the upper bounds provided by Krätschmer [2024a] do allow for non-linear, unbounded goal functions and have explicit constants, but are only available for stochastic programs where the risk measures are expected loss, upper semi-deviations, and divergence risk measures. Our work significantly extends the results of these two papers by establishing non-asymptotic upper bounds for SAA of stochastic programs with (i) general risk functionals encompassing convex risk measures and rank-dependent utility functionals with inverse S-shaped probability weighting functions, where (ii) the goal functions may be non-linear and unbounded, and (iii) the bounds feature explicit constants. The explicitness of the constants enables the construction of confidence regions. In addition, when the goal functions G are uniformly bounded, our non-asymptotic upper bound yields convergence rates that are tighter than Bartl and Tangpi [2022].

As an alternative to SAA, recent work by Chu and Tangpi [2024] also investigated the approximation of risk-averse stochastic programs using the stochastic gradient Langevin dynamics method. Under the assumption that the goal function is bounded, Lipschitz continuous, with a bounded gradient, they obtained explicit non-asymptotic upper bounds on the mean squared approximation error that decay with a rate of $1/n$ to the remaining positive constants that depend on the choice of the parameters. As compared to our work, we focus on the deviation probability (3.3), providing a bound that decays faster to zero with n than any moment bounds through Markov's inequality. However, we do note that their bound is independent of the dimension of the decision space Θ.

Asymptotic approximations of SAA for risk-averse stochastic programs have also been studied by Krätschmer [2024b], in the case of expected loss, upper semi-deviations, and divergence risk measures. Central limit theorems for risk-averse stochastic programs have also been established by Chen et al. [2024] for risk measures that are composite functionals,

and by Guigues et al. [2018] for risk measures that have a discrete Kusuoka Kusuoka [2001] representation. A central limit theorem for law-invariant coherent risk measures in the non-optimization setting has been studied by Belomestny and Krätschmer [2012] and Pflug and Wozabal [2010].

3.1.2 Outline

This paper is organized as follows. Section 3.2 summarizes the main assumptions that are made in this work and introduces the necessary preliminaries. Section 3.3 presents the main result (3.5) in a formal setting. Moreover, we also discuss the main assumptions imposed on the risk measures that are made throughout this paper and show that they can be simplified for some canonical examples of convex risk measures, such as the distortion risk measures with concave distortion functions. Section 3.4 shows the application of our main result in Section 3.3 to the estimation of the deviation probability for distortion risk measures with an inverse S-shaped distortion function, which is a major class of non-convex risk measures that are extensively used in behavioral economics (see, e.g., Wakker 2010). Section 3.5 extends the main result, namely Theorem 3.10, to unbounded goal functions. Finally, the proofs of the main results are provided in Section 3.7.

3.2 Set up

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a fixed atomless complete probability space, and let $L^p(\Omega, \mathcal{F}, \mathbb{P})$ denote the usual L^p-space on $(\Omega, \mathcal{F}, \mathbb{P})$ ($p \in [1, \infty]$), where we tacitly identify random variables that are different on $\mathbb{P}$-null sets only. Furthermore, $\mathcal{X}$ stands for some $\mathbb{R}$-vector space of $\mathbb{P}$-integrable random variables, enclosing all $\mathbb{P}$-essentially bounded random variables. The vector space $\mathcal{X}$ will be equipped with a norm $\|\cdot\|_{\mathcal{X}}$ and the $\mathbb{P}$-a.s. order $\succeq_{\mathbb{P}}$ such that $(\mathcal{X}, \|\cdot\|_{\mathcal{X}}, \succeq_{\mathbb{Q}})$ is a Banach lattice which is solid, i.e., $X \in \mathcal{X}$ if $|Y| \geq |X|$ for some $Y \in \mathcal{X}$, and rearrangement invariant, meaning that a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ belongs to $\mathcal{X}$ whenever it has the same distribution as some $Y \in \mathcal{X}$. In addition, we assume

$$\lim_{k\to\infty} \|X - X \wedge k\|_{\mathcal{X}} = 0, \quad \text{for nonnegative } X \in \mathcal{X}. \tag{3.6}$$

Prominent examples are provided by $L^p(\Omega, \mathcal{F}, \mathbb{P})$ with the usual L^p-norm for $p \in [1, \infty]$.

Next, let us fix any law-invariant convex risk measure, $\rho : \mathcal{X} \to \mathbb{R}$, which by definition is a convex mapping satisfying, in addition, the following properties.

- **monotonicity:**

$$\rho(X) \leq \rho(Y) \quad \text{for } X, Y \in \mathcal{X} \text{ with } X \leq Y \ \mathbb{P}-\text{a.s.},$$

- **cash-invariance:**

$$\rho(X + c) = \rho(X) + c \quad \text{for } X \in \mathcal{X}, c \in \mathbb{R},$$

- **law-invariance:**

$$\rho(X) = \rho(Y) \quad \text{for indentically distributed } X, Y \in \mathcal{X}.$$

If in addition ρ is positively homogeneous (meaning $\rho(\lambda X) = \lambda\rho(X)$ for any $\lambda \geq 0$), then ρ is called a law-invariant coherent risk measure. Throughout this paper, we also impose the following additional property for ρ.

- **Lebesgue property**

$$\lim_{k\to\infty} \rho(X_k) = \rho(X)$$

for any sequence $(X_k)_{k\in\mathbb{N}}$ in $\mathcal{X}$ which is uniformly bounded w.r.t. the $\mathbb{P}$-essential sup-norm and converges $\mathbb{P}$-a.s. to $X \in \mathcal{X}$.

The Lebesgue property is always fulfilled if $\mathcal{X}$ contains at least one random variable which is not $\mathbb{P}$-essentially bounded (see Delbaen 2009, Theorem 3).

The outstanding example of a law-invariant convex risk measure with the Lebesgue property is provided by the Average Value at Risk, which we recall now.

Example 3.1. Let $\alpha \in (0, 1)$. Then the mapping $\rho = AV@R_\alpha : L^1(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$, defined by

$$AV@R_\alpha(X) = \frac{1}{1-\alpha} \int_{[\alpha,1)} F_X^{\leftarrow}(u)\, du,$$

where $F_X^{\leftarrow}$ denotes the left-continuous quantile function of the distribution function F_X, is called the *Average Value at Risk* w.r.t. α (e.g., Föllmer and Schied 2016, Shapiro et al. 2009). It has the following useful representation

$$AV@R_\alpha(X) = \inf_{x\in\mathbb{R}} \mathbb{E}\left[\frac{(X-x)^+}{1-\alpha} + x\right], \quad \text{for } X \in L^1(\Omega, \mathcal{F}, \mathbb{P}) \tag{3.7}$$

(see e.g., Kaina and Rüschendorf 2009). In particular, $AV@R_\alpha$ may be identified as a law-invariant coherent risk measure.

The Average Value at Risk is the building block of all law-invariant convex risk measures that satisfy the Lebesgue property due to the *Kusuoka representation*, given by

$$\rho(X) = \sup_{\mu\in\mathrm{dom}(\beta_\rho)} \left(\int_{[0,1)} AV@R_\alpha(X)\ \mu(d\alpha) - \beta_\rho(\mu)\right), \tag{3.8}$$

where $\mathrm{dom}(\beta_\rho)$ denotes the effective domain of the function β_ρ (see Proposition 3.A.1 in Appendix 3.A) defined as

$$\beta_\rho : \mathcal{M}\big([0,1)\big) \to \mathbb{R} \cup \{\infty\},\ \mu \mapsto \sup_{\substack{X \in L^\infty(\Omega,\mathcal{F},\mathbb{P}) \\ \rho(X) \le 0}} \int_{[0,1)} AV@R_\alpha(X)\ \mu(d\alpha), \tag{3.9}$$

where $\mathcal{M}\big([0,1)\big)$ is the set of all Borel probability measures on $[0,1)$, and $AV@R_0$ stands for the expectation.

The Kusuoka representation is a powerful tool for deriving an upper bound for the deviation probability (3.3), as it allows us to decompose ρ into smaller "pieces" of $AV@R_\alpha$ that are easier to analyze. Indeed, the SAA of $AV@R_\alpha$ is itself a standard stochastic optimization problem involving the expected value due to the representation (3.7). Convenient ways to find upper bounds of the deviation probabilities (3.3) for expected values may be provided by general devices from empirical process theory, which are based on covering numbers for classes of Borel measurable mappings from $\mathbb{R}^d$ into $\mathbb{R}$ w.r.t. L^p-norms. Therefore, we restrict ourselves to goal functions G that satisfy the following properties, which are conventional for the applications of empirical process theory in stochastic programming.

(A 1) $G(\theta,\cdot)$ is Borel measurable for every $\theta \in \Theta$.

(A 2) There exists a strictly positive Borel measurable mapping $\xi : \mathbb{R}^d \to \mathbb{R}$ such that $\xi(Z_1) \in \mathcal{X}$ and

$$\sup_{\theta\in\Theta} |G(\theta,z)| \le \xi(z), \quad \text{for } z \in \mathbb{R}^d.$$

(A 3) There exists an at most countable subset $\bar{\Theta} \subset \Theta$ and $(\mathbb{P}^z)^n$-null sets N_n such that

$$\inf_{\theta\in\bar{\Theta}} \mathbb{E}\left[|G(\theta,Z_1) - G(\tilde{\theta},Z_1)|\right] = \inf_{\theta\in\bar{\Theta}} \max_{j\in\{1,\dots,n\}} |G(\theta,z_j) - G(\tilde{\theta},z_j)| = 0,$$

for $n \in \mathbb{N}$, $\tilde{\theta} \in \Theta$, and $(z_1,\dots,z_n) \in \mathbb{R}^{dn} \setminus N_n$.

(A 1) and (A 2) are standard conditions imposed on function classes. (A 3) is used for establishing measurability for (3.3) (see e.g., Lemma 3.23).

The complexity of a function class is measured by its uniform entropy integral and covering numbers. We recall these concepts adapted to our situation. Let us fix any nonvoid function class $\mathbb{F}$ of Borel measurable mappings from $\mathbb{R}^d$ into $\mathbb{R}$ and any probability measure $\mathbb{Q}$ on $\mathcal{B}(\mathbb{R}^d)$ with metric $d_{\mathbb{Q},p}$ induced by the L^p-norm $\|\cdot\|_{\mathbb{Q},p}$ for $p \in [1,\infty)$.

- *Covering numbers for* $\mathbb{F}$
  We use $N\big(\eta,\mathbb{F},L^p(\mathbb{Q})\big)$ to denote the minimal number to cover $\mathbb{F}$ by closed $d_{\mathbb{Q},p}$-balls of radius $\eta > 0$ with centers in $\mathbb{F}$. We define $N\big(\eta,\mathbb{F},L^p(\mathbb{Q})\big) \triangleq \infty$ if no finite cover is available.

- An *envelope* of $\mathbb{F}$ is defined as some Borel measurable mapping $C_{\mathbb{F}}$ from $\mathbb{R}^d$ into $\mathbb{R}$ satisfying $\sup_{h\in\mathbb{F}}|h| \leq C_{\mathbb{F}}$. If an envelope $C_{\mathbb{F}}$ has strictly positive outcomes, we shall speak of a *positive envelope*.
- $\mathcal{M}_{\text{fin}}$ denotes the set of all probability measures on $\mathcal{B}(\mathbb{R}^d)$ with finite support.
- The uniform entropy integral for a class $\mathbb{F}$ of Borel measurable functions from $\mathbb{R}^d$ into $\mathbb{R}$ with an arbitrary positive envelope $C_{\mathbb{F}}$ of $\mathbb{F}$ is denoted as:

$$J(\mathbb{F}, C_{\mathbb{F}}, \delta) \triangleq \int_0^{\delta} \sup_{\mathbb{Q}\in\mathcal{M}_{\text{fin}}} \sqrt{\log\left(2N\big(\varepsilon \, \|C_{\mathbb{F}}\|_{\mathbb{Q},2}, \mathbb{F}, L^2(\mathbb{Q})\big)\right)}\, d\varepsilon, \tag{3.10}$$

  for any $\delta > 0$.[1]

For our purposes, the following function class is the most relevant one, which we denote by:

$$\mathbb{F}^{\Theta} \triangleq \{G(\theta, \cdot) \mid \theta \in \Theta\}. \tag{3.11}$$

The derivation of our upper bound on the deviation probability (3.3) requires a finite uniform entropy integral for the function class (3.11). This is satisfied if the covering number in the integrand has polynomial rate, i.e., there exist $K \geq e$ and $v \geq 1$, such that

$$N\big(\varepsilon \, \|C_{\mathbb{F}}\|_{\mathbb{Q},2}, \mathbb{F}^{\Theta}, L^2(\mathbb{Q})\big) \leq (K/\epsilon)^v$$

holds for $\mathbb{Q} \in \mathcal{M}_{\text{fin}}$ and $\epsilon \in (0,1)$. Then, through a change of variable and integration by parts, the uniform entropy integral can be upper bounded as follows:

$$\int_0^1 \sqrt{v \log\left(K/\epsilon\right)} d\epsilon \leq 2\sqrt{v \log(K)}.$$

Examples of function classes with explicit upper bounds on $J(\mathbb{F}, C_{\mathbb{F}}, \delta)$ are provided in Section 2 of Krätschmer [2024a]. Here, we briefly summarize two main examples from Krätschmer [2024a].

Example 3.2 (Hölder continuous functions)**.** If the goal function $G(\theta, z)$ satisfies the Hölder condition for $\theta \in \mathbb{R}^m, z \in \mathbb{R}^d$, namely there exists an $\tilde{\alpha} \in (0,1]$, and a square $\mathbb{P}^Z$-integrable strictly positive mapping $C : \mathbb{R}^d \to (0,\infty)$, such that for $z \in \mathbb{R}^d, \theta_1, \theta_2 \in \Theta$:

$$|G(\theta_1, z) - G(\theta_2, z)| \leq C(z)\|\theta_1 - \theta_2\|_2^{\tilde{\alpha}},$$

then, under some mild conditions given by Proposition 2.6 of Krätschmer [2024a], there

[1] Throughout this paper, $y \mapsto \log(y)$ is defined as the inverse function of $x \mapsto e^x$.

exists an envelope function ξ for $\mathbb{F}^\Theta$, such that for $\delta \in (0, 1/2]$,

$$J(\mathbb{F}^\Theta, \xi, \delta) \leq 2\delta \sqrt{(3m+1)\ln(2) + \frac{m}{\tilde{\alpha}} \ln(2/\delta)}.$$

Example 3.3 (Two stage mixed-integer stochastic programs)**.** In two stage mixed-integer stochastic programs, the goal functions can be typically represented by a member of the class of piecewise-linear Hölder continuous functions, where G has the following structure (see Krätschmer 2024a and Eichhorn and Römisch 2007, p.121):

$$G(\theta, z) = \sum_{i=1}^{k} \min_{l=1,\ldots,s_i} \mathbb{1}_{I_{il}}(\Lambda_{i,l}(\theta, z) + a_l^i) \cdot G^i(\theta, z),$$

where

- $k, s_1, \ldots, s_k \in \mathbb{N}$,
- G^i satisfies (A 1) and is Hölder continuous with $\tilde{\alpha}_i \in (0, 1]$ and square integrable strictly positive mapping C_i for $i \in \{1, \ldots, k\}$,
- Λ_{il} are Borel measurable and affine linear for $z \in \mathbb{R}^d$, and $i \in \{1, \ldots, k\}$, $l \in \{1, \ldots, s_i\}$,
- $a_l^i \in \mathbb{R}$ for $i \in \{1, \ldots, k\}$ and $l \in \{1, \ldots, s_i\}$,
- $I_{il} = (0, \infty)$ or $[0, \infty)$ for $i \in \{1, \ldots, k\}$ and $l \in \{1, \ldots, s_i\}$,
- The set

$$\left\{ \bigcap_{l=1}^{s_i} \{\Lambda_{il}(\theta, \cdot) + a_l^i \in I_{il}\} \mid i \in \{1, \ldots, k\},\ l \in \{1, \ldots, s_i\} \right\}$$

  forms a partition of $\mathbb{R}^d$.

By Proposition 2.8 of Krätschmer [2024a], if G has the above piecewise-linear structure, then there exists an envelope function ξ for $\mathbb{F}^\Theta$, such that for $\delta \in (0, 1]$,

$$J(\mathbb{F}^\Theta, \xi, \delta)$$
$$\leq 2\delta \sqrt{k + 2km\log(3) + m\log(4k/\delta) \sum_{i=1}^{k} 1/\beta_i + \log(2) + (5 + 2\log(4k/\delta))(m+2) \sum_{i=1}^{k} s_i}.$$

3.3 Deviation probabilities

Throughout this section, we restrict ourselves to law-invariant convex risk measures ρ which are normalized, meaning $\rho(0) = 0$, and goal functions $G(\theta, z)$ that are uniformly

bounded in θ, i.e., (A 2) is satisfied with $\xi \equiv B$ for some $0 < B < \infty$. We first present our main result for coherent risk measures, for which an upper bound on the deviation probability (3.3) can be established under quite explicit conditions imposed on ρ. Extension to the case of general convex risk measures will be discussed in Subsection 3.3.3.

3.3.1 The Coherent Case

Let ρ be a coherent risk measure. To derive an upper bound on the deviation probability, we impose the following condition on ρ:

(A 4) There exists some $q \in [1, \infty)$ such that

$$\overline{N}_q \triangleq \sup_{t \in (0,1]} \frac{\overline{h}_\rho(t)}{t^{1/q}} < \infty,$$

where we introduce the auxiliary mapping

$$\overline{h}_\rho : [0,1] \to \mathbb{R},\ t \mapsto \rho\big(\mathbb{1}_{([1-t],1)}(U)\big),$$

where U denotes any random variable which is uniformly distributed on $(0,1)$, and therefore $\mathbb{1}_{([1-t],1)}(U)$ is a Bernoulli random variable that takes the value 1 with probability t. Under (A 4), we may state the following theorem.

Theorem 3.4. *Let ρ be a law-invariant coherent risk measure which satisfies the Lebesgue property, and let assumptions (A 1), (A 2), (A 3) and (A 4) be fulfilled with $q \in [1, \infty)$, $\xi \equiv B \in \mathbb{R}$, and constant $\overline{N}_q \in \mathbb{R}$. If $J(\mathbb{F}^\Theta, B, 1/8)$ is finite, then,*

$$\mathbb{P}\left(\left\{\left|\inf_{\theta\in\Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta\in\Theta} \mathcal{R}_\rho(F_\theta)\right| \geq \varepsilon\right\}\right) \leq \mathbb{P}^*\left(\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| \geq \epsilon\right)$$
$$\leq \exp\left(-\frac{n\ t^2\ \varepsilon^{2q}}{2^{4q+1}\ (t+1)^2\ B^{2q}\ \overline{N}_q^{2q}}\right)$$

holds for $t > 0$ if

$$\varepsilon > \overline{N}_q \cdot B\ [2^{2q-1}\ 128\ (t+1)\ \sqrt{2}]^{1/q}\ \big[J\big(\mathbb{F}^\Theta, B, 1/8\big) + 3/2\big]^{1/q} / n^{1/(2q)}.$$

Here $\mathbb{P}^*$ denotes the outer probability w.r.t. $\mathbb{P}$.

The proof may be found in Section 3.7.1, where Theorem 3.4 is proved together with Theorem 3.10, which is stated under the more general assumption of ρ being a convex risk measure.

Remark 3.5. The mapping $\inf_{\theta\in\Theta}\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta\in\Theta}\mathcal{R}_\rho(F_\theta)$ is measurable on $(\Omega,\mathcal{F},\mathbb{P})$, and thus the deviation probability does not need to be replaced by an outer measure $\mathbb{P}^*$. This is stated in Lemma 3.23.

Remark 3.6. The upper bound in Theorem 3.4 depends on the parameter q through ϵ^{2q}. As we will discuss in Subsection 3.3.4, the exponent of ϵ determines the convergence rate of $\inf_{\theta\in\Theta}\mathcal{R}_\rho(\widehat{F}_{n,\theta})$. In particular, we provide a lower bound on the deviation probability in Theorem 3.15, which shows that a dependence on q is necessary for the upper bound in Theorem 3.4.

The idea behind condition (A 4) is that it measures how much the coherent risk measure ρ overweights the right-tail probability of a loss distribution. This is best illustrated by the class of distortion risk measures, where the auxiliary function $\overline{h}_\rho$ simplifies to the distortion function h, as we elaborate in the next example.

Example 3.7. A concave distortion function is a concave non-decreasing mapping $h : [0,1] \to [0,1]$ satisfying $h(0) = 0$ and $h(1) = 1$. Every such mapping h induces via

$$\rho_h(X) \triangleq \int_{-\infty}^{0} \big[h(\mathbb{P}(X > x)) - 1\big]\, dx + \int_0^\infty h(\mathbb{P}(X > x))\, dx$$

a coherent risk measure on $\mathcal{X}$ consisting of all random variables X on $(\Omega,\mathcal{F},\mathbb{P})$ satisfying

$$\|X\|_h \triangleq \int_0^\infty h(\mathbb{P}(|X| > x))\, dx < \infty \tag{3.12}$$

(see e.g. Denneberg 1994 or Föllmer and Schied 2016). This is because ρ_h may be viewed as a Choquet integral w.r.t. the submodular set function $h(\mathbb{P})$.

The set $\mathcal{X}$ is a vector space, and the mapping $\|\cdot\|_h$, defined by (3.12), is a norm on $\mathcal{X}$ (see Denneberg 1994, Proposition 9.4). Moreover, if h is continuous, then $\mathcal{X}$, endowed with $\|\cdot\|_h$ and the $\mathbb{P}$-a.s. order $\succeq_{\mathbb{P}}$ is a solid, rearrangement invariant Banach lattice meeting property (3.6) (see Denneberg 1994, Proposition 9.5 with Theorem 8.9). Continuity of h also implies the Lebesgue property (see Denneberg 1994, Theorem 8.9).

For any Bernoulli variable Y that takes the value 1 with probability p, we have that $\rho_h(Y) = h(p)$. Therefore, the auxiliary function $\overline{h}_{\rho_h}$ coincides with the distortion function h. Hence, in the case of distortion risk measures, condition (A 4) is equivalent to the existence of a $q \geq 1$, such that

$$\sup_{t\in(0,1]} \frac{h(t)}{t^{1/q}} < \infty.$$

In Table 3.1, we provide examples of distortion functions with upper bounds for the terms $\overline{N}_q$ and the range $rg(q)$ of q such that (A 4) holds.

Distortion Family	$h(p),\ p \in [0,1]$	$\overline{N}_q$	rg(q)
Proportional Hazard	$p^a,\ a \in (0,1)$	1	$q \geq 1/a$
Gini Principles	$(1+a)p - ap^2,\ a \in (0,1)$	1 + a	$q \geq 1$
Dual Moments	$1-(1-p)^k,\ k > 1$	k	$q \geq 1$
MAXMINVAR	$(1-(1-p)^n)^{1/k},\ k > 1$	$k^{1/k}$	$q \geq k$
LB-transform	$p^a(1-\log(p^a)),\ a \in (0,1)$	$\exp\left(\frac{-a+1/q}{a(aq-1)}\right)\frac{aq}{aq-1}$	$q > 1/a$

Table 3.1: Examples of distortion functions with upper bounds on $\overline{N}_q$ and the range $rg(q)$ of q such that (A 4) holds. For Proportional Hazard, see Wang [1995]; for Gini Principles, see Denneberg [1990b]; for Dual Moments, see Muliere and Scarsini [1989] and Eeckhoudt and Laeven [2021], Eeckhoudt et al. [2020]; for MAXMINVAR, see Cherny and Madan [2009]; for LB-Transform, see Denneberg [1990a].

Moreover, we note that for any coherent risk measures ρ with a Kusuoka representation (3.8), there exists a set $\mathcal{H}_\rho$ of distortion functions, such that $\rho(X) = \sup_{h\in\mathcal{H}_\rho} \rho_h(X)$ (see Section 4.6, Föllmer and Schied 2016). Therefore, condition (A 4) is also equivalent to the existence of a $q \geq 1$, such that $\sup_{t\in(0,1]} \sup_{h\in\mathcal{H}_\rho} h(t)/t^{1/q} < \infty$.

Next, we shall focus on expectiles, genuinely introduced in the paper Newey and Powell [1987], for which the condition (A 4) can also be verified explicitly.

Example 3.8. For $\alpha \in (0,1)$, the mapping

$$\rho^\alpha : L^2(\Omega,\mathcal{F},\mathbb{P}) \to \mathbb{R},\ X \mapsto \operatorname*{argmin}_{x\in\mathbb{R}} \left[\alpha\|(X-x)^+\|_2^2 + (1-\alpha)\|(X-x)^-\|_2^2\right],$$

is well-defined and known as the *expectile w.r.t.* α. It has been shown in Bellini et al. [2014] that it is a law-invariant coherent risk measure for any $\alpha \in [1/2,1)$. Fixing $\alpha \in [1/2,1)$, the associated auxiliary function $\overline{h}_{\rho^\alpha}$ of ρ^α satisfies

$$\overline{h}_{\rho^\alpha}(t) = \frac{\alpha t}{1-\alpha+t(2\alpha-1)}.$$

In particular, since $\alpha \geq 1/2$,

$$\overline{N}_q \triangleq \sup_{t\in(0,1]} \frac{\overline{h}_{\rho^\alpha}(t)}{t^{1/q}} \leq \sup_{t\in(0,1]} \frac{\alpha}{1-\alpha+t(2\alpha-1)} = \frac{\alpha}{1-\alpha}, \quad \text{for every } q \in [1,\infty).$$

3.3.2 Alternative Bound for Coherent Risk Measures

Until now, we have derived upper bounds on the deviation probability (3.3) under assumption (A 4) for coherent risk measures, namely that the auxiliary function $\overline{h}_\rho(t)$ must approach zero at a rate of $t^{1/q}$ for some $q \geq 1$. For coherent risk measures that do not

satisfy (A 4), we can still derive an upper bound on the deviation probability by modifying the estimation technique in the proof of Theorem 3.10. This is stated in the following theorem.

Theorem 3.9. *Let ρ be a normalized law-invariant coherent risk measure that satisfies the Lebesgue property, and assumptions (A 1), (A 2), (A 3) be fulfilled with $\xi \equiv B \in \mathbb{R}$ and finite $J(\mathbb{F}^\Theta, B, 1/8)$. Then $K_\varepsilon \triangleq \inf\left\{K \in \mathbb{N} \mid 4B\ \overline{h}_\rho(2^{-K+1}) < \varepsilon\right\}$ is well-defined for any $\epsilon > 0$, and*

$$\mathbb{P}\left(\left\{\left|\inf_{\theta\in\Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta\in\Theta} \mathcal{R}_\rho(F_\theta)\right| \geq \varepsilon\right\}\right) \leq \mathbb{P}^*\left(\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| \geq \epsilon\right)$$
$$\leq \frac{1}{2304\ t^2\ \ln(4)}\ \exp\left(-\frac{n\ t^2\ \varepsilon^2}{8\ 4^{K_\varepsilon}\ (t+1)^2\ B^2}\right)$$

holds for $t > 0$ and

$$\varepsilon > 2^{K_\varepsilon - 1}\ 128\ (t+1)\ \sqrt{2}\ B\big[J\big(\mathbb{F}^\Theta, B, 1/8\big) + 3/2\big]/\sqrt{n}.$$

Here $\mathbb{P}^*$ denotes the outer probability w.r.t. $\mathbb{P}$.

The proof may be found in Section 3.7.3. We note that Theorem 3.9 holds under very general conditions, namely that K_ϵ always exists due to the Lebesgue property. As a tradeoff, the bound can be conservative for small ϵ, since it depends exponentially on K_ϵ. An application of Theorem 3.9 will be given in Section 3.4, when we estimate the deviation probability for inverse S-shaped distortion risk measures.

3.3.3 Extension to General Convex Risk Measures

If ρ is a general convex risk measure that is not coherent, then (A 4) is extended to the following condition imposed on the Kusuoka representation (3.8):

(A 5) There exists some $q \in [1, \infty)$ such that

$$N_{q,b} \triangleq \sup_{t\in(0,1]} \sup_{\substack{\mu\in\mathcal{M}([0,1)) \\ \beta_\rho(\mu)\leq b}} \frac{\int_{[0,1)} \frac{t\wedge(1-\alpha)}{1-\alpha}\ \mu(d\alpha)}{t^{1/q}} < \infty \quad \text{for } b > 0.$$

We note that for coherent risk measures, (A 5) is equivalent to (A 4), since the function β_ρ vanishes on its effective domain due to positive homogeneity of ρ. Under (A 5), we obtain the following theorem concerning the deviation probability (3.3), which directly implies Theorem 3.4 when ρ is coherent (by substituting $N_{q,4B+\delta}$ with $\overline{N}_q$ in the following statement).

Theorem 3.10. *Let ρ be a normalized law-invariant convex risk measure which satisfies the Lebesgue property, and let assumptions (A 1), (A 2), (A 3) and (A 5) be fulfilled with $q \in [1,\infty)$, $\xi \equiv B \in \mathbb{R}$, and constants $N_{q,b} \in \mathbb{R}$ $(b>0)$ as in (A 5). If $J(\mathbb{F}^\Theta, B, 1/8)$ is finite, then*

$$\mathbb{P}\left(\left\{\left|\inf_{\theta\in\Theta}\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta\in\Theta}\mathcal{R}_\rho(F_\theta)\right| \geq \varepsilon\right\}\right) \leq \mathbb{P}^*\left(\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| \geq \epsilon\right)$$
$$\leq \exp\left(-\frac{n\ t^2\ \varepsilon^{2q}}{2^{4q+1}\ (t+1)^2\ B^{2q}\ N_{q,(4B+\delta)}^{2q}}\right)$$

holds for $t,\delta>0$ whenever

$$\varepsilon > N_{q,(4B+\delta)} \cdot B\ [2^{2q-1}\ 128\ (t+1)\ \sqrt{2}]^{1/q}\ \left[J\big(\mathbb{F}^\Theta, B, 1/8\big) + 3/2\right]^{1/q} / n^{1/(2q)}.$$

Here $\mathbb{P}^*$ denotes the outer probability w.r.t. $\mathbb{P}$.

The proof may be found in Section 3.7.1.

Remark 3.11. Theorem 3.10 allows us to construct a confidence region for the unknown optimal value $\inf_{\theta\in\Theta}\mathcal{R}_\rho(F_\theta)$. Indeed, given a confidence parameter β, for any ϵ that satisfies the condition of Theorem 3.10, if the sample size $n \geq \frac{\log(1/\beta)}{Ct^2\epsilon^{2q}}$ for $C = 1/(2^{4q+1}(t+1)^2 B^{2q} N_{q,4B+\delta}^{2q})$, then the interval $\inf_{\theta\in\Theta}\mathcal{R}_\rho(\hat{F}_{n,\theta}) \pm \epsilon$ contains $\inf_{\theta\in\Theta}\mathcal{R}_\rho(F_\theta)$ with probability $1-\beta$. Moreover, we note that $\inf_{\theta\in\Theta}\mathcal{R}_\rho(\hat{F}_{n,\theta})$ is a convex optimization problem for many risk measures (see, e.g., Ben-Tal and Teboulle 2007, Ben-Tal et al. 2013 and Jin et al. 2025b).

Although assumption (A 5) is less explicit than (A 4) in the coherent case, we can still provide some sufficient conditions on (A 5) that are more assessable. This is discussed in the following remark.

Remark 3.12. Let $q \in [1,\infty)$. Suppose that representation (3.8) satisfies

$$\overline{M}_q \triangleq \sup_{\mu\in\mathcal{M}([0,1))}\left(\int_{[0,1)}\frac{1}{(1-\alpha)^{1/q}}\ \mu(d\alpha) - \beta_\rho(\mu)\right) < \infty. \tag{3.13}$$

Then, for any μ from the effective domain of β_ρ and every $t \in [0,1]$,

$$\int_{[0,1)}\frac{t\wedge(1-\alpha)}{1-\alpha}\ \mu(d\alpha) \leq t^{1/q}\ \int_{[0,1)}\frac{1}{(1-\alpha)^{1/q}}\ \mu(d\alpha) = t^{1/q}\ (\overline{M}_q + \beta_\rho(\mu)).$$

Hence, (A 5) is satisfied with $N_{q,b} = \overline{M}_q + b$ for $b>0$.

A criterion to ensure (3.13) for $q \in (1,\infty)$ is provided by the property of q-regularity, which was introduced in Bartl and Tangpi [2022]. By definition, ρ is called q-regular with $q \in (1,\infty)$ if for any random variable X_q on $(\Omega,\mathcal{F},\mathbb{P})$ which is Pareto-distributed with

location parameter 1 and scale parameter q,

$$\widehat{M}_q \triangleq \sup_{k\in\mathbb{N}} \rho(X_q \wedge k) < \infty.$$

In terms of the representation (3.8), the application of the monotone convergence theorem yields

$$\begin{aligned}\widehat{M}_q &= \sup_{\mu\in\mathcal{M}([0,1))} \left(\int_{[0,1)} AV@R_\alpha(X_q)\ \mu(d\alpha) - \beta_\rho(\mu) \right) \\ &= \sup_{\mu\in\mathcal{M}([0,1))} \left(\int_{[0,1)} \frac{q}{q-1}\ \frac{1}{(1-\alpha)^{1/q}}\ \mu(d\alpha) - \beta_\rho(\mu) \right) \\ &= \frac{q}{q-1} \sup_{\mu\in\mathcal{M}([0,1))} \left(\int_{[0,1)} \frac{1}{(1-\alpha)^{1/q}}\ \mu(d\alpha) - \frac{q-1}{q}\ \beta_\rho(\mu) \right).\end{aligned}$$

Finally, $0 = \rho(0) = \sup_{\mu\in\mathcal{M}([0,1))}[-\beta_\rho(\mu)]$ and therefore β_ρ is nonnegative. Hence (3.13) is fulfilled with $\overline{M}_q \leq (q-1)\ \widehat{M}_q/q$.

As an application of Remark 3.12, we shall examine utility-based shortfall risk measures.

Example 3.13. Let $l : \mathbb{R} \to \mathbb{R}$ denote a non-constant, non-decreasing convex function, and x_0 be an element from the topological interior of the range of l. Furthermore, it is assumed that there are $N_{(l)} > 0$ and $p \in [1, \infty)$ such that $l(x) \leq N_{(l)}\ (x^+)^p$ holds for $x \in \mathbb{R}$. Then with $c_0 \triangleq \inf\{c \in \mathbb{R} \mid l(-c) \leq x_0\}$,

$$\rho_{l,x_0} : L^p(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R},\ X \mapsto \inf\left\{x \in \mathbb{R} \mid \mathbb{E}[l(X-x)] \leq x_0\right\} - c_0$$

defines a normalized law-invariant convex risk measure which satisfies the Lebesgue property. As a convex risk measure on a Banach lattice, it is also continuous w.r.t. the L^p-norm (see Ruszczynski and Shapiro 2006, Proposition 3.1).

For any $q \in (p, \infty)$, every random variable X_q on $(\Omega, \mathcal{F}, \mathbb{P})$ which is Pareto-distributed with location parameter 1 and scale parameter q belongs to $L^p(\Omega, \mathcal{F}, \mathbb{P})$. In particular, $\sup_{k\in\mathbb{N}} \rho_{l,x_0}(X_q \wedge k) = \rho_{l,x_0}(X_q)$ for $q \in (p, \infty)$ due to L^p-norm continuity of ρ_{l,x_0}. According to Remark 3.12, condition (A 5) is fulfilled for every $q \in (p, \infty)$, where the corresponding constants $\overline{M}_q$ satisfy the following upper estimation

$$\overline{M}_q \leq \frac{q-1}{q}\ \rho_{l,x_0}(X_q) \leq \frac{q-1}{q}\ \inf\left\{x \in \mathbb{R} \mid \mathbb{E}\big[\big((X_q - x)^+\big)^p\big] \leq N_{(l)}\ x_0\right\} - \frac{q-1}{q}\ c_0.$$

3.3.4 Sharpness of Convergence Rates

In this subsection, we discuss the convergence rate of the sequence

$$\left(\inf_{\theta\in\Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta\in\Theta} \mathcal{R}_\rho(F_\theta)\right)_{n\in\mathbb{N}} \tag{3.14}$$

that can be obtained from the bounds given in Theorems 3.4 and 3.10. An immediate consequence is that we may provide the following simple criterion for the tightness rates of the sequence (3.14).

Theorem 3.14. *Let ρ be a law-invariant convex risk measure which satisfies the Lebesgue property, and let assumptions (A 1), (A 2), (A 3) be fulfilled with $\xi \equiv B \in \mathbb{R}$ and $J(\mathbb{F}^\Theta, B, 1/8) < \infty$. If ρ satisfies (A 5) $q \in [1,\infty)$, or if ρ is coherent and meets (A 4) with the same q, then the sequence*

$$\left(n^{1/(2q)} \left[\inf_{\theta\in\Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta\in\Theta} \mathcal{R}_\rho(F_\theta)\right]\right)_{n\in\mathbb{N}}$$

is uniformly tight.

Proof. Let $q \in [1,\infty)$ such that ρ satisfies (A 5) with q, or (A 4) with q if ρ is coherent. Then, applying Theorem 3.10 or Theorem 3.4 both with $t = 1$, there is some $\widehat{N}_q > 0$ such that

$$\mathbb{P}\left(\left\{\left|\inf_{\theta\in\Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta\in\Theta} \mathcal{R}_\rho(F_\theta)\right| \geq \frac{\varepsilon}{n^{1/(2q)}}\right\}\right) \leq \exp\left(-\frac{\varepsilon^{2q}}{2^{4q+3}\ 4\ B^{2q}\ \widehat{N}_q^{2q}}\right)$$

holds for every $\varepsilon > \widehat{N}_q \cdot B\ [2^{2q-1}\ 256\ \sqrt{2}]^{1/q}\ \left[J\big(\mathbb{F}^\Theta, B, 1/8\big) + 3/2\right]^{1/q}$ and any $n \in \mathbb{N}$. In particular,

$$\lim_{\varepsilon\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\left\{n^{1/(2q)} \left|\inf_{\theta\in\Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta\in\Theta} \mathcal{R}_\rho(F_\theta)\right| \geq \varepsilon\right\}\right) = 0,$$

which completes the proof. □

Theorem 3.14 provides a means of characterizing the convergence rate of the sequence (3.14), in the sense that (3.14) is bounded by $\epsilon/n^{1/(2q)}$ with high probability, uniformly over all distributions, if ϵ is a sufficiently large constant. A similar conclusion on the $n^{-1/(2q)}$ convergence rate is established in Theorem 5.1 of Bartl and Tangpi [2022], which only holds for linear goal functions and $q > 1$. We emphasize that compared to Bartl and Tangpi [2022], our Theorem 3.10 provides a strict improvement for the convergence rate. Indeed, besides the fact that Theorem 3.10 holds for non-linear goal functions, it also includes cases where the best possible $n^{-1/2}$ rate can be achieved. This is true for risk measures such as the Average Value at Risk, the Expectiles from Example 3.8, and

the Dual Moments distortion risk measures as in Table 3.1, where Theorem 3.14 holds for $q = 1$. The rate improvement is not only for risk measures that satisfy (A 5) for $q = 1$. For example, the MAXMINVAR distortion risk measure of order $k > 1$ in Table 3.1 satisfies the condition of Theorem 5.1 in Bartl and Tangpi [2022] only for $q > k$, whereas Table 3.1 shows that Theorem 3.14 also holds for MAXMINVAR with $q = k$. As we mentioned earlier in Remark 3.12, (A 5) is a more general condition than the "q-regularity" imposed by Bartl and Tangpi [2022].

It is also important to investigate whether the factor ϵ^{2q} in the bounds of Theorem 3.4 and 3.10 can be replaced by ϵ^2, which would then yield the usual $1/\sqrt{n}$ convergence rate for $q > 1$. In the following theorem, we state that it is impossible to achieve this independence of q, meaning our bounds provide convergence rates that are tight up to a factor of two, uniformly over all distributions.

Theorem 3.15. *For any $q \geq 1$, there exists a coherent risk measure ρ, such that $\overline{N}_q < \infty$ as in (A 4), and*

$$\sup_{\mathbb{Q}} \mathbb{Q}\left(\left|\mathcal{R}_\rho(\widehat{F}_n) - \mathcal{R}_\rho(F)\right| \geq \epsilon\right) \geq \frac{1}{\sqrt{2^{q+3}}} \exp\left(-\tilde{C}_q n \epsilon^q (1 + O(\epsilon^2))\right),$$

for all $n \geq 3$ and $1/(2n^{1/q}) \leq \epsilon \leq (1 - 1/n)^{1/q}/2 \wedge 1/4$, where the supremum is taken over all probability measure $\mathbb{Q}$ on $(\Omega, \mathcal{F})$, $\hat{F}_n, F$ are the empirical and true distribution functions under $\mathbb{Q}$, and

$$\tilde{C}_q \triangleq 2^{q+1}(q+1)\log(2) - \frac{2^q - 1}{1 - 1/2^q} + \frac{1}{2}.$$

The proof is provided in Section 3.7.2.

3.4 Distortion Risk Measures with Inverse S-Shaped Distortion Function

In this section, we examine distortion risk measures with an inverse S-shaped distortion function, which are widely used in behavioral economics, as they capture risk preferences that are more consistent with empirical findings in decision-making (see Wakker 2010). A distortion function h is inverse S-shaped, if there exists a $p_0 \in (0, 1)$ such that h is concave on $[0, p_0]$, but convex on $[p_0, 1]$.

To provide an upper bound on the deviation probability (3.3) for distortion risk measures with an inverse S-shaped distortion function, we will represent ρ as the difference of two distortion risk measures with concave distortion functions. Each of these is a convex risk measure to which Theorem 3.4 can be applied individually. This is achieved by considering the dual function $\bar{h}(p) \triangleq 1 - h(1-p)$ that is associated with a distortion function h.

Given an inverse S-shaped distortion function h, we define $h_0(p) \triangleq \min\{h(p), h(p_0)\}$ and $\bar{h}_0(p) \triangleq \min\{\bar{h}(p), 1 - h(p_0)\}$, for $p \in [0,1]$. Note that both h_0 and $\bar{h}_0$ are non-decreasing concave functions on $[0,1]$, and we denote their normalized versions as $h_{cc} \triangleq h_0/h(p_0)$ and $h_{cv} \triangleq \bar{h}_0/(1-h(p_0))$. The following lemma states the representation of a distortion risk measure with inverse-S-shaped distortion function in terms of h_{cc} and h_{cv}.

Lemma 3.16. *Let h be an inverse S-shaped distortion function. Then, we have that*

$$\rho_h(X) = h(p_0)\rho_{h_{cc}}(X) - (1 - h(p_0))\rho_{h_{cv}}(-X).$$

Therefore, we see that upper bounds on the deviation probability for distortion risk measures with inverse S-shaped distortion functions can be provided by bounding their concave and convex parts given by h_{cc} and h_{cv}. We state this in the following theorem, where $\overline{N}_{q,cc}, \overline{N}_{q,cv}$ denote the constants in (A 4) for h_{cc} and h_{cv} (recall that by Example 3.7, $\overline{h}_{\rho_h}(t) = h(t)$).

Theorem 3.17. *Let ρ_h be a distortion risk measure with an inverse S-shaped distortion function. Let (A 1), (A 2) with $\xi \equiv B \in \mathbb{R}$, (A 3), (A 4) be fulfilled with $q_1, q_2 \in [1,\infty)$ such that $\overline{N}_{q_1,cc}, \overline{N}_{q_2,cv} < \infty$. Suppose that $J(\mathbb{F}^\Theta, B, 1/8)$ is finite. Then,*

$$\begin{aligned}
&\mathbb{P}^*\Big(\Big\{\sup_{\theta\in\Theta}\big|\mathcal{R}_{\rho_h}\big(\widehat{F}_{n,\theta}\big) - \mathcal{R}_{\rho_h}\big(F_\theta\big)\big| > \epsilon\Big\}\Big) \\
&\leq \exp\left(-\frac{n\ t^2\ \varepsilon^{2q_1}}{2^{6q_1+1}\ (t+1)^2\ B^{2q_1}\ h(p_0)^{2q_1}\overline{N}_{q_1,cc}^{2q_1}}\right) \\
&\quad + \exp\left(-\frac{n\ t^2\ \varepsilon^{2q_2}}{2^{6q_2+1}\ (t+1)^2\ B^{2q_2}\ (1-h(p_0))^{2q_2}\overline{N}_{q_2,cv}^{2q_2}}\right)
\end{aligned}$$

holds for $t > 0$ if

$$\begin{aligned}
\varepsilon > \max_{(q,c)\in\{(q_1,cc),(q_2,cv)\}} &\overline{N}_{q,c} \cdot B\ [2^{2q-1}\ 128\ (t+1)\ \sqrt{2}]^{1/q}\ \big[J\big(\mathbb{F}^\Theta, B, 1/8\big) + 3/2\big]^{1/q}/n^{1/(2q)} \\
&\cdot \max\{h(p_0), 1-h(p_0)\}.
\end{aligned}$$

Here, $\mathbb{P}^*$ denotes the outer probability w.r.t. $\mathbb{P}$.

Example 3.18. Let h_1, h_2 be two distortion functions that satisfy (A 4) with $q_1, q_2 \geq 1$ (such as the examples in Table 3.1). Let $p_0 \in (0,1)$. Then, the function

$$h^S(p) = \begin{cases} h_1(p) & p \leq p_0 \\ 1 - \frac{1-h_1(p_0)}{h_2(1-p_0)} h_2(1-p) & p \geq p_0, \end{cases}$$

is an inverse S-shaped distortion function, such that the corresponding h^S_{cc}, h^S_{cv} satisfies $\overline{N}_{q_1,cc}, \overline{N}_{q_2,cv} < \infty$.

Example 3.19. Consider the Kahneman-Tversky function (see Tversky and Kahneman 1992), which is defined as

$$h(p) = 1 - \frac{(1-p)^\beta}{((1-p)^\beta + p^\beta)^{1/\beta}},\ 0 < \beta < 1.$$

One can show that the Kahneman-Tversky function satisfies $\overline{N}_{q,cc}, \overline{N}_{q,cv} < \infty$, for $q = 1/\beta$. The details are provided in Section 3.8.2.

However, there are inverse S-shaped distortion functions for which their convex part h_{cv} does not satisfy $\overline{N}_{q,cv} < \infty$ for any $q \geq 1$. One prominent example of this is Prelec's distortion function (see Prelec 1998), defined as

$$h(p) = 1 - \exp\left(-(-\log(1-p))^\alpha\right),\ 0 < \alpha < 1.$$

Indeed, for Prelec's distortion function, we have that for $p \leq 1/e$,

$$h_{cv}(p) = \exp\left(-(-\log(p))^\alpha\right) \cdot \frac{1}{1 - h(1/e)},\ 0 < \alpha < 1.$$

However, since for any $q \geq 1$,

$$\lim_{p \downarrow 0} \frac{\exp\left(-(-\log(p))^\alpha\right)}{p^{1/q}} = \lim_{p \downarrow 0} \exp\left(\frac{1}{q} \log\left(\frac{1}{p}\right)\left(1 - \log^{\alpha-1}\left(\frac{1}{p}\right)\right)\right) = \infty,$$

we have that $\overline{N}_{q,cv} = \infty$ for any $q \geq 1$. Hence, Theorem 3.17 cannot be applied to Prelec's function. Instead, one should use Theorem 3.9 to estimate the deviation probability for the convex part h_{cv} of Prelec's function.

3.5 Extension to Goal Functions with Unbounded Support

The previous sections concern only uniformly bounded goal functions. In this section, we provide an upper bound on the deviation probability (3.3) when the goal functions G have unbounded support. This requires several technical adjustments to the proof of Theorem 3.10, namely

1. Instead of a uniform bound B as the envelope function, we assume that ξ is integrable up to some order $r \geq 2$ (i.e., $\|\xi\|_{\mathbb{P}^Z,r} < \infty$).

2. The deviation quantity $\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)$ is decomposed into a part with bounded goal functions for the application of Theorem 3.10, and a remaining part that requires further estimation.

3. If ρ is a convex risk measure that is not coherent, then (A 5) has to be modified, since one can no longer use $4B + \delta$ as an upper bound for the penalty function in the Kusuoka representation.

Given an envelope function ξ with $\|\xi\|_{\mathbb{P}^Z,r} < \infty$, the decomposition step of the aforementioned technical adjustment is carried out by introducing an auxiliary function $\phi_{r,n} : \mathbb{R} \to \mathbb{R}$ defined by $\phi_{r,n}(x) \triangleq (x \wedge w_{r,n}) \vee (-w_{r,n})$, where for $n \in \mathbb{N}$, we set $w_{r,n} \triangleq (2n)^{1/r} \, \|\xi\|_{\mathbb{P}^Z,r}$. Note that the image of $\phi_{r,n}$ is always contained in the interval $[-w_{r,n}, w_{r,n}]$. The purpose of $\phi_{r,n}$ is solely to restrict the objective function G on a compact support, as we will consider the class of functions $\phi_{r,n}(G(\theta, Z))$ for $\theta \in \Theta$. We let $F_\theta^{\phi_{r,n}}$ to denote the distribution function of $\phi_{r,n}(G(\theta, Z))$, and $\widehat{F}_\theta^{\phi_{r,n}}$ the corresponding empirical distribution function based on $Z_1, \ldots, Z_n$.

Finally, besides the auxiliary function $\phi_{r,n}$, we also introduce the auxiliary events,

$$B_{n,r}^{\xi} \triangleq \Big\{ \frac{1}{n} \sum_{j=1}^{n} \xi(Z_j)^r \leq 2\mathbb{E}[\xi(Z_1)^r] \Big\} \quad (n \in \mathbb{N}). \tag{3.15}$$

After introducing all the necessary notations, we can state our main theorems on the deviation probability (3.3), when G has unbounded support. Similar to Section 3.3, we separate the coherent case and the general convex case.

If ρ is coherent, then we may provide the following upper bound on (3.3) under assumption (A 4).

Theorem 3.20. *Let ρ be a normalized law-invariant coherent risk measure which satisfies the Lebesgue property, and let assumptions (A 1), (A 2), (A 3) and (A 4) be fulfilled for some $q \in [1, \infty)$. Suppose the envelope function ξ as in (A 2) satisfies $\|\xi\|_{\mathbb{P}^Z,r} < \infty$ for some $r \geq q \vee 2$. If $J(\mathbb{F}^\Theta, \xi, 1/4) < \infty$, then for any $t, \delta > 0$,*

$$\begin{aligned} &\mathbb{P}\left(\left| \inf_{\theta \in \Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta \in \Theta} \mathcal{R}_\rho(F_\theta) \right| \geq \epsilon \right) \\ &\leq \exp\left(- \frac{n^{1-2q/r} \; t^2 \; \varepsilon^{2q}}{2^{4q+3} \; (t+1)^2 \overline{N}_q^{2q} \|\xi\|_{\mathbb{P}^Z,r}^{2q}} \right) + \mathbb{P}\big(\Omega \setminus B_{n,r}^{\xi}\big), \end{aligned}$$

whenever $t, \delta > 0$ satisfies

$$\begin{aligned} \epsilon \geq \; & \frac{(256(t+1)\|\xi\|_{\mathbb{P}^Z,r}^q [J\big(\mathbb{F}^\Theta, \xi, 1/8\big) + 3/2])^{1/q} 8\overline{N}_q}{n^{1/(2q)-1/r}} \\ & + \frac{4\overline{N}_q \|\xi\|_{\mathbb{P}^Z,r} q/(r-q)}{n^{1/q-1/r}} \end{aligned}$$

The proof is presented in Section 3.7.6.

Let us discuss the further estimation of the probability $\mathbb{P}(\Omega \setminus B_{n,r}^{\xi})$, which is similar to

the case of $r = 2$ as discussed in Remark 3.2 of Krätschmer [2024a], using the substitution $\tilde{\xi} = \xi^{r/2}$.

Remark 3.21. We have the following estimations on $\mathbb{P}(\Omega \setminus B^{\xi}_{n,r})$.

(1) If $\xi(Z_1)$ is integrable of order $2r$, then Cantelli's inequality implies

$$\mathbb{P}(\Omega \setminus B^{\xi}_{n,r}) \leq \frac{\mathrm{Var}[\xi^r(Z_1)]}{n\mathbb{E}[\xi^r(Z_1)]^2 + \mathrm{Var}[\xi^r(Z_1)]}.$$

(2) If $\mathbb{E}[\exp(\lambda \xi^r(Z_1))] < \infty$ for some $\lambda > 0$, then by Remark 3.2 of Krätschmer [2024a], we have that

$$\mathbb{P}(\Omega \setminus B^{\xi}_{n,r}) \leq \exp\left(-n\mathbb{E}[\xi^r(Z_1)]^2/(8\delta_0^2)\right) \vee \exp\left(-n\mathbb{E}[\xi^r(Z_1)]/(4\delta_0)\right),$$

for $n \in \mathbb{N}$ and

$$\delta_0 \geq \sup_{k \in \mathbb{N}, k \geq 2} \left(\left|\mathbb{E}\left[(\xi^r(Z_1) - \mathbb{E}[\xi^r(Z_1)])^k\right]\right| / k!\right)^{1/k}.$$

If ρ is a convex risk measure that is not coherent, then we impose the following condition on the Kusuoka representation (3.8), which, by Remark 3.12, also provides a sufficient condition on (A 5).

(A 6) There exists $r, q \in [1, \infty)$ with $r \geq q \vee 2$, such that $\|\xi\|_{\mathbb{P}^Z, r} < \infty$ and

$$\overline{M}^{\xi}_{r,q} \triangleq \sup_{\mu \in \mathrm{dom}(\beta_\rho)} \left(\int_{[0,1)} \frac{(4\ \|\xi\|_{\mathbb{P}^Z,r}) \vee 1}{(1-\alpha)^{1/q}}\ \mu(d\alpha) - \beta_\rho(\mu) \right) < \infty.$$

We note that similarly as in Remark 3.12, an upper bound on $\overline{M}^{\xi}_{r,q}$ can also be provided by the q-regularity condition introduced in Bartl and Tangpi [2022], namely one has

$$\overline{M}^{\xi}_{r,q} \leq \frac{q-1}{q}\ \sup_{k \in \mathbb{N}} \rho(X^*_{[4\|\xi\|_{\mathbb{P}^Z,r}] \vee 1, q} \wedge k),$$

where $X^*_{\lambda,q}$ is the Pareto distribution with scale parameter $\lambda > 0$ and shape parameter q.

Theorem 3.22. *Let ρ be a normalized law-invariant convex risk measure which satisfies the Lebesgue property, and let assumptions (A 1), (A 2), (A 3) and (A 6) be fulfilled for some $q \in [1, \infty)$ and $r \geq q \vee 2$. Denote $b_{r,\delta} \triangleq 2\overline{M}^{\xi}_{r,q} + \delta$. Then, the mapping*

$$\inf_{\theta \in \Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta \in \Theta} \mathcal{R}_\rho(F_\theta)$$

is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. *Moreover, if* $J(\mathbb{F}^\Theta, \xi, 1/4) < \infty$, *then for any* $t, \delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\inf_{\theta\in\Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta\in\Theta} \mathcal{R}_\rho(F_\theta)\right| \geq \epsilon\right) \\
&\leq \mathbb{P}^*\left(\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| \geq \epsilon\right) \\
&\leq \exp\left(-\frac{n^{1-2q/r}\ t^2\ \varepsilon^{2q}}{2^{6q+1}\ (t+1)^2\ N_{q,b_{r,\delta}}^{2q}\ \|\xi\|_{\mathbb{P}_Z,r}^{2q}}\right) + \mathbb{P}\big(\Omega \setminus B_{n,r}^\xi\big)
\end{aligned}$$

whenever $t, \delta > 0$ *and*

$$\begin{aligned}
\varepsilon > &\frac{\big((256(t+1)\|\xi\|_{\mathbb{P}_Z,r}^q[J\big(\mathbb{F}^\Theta, \xi, 1/8\big) + 3/2])^{1/q} 8 N_{q,b_{r,\delta}}}{n^{1/(2q)-1/r}} \\
&+ \frac{4[3\overline{M}_{r,q}^\xi + \delta]\ \|\xi\|_{\mathbb{P}^Z,r}}{n^{1/q-1/r}}.
\end{aligned}$$

The proof is presented in Section 3.7.6.

Finally, let us comment on the bounds in Theorems 3.20 and 3.22. Note that if G is uniformly bounded in θ, then we may take $\xi \equiv B < \infty$, which means $r = \infty$ and $\mathbb{P}(\Omega \setminus B_{n,r}^\xi) = 0$. Then, the rate in Theorems 3.20 and 3.22 recovers back to $n^{-1/(2q)}$, which is the same as in Theorem 3.10 for the bounded case. Moreover, combining the estimation of $\mathbb{P}(\Omega \setminus B_{n,r}^\xi)$ in Remark 3.21, the bound in Theorems 3.20 and 3.22 yields a smaller confidence region for the true optimal value $\inf_{\theta\in\Theta} \mathcal{R}_\rho(F_\theta)$ than the moment bound provided in Bartl and Tangpi [2022], which through Markov's inequality decays in the order of $1/n^{1/(q(2-1/r))-1/(r-2)}$, which is always slower than $1/\sqrt{n}$.

3.6 Conclusion and Discussion

By utilizing the Kusuoka representation of law-invariant convex risk measures and tools from empirical process theory, we obtain non-asymptotic upper estimations on the deviation probability (3.3) under more general conditions imposed on the goal functions and risk measures, as compared to Bartl and Tangpi [2022] and Krätschmer [2024a]. As a direction of future research, it would be valuable to investigate further if the convergence rate $n^{-1/(2q)}$ established in Theorem 3.10 can be improved to $n^{-1/q}$.

3.7 Proofs

Lemma 3.23. *Let assumptions (A 1) - (A 3) be fulfilled. If* ρ *satisfies the Lebesgue property, then the mapping*

$$\inf_{\theta\in\Theta} \mathcal{R}_\rho(F_\theta) - \inf_{\theta\in\Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta})$$

is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ *for every* $n \in \mathbb{N}$*. If in addition* G *is lower semicontinuous in* θ*, then the mapping*

$$\sup_{\theta\in\Theta} \big|\mathcal{R}_\rho(F_\theta) - \mathcal{R}_\rho(\widehat{F}_{n,\theta})\big|$$

is also a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$*.*

Proof. First of all, $\{\mathcal{R}_\rho(F_\theta) \mid \theta \in \Theta\}$ is bounded from below due to (A 2). Fix any $n \in \mathbb{N}$. Let ξ be from (A 2), and let $\overline{\Theta}$ be some at most countable subset of Θ as in assumption (A 3). The left-continuous quantile function of a distribution function F will be denoted by $F^{\leftarrow}$. We may select some random variable U on the atomless probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which is uniformly distributed on $(0, 1)$.

By assumption (A 3), there is some $A_n \in \mathcal{F}$ with $\mathbb{P}(A_n) = 1$ such that

$$\inf_{\vartheta\in\overline{\Theta}} \max_{j\in\{1,\dots,n\}} \big\{\big|G\big(\theta, Z_j(\omega)\big) - G\big(\vartheta, Z_j(\omega)\big)\big|\big\} = 0, \quad \text{for } \omega \in A_n, \theta \in \Theta.$$

Let $\omega \in A_n$ and $\theta \in \Theta$ be arbitrary. We may find some sequence $(\vartheta_k)_{k\in\mathbb{N}}$ in $\overline{\Theta}$ such that $\widehat{F}_{n,\vartheta_k|\omega}(x) \to \widehat{F}_{n,\theta|\omega}(x)$ at every continuity point x of $\widehat{F}_{n,\theta|\omega}$. Then $\widehat{F}^{\leftarrow}_{n,\vartheta_k|\omega}(\alpha) \to \widehat{F}^{\leftarrow}_{n,\theta|\omega}(\alpha)$ at every continuity point α of $\widehat{F}^{\leftarrow}_{n,\theta|\omega}(\alpha)$ (see e.g., van der Vaart 1998, Lemma 21.2). In particular, $\widehat{F}^{\leftarrow}_{n,\vartheta_k|\omega}(U) \to \widehat{F}^{\leftarrow}_{n,\theta|\omega}(U)$ $\mathbb{P}$-a.s.. Since in addition the sequence $\big(\widehat{F}^{\leftarrow}_{n,\vartheta_k|\omega}(U)\big)_{k\in\mathbb{N}}$ is uniformly bounded by $\max_{j\in\{1,\dots,n\}} \xi\big(Z_j(\omega)\big)$, Lebesgue property of ρ implies

$$\rho\big(\widehat{F}^{\leftarrow}_{n,\vartheta_k|\omega}(U)\big) \to \rho\big(\widehat{F}^{\leftarrow}_{n,\theta|\omega}(U)\big). \tag{3.16}$$

Finally, note that $\widehat{F}^{\leftarrow}_{n,\vartheta|\omega}(U)$ is distributed according to the empirical distribution based on $G\big(\vartheta, Z_1(\omega)\big), \dots, G\big(\vartheta, Z_n(\omega)\big)$ for $\vartheta \in \Theta$ and ω. Thus, we have shown

$$\inf_{\vartheta\in\overline{\Theta}} \mathcal{R}_\rho(\widehat{F}_{n,\vartheta|\omega}) = \inf_{\theta\in\Theta} \mathcal{R}_\rho(\widehat{F}_{n,\theta|\omega}), \quad \text{for } \omega \in A_n.$$

Since $\overline{\Theta}$ is at most countable, and $(\Omega, \mathcal{F}, \mathbb{P})$ is complete, it remains to show that $\mathcal{R}_\rho(\widehat{F}_{n,\theta})$ is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ for every $\theta \in \Theta$.

For this purpose, let $F^{\leftarrow}_{(x_1,\dots,x_n)}$ denote the left-continuous quantile function of the empirical distribution function based on $(x_1, \dots x_n) \in \mathbb{R}^n$. In the same way as we derived convergence (3.16), we may use the Lebesgue property again to verify the continuity of the mapping,

$$\varphi : \mathbb{R}^n \to \mathbb{R}, \ (x_1, \dots, x_n) \mapsto \rho\big(F^{\leftarrow}_{(x_1,\dots,x_n)}(U)\big).$$

Thus $\mathcal{R}_\rho(\widehat{F}_{n,\theta}) = \varphi \circ \big(G(\theta, Z_1), \dots, G(\theta, Z_n)\big)$ is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ for every $\theta \in \Theta$.

For the remaining part of the proof, let us assume that $G(\cdot, z)$ is lower semicontinuous for every $z \in \mathbb{R}^d$. Note that $G(\cdot, z)$ is bounded due to (A 2), and by (A 3) we may find

some $A \in \mathcal{F}$ with $\mathbb{P}(A) = 1$ such that

$$\inf_{\theta\in\Theta} \big(G\big(\theta, Z_1(\omega)\big) - \psi(\theta)\big) = \inf_{\theta\in\overline{\Theta}} \big(G\big(\theta, Z_1(\omega)\big) - \psi(\theta)\big)\,, \quad \text{for } \omega \in A,$$

for any continuous mapping $\psi : \Theta \to \mathbb{R}$. This implies that

$$\inf_{\theta\in\Theta} \big(G\big(\theta, Z_1\big) - \psi(\theta)\big)$$

is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ for any continuous mapping $\psi : \Theta \to \mathbb{R}$, because $\overline{\Theta}$ is countable and $(\Omega, \mathcal{F}, \mathbb{P})$ is complete. Since Θ is compactly metrizable, we may draw on Theorem 2.1 and Remark 1.1 both from Levin [1994] to conclude that G is measurable w.r.t. the product σ-algebra $\mathcal{B}(\Theta) \otimes \mathcal{B}(\mathbb{R}^d)$ of the Borel σ-algebra $\mathcal{B}(\Theta)$ on Θ and the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ on $\mathbb{R}^d$. As a further consequence, $\varphi \circ \big(G(\cdot, Z_1), \ldots, G(\cdot, Z_n)\big)$ is also $\mathcal{B}(\Theta) \otimes \mathcal{B}(\mathbb{R}^d)$ measurable. Moreover, in view of Corollary 3.A.3 in Appendix 3.A along with assumption (A 2), the mapping

$$\Theta \to \mathbb{R},\ \theta \mapsto \rho\big(G(\theta, Z_1)\big) \tag{3.17}$$

is lower semicontinuous. Therefore, the mapping

$$\Theta \times \Omega \to \mathbb{R},\ (\theta, \omega) \mapsto \varphi \circ \Big(G\big(\theta, Z_1(\omega)\big), \ldots, G\big(\theta, Z_n(\omega)\big)\Big) - \rho\big(G(\theta, Z_1)\big)$$

is $\mathcal{B}(\Theta) \otimes \mathcal{B}(\mathbb{R}^d)$ measurable, and it is also bounded in θ for every $\omega \in \Omega$ due to (A 2) along with the monotonicity of ρ. Hence, completeness of $(\Omega, \mathcal{F}, \mathbb{P})$ implies that

$$\sup_{\theta\in\Theta} \Big|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\Big| = \sup_{\theta\in\Theta} \big|\varphi \circ \big(G(\theta, Z_1), \ldots, G(\theta, Z_n)\big) - \rho\big(G(\theta, Z_1)\big)\big|$$

is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ (see van der Vaart and Wellner 2023, Example 1.7.5). This completes the proof. □

In the following proofs, we shall make use of the usual notation from empirical process theory

$$(\mathbb{P}_n - \mathbb{P})(f) \triangleq \frac{1}{n} \sum_{j=1}^{n} \Big(f(Z_j) - \mathbb{E}[f(Z_j)]\Big) \tag{3.18}$$

for $\mathbb{P}^Z$-integrable mappings $f : \mathbb{R}^d \to \mathbb{R}$.

3.7.1 Proof of Theorem 3.10

For any $\epsilon > 0$, we have the following inequality:

$$\mathbb{P}\left(\left\{\left|\inf_{\theta\in\Theta}\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \inf_{\theta\in\Theta}\mathcal{R}_\rho(F_\theta)\right| \geq \varepsilon\right\}\right) \leq \mathbb{P}^*\left(\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| \geq \epsilon\right).$$

Recall that G is uniformly bounded by B. Hence, for every, $\theta \in \Theta$,

$$\begin{aligned}
\mathcal{R}_\alpha(F_\theta) &= \inf_{|x|\leq B} \mathbb{E}\big[\big(G(\theta, Z_1) - x\big)^+ + (1-\alpha)x\big]/(1-\alpha),\\
\mathcal{R}_\alpha(\widehat{F}_{n,\theta}) &= \inf_{|x|\leq B} \frac{1}{n}\sum_{j=1}^{n}\Big[\Big(G\big(\theta, Z_j\big) - x\Big)^+ + (1-\alpha)x\Big]/(1-\alpha).
\end{aligned}$$

Define the function class $\mathbb{F}_B^\Theta \triangleq \{[G(\theta,\cdot) - x]^+ \mid \theta \in \Theta, x \in [-B, B]\}$, and furthermore $\Delta_n \triangleq \sup_{f\in\mathbb{F}_B^\Theta} |(\mathbb{P}_n - \mathbb{P})(f)|$. We obtain

$$\begin{aligned}
\left|\mathcal{R}_\alpha(F_\theta) - \mathcal{R}_\alpha(\widehat{F}_{n,\theta})\right| &\leq \frac{1}{1-\alpha}\sup_{|x|\leq B}\left|(\mathbb{P}_n - \mathbb{P})\big(\big[G(\theta,\cdot) - x\big]^+\big)\right|\\
&\leq \frac{1}{1-\alpha}\sup_{f\in\mathbb{F}_B^\Theta}|(\mathbb{P}_n - \mathbb{P})(f)| = \frac{\Delta_n}{1-\alpha}, \quad \text{for } \alpha \in [0,1), \theta \in \Theta.
\end{aligned}$$

Let $\mathcal{M}_{B,\delta} \triangleq \{\mu \in \mathcal{M} \mid \beta(\mu) \leq (4B + \delta)\}$, for $\delta > 0$ (note that if ρ is *coherent*, then we may simply replace $\mathcal{M}_{B,\delta}$ with $\{\mu \in \mathcal{M} \mid \beta(\mu) = 0\}$). We obtain from Proposition 3.A.2 in Appendix 3.A, that

$$\begin{aligned}
\mathcal{R}_\alpha(F_\theta) &= \sup_{\mu\in\mathcal{M}_{B,\delta}}\left(\int_{[0,1)}\mathcal{R}_\alpha(F_\theta)\ \mu(d\alpha) - \beta_\rho(\mu)\right),\\
\mathcal{R}_\alpha(\widehat{F}_{n,\theta}) &= \sup_{\mu\in\mathcal{M}_{B,\delta}}\left(\int_{[0,1)}\mathcal{R}_\alpha(\widehat{F}_{n,\theta})\ \mu(d\alpha) - \beta_\rho(\mu)\right).
\end{aligned}$$

Since the class $\mathbb{F}_B^\Theta$ is uniformly bounded by the constant $2B$, we have for any θ and every $\delta > 0$:

$$\begin{aligned}
\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| &\leq \sup_{\mu\in\mathcal{M}_{B,\delta}}\int_{[0,1)}\left|\mathcal{R}_\alpha(F_\theta) - \mathcal{R}_\alpha(\widehat{F}_{n,\theta})\right|\mu(d\alpha)\\
&\leq \sup_{\mu\in\mathcal{M}_{B,\delta}}\int_{[0,1)}\left(\frac{\Delta_n}{1-\alpha}\wedge 2B\right)\mu(d\alpha). \qquad (3.19)
\end{aligned}$$

Next, note that $\Delta_n \leq 2B$ holds pointwise so that in view of (3.19)

$$\left\{\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| > 0\right\} \subseteq \{0 < \Delta_n \leq 2B\} = \bigcup_{K=0}^{\infty}\Omega_{nK}, \qquad (3.20)$$

where $\Omega_{nK} \triangleq \{2^{K-1}\Delta_n < 2B \le 2^K\Delta_n\}$. For $K \in \mathbb{N}$ and $\omega \in \Omega_{nK}$ we obtain

$$\frac{\Delta_n}{1-\alpha} \wedge 2B \begin{cases} = \dfrac{\Delta_n}{1-\alpha} & \text{for } \alpha < 1 - 2^{1-K} \\ \le 2B & \text{for } \alpha \ge 1 - 2^{1-K}. \end{cases}$$

This implies by (3.19),

$$\begin{aligned} &\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta|_\omega}) - \mathcal{R}_\rho(F_\theta)\right| \\ &\le \sup_{\mu\in\mathcal{M}_{B,\delta}} \left(\Delta_n(\omega) \int_{[0,1-2^{1-K})} \frac{1}{1-\alpha}\, \mu(d\alpha) + 2B\, \mu\big([1-2^{1-K},1)\big)\right), \quad \text{for } \theta\in\Theta, \delta > 0. \end{aligned}$$

Furthermore, for any Borel probability measure μ on $[0,1)$ and every $t \in [0,1]$,

$$\int_{[0,1)} \frac{t \wedge (1-\alpha)}{1-\alpha}\, \mu(d\alpha) = \mu\big([1-t,1)\big) + t \int_{[0,1-t)} \frac{1}{1-\alpha}\, \mu(d\alpha).$$

Hence, recalling $\Delta_n(\omega) < 2^{2-K}\ B$ and $2B \le 2^K\Delta_n(\omega)$,

$$\begin{aligned} &\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta|_\omega}) - \mathcal{R}_\rho(F_\theta)\right| \\ &\le \sup_{\mu\in\mathcal{M}_{B,\delta}} \left(\int_{[0,1-2^{1-K})} \frac{\Delta_n(\omega) - 2^{2-K}B}{1-\alpha}\, \mu(d\alpha) + 2B \int_{[0,1)} \frac{2^{1-K}\wedge(1-\alpha)}{1-\alpha}\, \mu(d\alpha)\right) \\ &\le 2B \sup_{\mu\in\mathcal{M}_{B,\delta}} \int_{[0,1)} \frac{2^{1-K}\wedge(1-\alpha)}{1-\alpha}\, \mu(d\alpha) \\ &\le 2^K\Delta_n(\omega) \sup_{\mu\in\mathcal{M}_{B,\delta}} \int_{[0,1)} \frac{2^{1-K}\wedge(1-\alpha)}{1-\alpha}\, \mu(d\alpha), \quad \text{for } \theta\in\Theta, \delta > 0. \end{aligned} \tag{3.21}$$

Moreover, for $\omega \in \Omega$ with $2^{-1}\Delta_n(\omega) < 2B \le \Delta_n(\omega)$, we have $2B \le \Delta_n(\omega)/(1-\alpha)$ for every $\alpha \in [0,1)$. Therefore,

$$\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta|_\omega}) - \mathcal{R}_\rho(F_\theta)\right| \le 2B \le \Delta_n(\omega). \tag{3.22}$$

Now, let $\varepsilon > 0$. Putting together (3.20) with (3.21) and (3.22), we may observe

$$\begin{aligned} &\sup_{\theta\in\Theta} \left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| \\ &\le \mathbb{1}_{\Omega_{n0}} \cdot \Delta_n + \sum_{K=1}^{\infty} \mathbb{1}_{\Omega_{nK}} \cdot 2^K\Delta_n \sup_{\mu\in\mathcal{M}_{B,\delta}} \int_{[0,1)} \frac{2^{1-K}\wedge(1-\alpha)}{1-\alpha}\, \mu(d\alpha), \end{aligned} \tag{3.23}$$

for every $\delta > 0$. Invoking assumption (A 5) with $q \in [1, \infty)$ and the constant $N_{q,b}$ (if ρ is coherent, then use assumption (A 4) with constant $\overline{N}_q$), we have

$$2^K \sup_{\mu \in \mathcal{M}_{B,\delta}} \int_{[0,1)} \frac{2^{1-K} \wedge (1-\alpha)}{1-\alpha} \, \mu(d\alpha) \leq 2^K \; N_{q,(4B+\delta)} \; 2^{(1-K)/q}, \quad \text{for } \delta > 0, K \in \mathbb{N}.$$

This implies

$$\begin{aligned}
&\mathbb{1}_{\Omega_{nK}} \cdot 2^K \Delta_n \sup_{\mu \in \mathcal{M}_{B,\delta}} \int_{[0,1)} \frac{2^{1-K} \wedge (1-\alpha)}{1-\alpha} \, \mu(d\alpha) \\
&\leq \mathbb{1}_{\Omega_{nK}} \cdot 2 \cdot 2^{(K-1)(q-1)/q} \; N_{q,(4B+\delta)} \; \Delta_n \\
&= \mathbb{1}_{\Omega_{nK}} \cdot N_{q,(4B+\delta)} \; 2 \cdot 2^{(K-1)(q-1)/q} \; \Delta_n^{(q-1)/q} \; \Delta_n^{1/q} \\
&\leq \mathbb{1}_{\Omega_{nK}} \cdot N_{q,(4B+\delta)} \; 2 \cdot 2^{(K-1)(q-1)/q} \; 2^{(2-K)(q-1)/q} \; B^{(q-1)/q} \; \Delta_n^{1/q} \\
&\leq \mathbb{1}_{\Omega_{nK}} \cdot N_{q,(4B+\delta)} \cdot 2^{(2q-1)/q} \; B^{(q-1)/q} \; \Delta_n^{1/q} \quad \text{for } \delta > 0, K \in \mathbb{N}.
\end{aligned}$$

Since $N_{q,(4B+\delta)} \geq 1$ for $\delta > 0$, and since $\Delta_n \leq 2B$ we may also observe

$$\begin{aligned}
\mathbb{1}_{\Omega_{n0}} \cdot \Delta_n = \mathbb{1}_{\Omega_{n0}} \cdot \Delta_n^{(q-1)/q} \; \Delta_n^{1/q} &= \mathbb{1}_{\Omega_{nK}} \cdot 2^{(q-1)/q} \; B^{(q-1)/q} \; \Delta_n^{1/q} \\
&\leq \mathbb{1}_{\Omega_{n0}} \cdot N_{q,(4B+\delta)} \cdot 2^{(2q-1)/q} \; B^{(q-1)/q} \; \Delta_n^{1/q}, \quad \text{for } \delta > 0.
\end{aligned}$$

Hence, combining (3.23) with (3.20), we end up with

$$\begin{aligned}
&\mathbb{P}^* \left(\left\{ \sup_{\theta \in \Theta} \left| \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta) \right| \geq \varepsilon \right\} \right) \\
&\leq \mathbb{P}^* \left(\left\{ \sum_{K=0}^{\infty} \mathbb{1}_{\Omega_{nK}} \cdot N_{q,(4B+\delta)} \cdot 2^{(2q-1)/q} \; B^{(q-1)/q} \; \Delta_n^{1/q} \geq \varepsilon \right\} \right) \\
&\leq \mathbb{P}^* \left(\left\{ \; \Delta_n \geq \frac{\varepsilon^q}{N_{q,(4B+\delta)}^q \cdot 2^{(2q-1)} \; B^{(q-1)}} \right\} \right), \quad \text{for } \delta > 0. \qquad (3.24)
\end{aligned}$$

Recall the at most countable subset $\overline{\Theta} \subseteq \Theta$ from assumption (A 3). In view of Lemma 3.24 we may further conclude from (3.24)

$$\begin{aligned}
&\mathbb{P}^* \left(\left\{ \sup_{\theta \in \Theta} \left| \mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta) \right| \geq \varepsilon \right\} \right) \\
&\leq \mathbb{P} \left(\left\{ \sup_{f \in \widehat{\mathbb{F}}_I} \left| (\mathbb{P}_n - \mathbb{P})(f) \right| \geq \frac{\varepsilon^q}{N_{q,(4B+\delta)}^q \cdot 2^{(2q-1)} \; B^{(q-1)}} \right\} \right), \qquad (3.25)
\end{aligned}$$

where $\widehat{F}_I \triangleq \{[G(\theta, \cdot) - x]^+ \mid \theta \in \overline{\Theta}, x \in I\}$ for some countable dense subset $I \subseteq [-B, B]$. We want to apply the concentration inequality in Theorem 3.B.4 from Appendix 3.B to upper estimate the deviation probabilities on the right-hand of the inequality (3.25).

For this purpose, note that the constant function $2B$ is a positive envelope of $\widehat{\mathbb{F}}_I$, and recall from Lemma 3.24

$$J\big(\widehat{\mathbb{F}}_I, 2B, 1/2\big) \leq 4\ J(\mathbb{F}^\Theta, B, 1/8\big) + 6.$$

Since in addition, the function class $\widehat{\mathbb{F}}_I$ is at most countable, we may conclude from Theorem 2.1 in Krätschmer [2024a]

$$\mathbb{E}\left[\sup_{f\in\widehat{\mathbb{F}}_I}\big|(\mathbb{P}_n - \mathbb{P})(f)\big|\right] \leq \frac{32\ \sqrt{2}\ B}{\sqrt{n}}\ J\big(\widehat{\mathbb{F}}_I, 2B, 1/2\big) \leq \frac{128\ \sqrt{2}\ B\big[J\big(\mathbb{F}^\Theta, B, 1/8\big) + 3/2\big]}{\sqrt{n}}.$$

If $t > 0$ and $\varepsilon > N_{q,(4B+\delta)} \cdot B\ [2^{2q-1}\ 128\ (t+1)\ \sqrt{2}]^{1/q}\ \big[J\big(\mathbb{F}^\Theta, B, 1/8\big) + 3/2\big]^{1/q}/n^{1/(2q)}$, then

$$\begin{aligned}
&\frac{\varepsilon^q}{N^q_{q,(4B+\delta)} \cdot 2^{(2q-1)}\ B^{(q-1)}} \\
&= \frac{t\ \varepsilon^q}{(t+1)\ N^q_{q,(4B+\delta)} \cdot 2^{(2q-1)}\ B^{(q-1)}} + \frac{\varepsilon^q}{(t+1)\ N^q_{q,(4B+\delta)} \cdot 2^{(2q-1)}\ B^{(q-1)}} \\
&> \frac{t\ \varepsilon^q}{(t+1)\ N^q_{q,(4B+\delta)} \cdot 2^{(2q-1)}\ B^{(q-1)}} + \mathbb{E}\left[\sup_{f\in\widehat{\mathbb{F}}_I}\big|(\mathbb{P}_n - \mathbb{P})(f)\big|\right], \quad \text{for } \delta > 0.
\end{aligned}$$

Therefore in this case, by Theorem 3.B.4 from Appendix 3.B

$$\begin{aligned}
&\mathbb{P}\left(\left\{\sup_{f\in\widehat{\mathbb{F}}_I}\big|(\mathbb{P}_n - \mathbb{P})(f)\big| \geq \frac{\varepsilon^q}{N^q_{q,(4B+\delta)} \cdot 2^{(2q-1)}\ B^{(q-1)}}\right\}\right) \\
&\leq \exp\left(-\frac{n\ t^2\ \varepsilon^{2q}}{8\ (t+1)^2\ B^2\ N^{2q}_{q,(4B+\delta)} \cdot 4^{(2q-1)}\ B^{2(q-1)}}\right) \\
&\leq \exp\left(-\frac{n\ t^2\ \varepsilon^{2q}}{2^{4q+1}\ (t+1)^2\ B^{2q}\ N^{2q}_{q,(4B+\delta)}}\right), \quad \text{for } \delta > 0.
\end{aligned}$$

This completes the proof due to (3.25). □

3.7.2 Proof of Theorem 3.15

Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli variables with parameter p. Let F_n be the empirical distribution function induced by $X_1, \ldots, X_n$ and F the distribution function of X_1. Let

ρ be a distortion risk measure with $h(t) = t^{1/q}$ for $q \geq 1$. Then, we have that,

$$\begin{aligned}
\mathbb{P}\left(\left|\mathcal{R}_\rho(\hat{F}_n) - \mathcal{R}_\rho(F)\right| \geq \epsilon\right) &= \mathbb{P}\left(\left|\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^{1/q} - p^{1/q}\right| \geq \epsilon\right) \\
&\geq \mathbb{P}\left(\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^{1/q} - p^{1/q} \geq \epsilon\right) \\
&= \mathbb{P}\left(\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^{1/q} \geq \epsilon + p^{1/q}\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq \left(\epsilon + p^{1/q}\right)^q\right) \\
&= \mathbb{P}\left(\sum_{i=1}^n X_i \geq n\left(\epsilon + p^{1/q}\right)^q\right).
\end{aligned}$$

Let $p = \epsilon^q$. Then the above becomes

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq n\left(\epsilon + p^{1/q}\right)^q\right) = \mathbb{P}\left(\sum_{i=1}^n X_i \geq n(2\epsilon)^q\right).$$

We assume $1/(2n^{1/q}) \leq \epsilon \leq (1-1/n)^{1/q}/2 \wedge 1/4$. Then by Lemma 4.7.2 of Ash [1990] (see also Ferrante 2021), we have the following lower bound on the binomial distribution:

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^n X_i \geq n(2\epsilon)^q\right) \\
&\geq \mathbb{P}\left(\sum_{i=1}^n X_i \geq n\lceil(2\epsilon)^q n\rceil/n\right) \\
&\geq \frac{1}{\sqrt{8\lceil(2\epsilon)^q n\rceil(1 - \lceil(2\epsilon)^q n\rceil/n)}} \exp\left(-nD(\lceil(2\epsilon)^q n\rceil/n \mid\mid \epsilon^q)\right) \\
&\geq \frac{1}{\sqrt{2^{q+4}n}} \exp\left(-nD(\lceil(2\epsilon)^q n\rceil/n \mid\mid \epsilon^q)\right), \qquad \text{(note that } \lceil(2\epsilon)^q n\rceil \leq 2(2\epsilon)^q n),
\end{aligned}$$

where

$$\begin{aligned}
D(\lceil(2\epsilon)^q n\rceil/n \mid\mid \epsilon^q) &= \frac{\lceil(2\epsilon)^q n\rceil}{n}\log\left(\frac{\lceil(2\epsilon)^q n\rceil}{n\epsilon^q}\right) - \left(1 - \frac{\lceil(2\epsilon)^q n\rceil}{n}\right)\log\left(\frac{1-\epsilon^q}{1 - \frac{\lceil(2\epsilon)^q n\rceil}{n}}\right) \\
&\leq \frac{2(2\epsilon)^q n}{n}\log\left(\frac{2(2\epsilon)^q n}{n\epsilon^q}\right) - \left(1 - \frac{(2\epsilon)^q n}{n}\right)\log\left(\frac{1-\epsilon^q}{1 - \frac{(2\epsilon)^q n}{n}}\right) \\
&= 2^{q+1}(q+1)\epsilon^q \log(2) - (1-(2\epsilon)^q)\log\left(\frac{1-\epsilon^q}{1-(2\epsilon)^q}\right).
\end{aligned}$$

We note that

$$\begin{aligned}\log\left(\frac{1-\epsilon^q}{1-(2\epsilon)^q}\right) &= \log\left(1+\frac{1-\epsilon^q}{1-(2\epsilon)^q}-1\right)\\ &= \log\left(1+\frac{2^q-1}{1-(2\epsilon)^q}\epsilon^q\right)\\ &\le \log\left(1+\frac{2^q-1}{1-1/2^q}\epsilon^q\right)\\ &= \frac{2^q-1}{1-1/2^q}\epsilon^q + O(\epsilon^{2q}).\end{aligned}$$

Hence,

$$\begin{aligned}D(\lceil(2\epsilon)^q n\rceil/n \mid\mid \epsilon^q) &\le 2^{q+1}(q+1)\epsilon^q\log(2) - (1-(2\epsilon)^q)\left(\frac{2^q-1}{1-1/2^q}\epsilon^q + O(\epsilon^{2q})\right)\\ &= \left(2^{q+1}(q+1)\log(2) - \frac{2^q-1}{1-1/2^q}\right)\epsilon^q + O(\epsilon^{2q}).\end{aligned}$$

Hence, we have that

$$\begin{aligned}&\mathbb{P}\left(\sum_{i=1}^n X_i \ge n(2\epsilon)^q\right)\\ &\ge \frac{1}{\sqrt{2^{q+3}}}\exp\left(-n\left(\left(2^{q+1}(q+1)\log(2) - \frac{2^q-1}{1-1/2^q}\right)\epsilon^q(1+O(\epsilon^2)) + \frac{1}{2}\log(n)/n\right)\right)\\ &\ge \frac{1}{\sqrt{2^{q+3}}}\exp\left(-n\left(2^{q+1}(q+1)\log(2) - \frac{2^q-1}{1-1/2^q} + \frac{1}{2}\right)\epsilon^q(1+O(\epsilon^2))\right),\end{aligned}$$

where we note that $\log(n)/n \le \epsilon^q$ iff $\log(n) \le n\epsilon^q$. Since $n\epsilon^q \le n-1$, we have $\log(n) \le n-1$, which is iff $1 \le (n-1)/\log(n)$, which holds for all $n \ge 3$. □

3.7.3 Proof of Theorem 3.9

We re-examine the proof of Theorem 3.10. Let $\{\Omega_{nK}\}_{K\ge 0}$ be the disjoint partitions introduced in the proof of Theorem 3.10, where $\Omega_{nK} \triangleq \left\{2^{K-1}\Delta_n < 2B \le 2^K\Delta_n\right\}$. Using the fact that ρ is coherent (hence, β_ρ vanishing on its effective domain), we have

$$\begin{aligned}&\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right|\\ &\le \mathbb{1}_{\Omega_{n0}}\cdot\Delta_n + \sum_{K=1}^\infty \mathbb{1}_{\Omega_{nK}}\cdot 2^K\Delta_n \sup_{\mu\in\mathcal{M},\beta_\rho(\mu)=0}\int_{[0,1)}\frac{2^{1-K}\wedge(1-\alpha)}{1-\alpha}\,\mu(d\alpha)\\ &= \mathbb{1}_{\Omega_{n0}}\cdot\Delta_n + \sum_{K=1}^\infty \mathbb{1}_{\Omega_{nK}}\cdot 2^K\Delta_n\,\overline{h}_\rho(2^{1-K}).\end{aligned}$$

For $K \geq 1$, and $\omega \in \Omega_{nK}$, we have that

$$2^K \Delta_n \ \overline{h}_\rho(2^{1-K}) \leq 4B\overline{h}_\rho(2^{1-K}).$$

For a given $\epsilon > 0$, let $K_\epsilon \triangleq \min\{K \in \mathbb{N} \mid 4B\overline{h}_\rho(2^{1-K}) < \epsilon\}$. Then, for $K \geq K_\epsilon$, we have that

$$\sup_{\theta\in\Theta} \left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| \leq 4B\overline{h}_\rho(2^{1-K}) < \epsilon.$$

Hence,

$$\left\{\sup_{\theta\in\Theta} \left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| > \epsilon\right\} \cap \bigcup_{K\geq K_\epsilon} \Omega_{nK} = \emptyset.$$

Therefore, using $\overline{h}_\rho(t) \leq 1, \forall t \in [0,1]$, we have

$$\begin{aligned}
\mathbb{P}^*\left(\sup_{\theta\in\Theta} \left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| > \epsilon\right) &\leq \mathbb{P}^*\left(\mathbb{1}_{\Omega_{n0}} \cdot \Delta_n + \sum_{K=1}^{K_\epsilon-1} \mathbb{1}_{\Omega_{nK}} \cdot 2^K \Delta_n \ \overline{h}_\rho(2^{1-K}) > \epsilon\right) \\
&\leq \mathbb{P}^*\left(\mathbb{1}_{\Omega_{n0}} \cdot \Delta_n + \sum_{K=1}^{K_\epsilon-1} \mathbb{1}_{\Omega_{nK}} \cdot 2^K \Delta_n > \epsilon\right) \\
&\leq \sum_{K=0}^{K_\epsilon-1} \mathbb{P}^*\left(\Delta_n > \frac{\epsilon}{2^K}\right).
\end{aligned}$$

Now, following the same arguments as in the proof of Theorem 3.10, we may use Lemma 3.24 to obtain the further estimation:

$$\begin{aligned}
&\mathbb{P}^*\left(\left\{\sup_{\theta\in\Theta} \left|\mathcal{R}_\rho(\widehat{F}_{n,\theta}) - \mathcal{R}_\rho(F_\theta)\right| \geq \varepsilon\right\}\right) \\
&\leq \sum_{K=0}^{K_\varepsilon-1} \mathbb{P}\left(\left\{\sup_{f\in\widehat{\mathbb{F}}_I} \left|(\mathbb{P}_n - \mathbb{P})(f)\right| \geq \frac{\varepsilon}{2^K}\right\}\right),
\end{aligned} \tag{3.26}$$

where $\widehat{F}_I \triangleq \{[G(\theta,\cdot) - x]^+ \mid \theta \in \overline{\Theta}, x \in I\}$ for some countable dense subset $I \subseteq [-B, B]$. To apply the concentration inequality in Theorem 3.B.4, we note that the constant function $2B$ is a positive envelope of $\widehat{\mathbb{F}}_I$, and recall from Lemma 3.24

$$J\big(\widehat{\mathbb{F}}_I, 2B, 1/2\big) \leq 4 \ J\big(\mathbb{F}^\Theta, B, 1/8\big) + 6.$$

Since in addition, the function class $\widehat{\mathbb{F}}_I$ is at most countable, we may conclude from Theorem 2.1 in Krätschmer [2024a]

$$\mathbb{E}\left[\sup_{f\in\widehat{\mathbb{F}}_I}\left|(\mathbb{P}_n-\mathbb{P})(f)\right|\right]\leq\frac{32\ \sqrt{2}\ B}{\sqrt{n}}\ J\big(\widehat{\mathbb{F}}_I,2B,1/2\big)\leq\frac{128\ \sqrt{2}\ B\big[J\big(\mathbb{F}^{\Theta},B,1/8\big)+3/2\big]}{\sqrt{n}}.$$

If $t>0$ and $\varepsilon>2^{K_\varepsilon-1}\ 128\ (t+1)\ \sqrt{2}\ B\big[J\big(\mathbb{F}^{\Theta},B,1/8\big)+3/2\big]/\sqrt{n}$, then for arbitrary $K\in\{0,\ldots,K_\varepsilon-1\}$,

$$\frac{\varepsilon}{2^K}=\frac{t\ \varepsilon}{(t+1)\ 2^K}+\frac{\varepsilon}{(t+1)\ 2^K}>\frac{t\ \varepsilon}{(t+1)\ 2^K}+\mathbb{E}\left[\sup_{f\in\widehat{\mathbb{F}}_I}\left|(\mathbb{P}_n-\mathbb{P})(f)\right|\right].$$

Therefore in this case, by Theorem 3.B.4 from Appendix 3.B

$$\mathbb{P}\left(\left\{\sup_{f\in\widehat{\mathbb{F}}_I}\left|(\mathbb{P}_n-\mathbb{P})(f)\right|\geq\frac{\varepsilon}{2^K}\right\}\right)\leq\exp\left(-\frac{n\ t^2\ \varepsilon^2}{2\ 4^{K+1}\ (t+1)^2\ B^2}\right),$$

for $K\in\{0,\ldots,K_\varepsilon-1\}$. Therefore,

$$\begin{aligned}\mathbb{P}^*\left(\left\{\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(\widehat{F}_{n,\theta})-\mathcal{R}_\rho(F_\theta)\right|\geq\varepsilon\right\}\right)&\leq\sum_{k=0}^{K_\varepsilon-1}\exp\left(-\frac{n\ t^2\ \varepsilon^2}{2\ 4^{k+1}\ (t+1)^2\ B^2}\right)\\&\leq\sum_{k=0}^{K_\varepsilon-1}\int_k^{k+1}\exp\left(-\frac{n\ t^2\ \varepsilon^2}{4^u\ 8\ (t+1)^2\ B^2}\right)du\\&=\int_0^{K_\epsilon}\exp\left(-\frac{n\ t^2\ \varepsilon^2}{4^u\ 8\ (t+1)^2\ B^2}\right)du.\end{aligned}$$

Invoking the change of variable formula, we may further proceed

$$\begin{aligned}&\int_0^{K_\epsilon}\exp\left(-\frac{n\ t^2\ \varepsilon^2}{4^u\ 8\ (t+1)^2\ B^2}\right)du\\&\leq\int_{4^{-K_\epsilon}}^1\exp\left(-y\ \frac{n\ t^2\ \varepsilon^2}{8\ (t+1)^2\ B^2}\right)\ \frac{1}{y\ \ln(4)}dy\\&\leq\frac{4^{K_\varepsilon}}{\ln(4)}\ \int_{4^{-K_\varepsilon}}^1\exp\left(-y\ \frac{n\ t^2\ \varepsilon^2}{8\ (t+1)^2\ B^2}\right)dy\\&\leq\frac{1}{2304\ t^2\ \ln(4)}\ \exp\left(-\frac{n\ t^2\ \varepsilon^2}{8\ 4^{K_\varepsilon}\ (t+1)^2\ B^2}\right),\end{aligned}$$

where in the last step the assumption $\varepsilon>2^{K_\varepsilon-1}\ 128\ (t+1)\ \sqrt{2}\ B\big[J\big(\mathbb{F}^{\Theta},B,1/8\big)+3/2\big]/\sqrt{n}$ has been taken into account. The proof is now complete.

3.7.4 Auxiliary Lemma for Proofs of Theorems 3.10, 3.9 and 3.22

The following lemma is repeatedly used in proofs of Theorems 3.10, 3.9, and 3.22. In Theorems 3.10 and 3.9, the goal functions are assumed to be uniformly bounded by some $B \in \mathbb{R}$, whereas in Theorem 3.22, goal functions are allowed to have unbounded support. Therefore, to state the following lemma for both the bounded and unbounded cases, we adopt the following rules for interpreting the notations $\phi_{r,n}$ and $w_{r,n}$. Namely, for the proofs of Theorems 3.10 and 3.9, Lemma 3.24 is applied with $w_{r,n} = B$ and $\phi_{r,n}(x) = x$, whereas for the proof of Theorem 3.22, Lemma 3.24 is applied with $w_{r,n} = (2n)^{1/r}\|\xi\|_{\mathbb{P}^Z,r}$ and $\phi_{r,n}(x) = (x \wedge w_{r,n}) \vee (-w_{r,n})$. Moreover, we denote the function class

$$\mathbb{F}_{w_{r,n}}^{\Theta,\phi_{r,n}} \triangleq \Big\{ \big[\phi_{r,n}\big(G(\theta,\cdot)\big) - x\big]^+ \mid \theta \in \Theta, x \in [-w_{r,n}, w_{r,n}] \Big\}. \tag{3.27}$$

Lemma 3.24. *Let (A 1) - (A 3) be fulfilled. Furthermore, let $\overline{\Theta}$ be the at most countable subset of Θ from (A 3). Then for any $n \in \mathbb{N}$, the function class $\mathbb{F}_{w_{r,n}}^{\Theta,\phi_{r,n}}$ satisfies the following properties.*

1) For countable dense subset I of $[-w_{r,n}, w_{r,n}]$.

$$\sup_{f \in \mathbb{F}_{w_{r,n}}^{\Theta,\phi_{r,n}}} \big|(\mathbb{P}_n - \mathbb{P})(f)\big| = \sup_{(\theta,x)\in\overline{\Theta}\times I} \Big|(\mathbb{P}_n - \mathbb{P})\Big(\big(\phi_{r,n}/ \circ G(\theta,\cdot) - x\big)^+\Big)\Big| \ \mathbb{P} - a.s..$$

2) If I is some countable dense subset of $[-w_{r,n}, w_{r,n}]$, then a positive envelope of $\overline{\mathbb{F}}_I^{r,n} \triangleq \big\{\big(\phi_{r,n} \circ G(\theta,\cdot) - x\big)^+ \mid \theta \in \overline{\Theta}, x \in I\big\}$ is given by $\xi + w_{r,n}$, and

$$J(\overline{\mathbb{F}}_I^{r,n}, \xi + w_{r,n}, \delta) \le 4\ J(\mathbb{F}^{\Theta}, \xi, \delta/4) + 6 \quad \text{for } \delta \in (0,4].$$

Proof. Statement 1) follows immediately from assumption (A 3) along with the continuity of $\phi_{r,n}$ because

$$\begin{aligned}
&\big|\big(\phi_{r,n} \circ G(\theta,z) - x\big)^+ - \big(\phi_{r,n} \circ G(\vartheta,z) - y\big)^+\big| \\
&\le |\phi_{r,n} \circ G(\theta,z) - \phi_{r,n} \circ G(\vartheta,z)| + |x - y| \\
&\le |G(\theta,z) - G(\vartheta,z)| + |x - y|, \quad \text{for } \theta,\vartheta \in \Theta \text{ and } x, y \in \mathbb{R}.
\end{aligned}$$

Concerning statement 2), let I be any countable dense subset of $[-w_{r,n}, w_{r,n}]$. Condition (A 2) ensures that $\xi + w_{r,n}$ is a positive envelope of the function class $\overline{\mathbb{F}}_I^{r,n}$ defined in the display of statement 2).

For $\theta \in \Theta, x \in \mathbb{R}$, we set $F(\theta,x,\cdot) \triangleq \big(G(\theta,\cdot) - x\big)^+$, so that by definition the class $\overline{\mathbb{F}}_I^{r,n}$ consists of all functions $F(\theta,x,\cdot)$ with $(\theta,x) \in \overline{\Theta} \times I$. In the next step, we want to derive

upper estimates for the uniform entropy integrals $J\big(\overline{\mathbb{F}}_I^{r,n}, \xi + w_{r,n}, \delta\big)$. Therefore, let us fix $\eta > 0$ and $\mathbb{Q} \in \mathcal{M}_{\text{fin}}$ with support $supp(\mathbb{Q})$. Consider any $(\theta, x), (\vartheta, y) \in \overline{\Theta} \times I$ such that the inequalities $\|G(\theta,\cdot) - G(\vartheta,\cdot)\|_{\mathbb{Q},2} \leq \eta \, \|\xi\|_{\mathbb{P}^z,2}/2$ and $|x-y| \leq \eta \, w_{r,n}/2$ are valid. Since $\phi_{r,n}$ is 1-Lipschitz continuous, then

$$\begin{aligned}
\|F(\theta,x,\cdot) - F(\vartheta,y,\cdot)\|_{\mathbb{Q},2}^2 &= \sum_{z \in supp(\mathbb{Q})} \mathbb{Q}(\{z\}) \; |F(\theta,x,z) - F(\vartheta,y,z)|^2 \\
&\leq 4 \sum_{z \in supp(\mathbb{Q})} \mathbb{Q}(\{z\}) \; \big[|G(\theta,z) - G(\vartheta,z)|^2 + |x-y|^2\big] \\
&\leq 4 \, \|G(\theta,\cdot) - G(\vartheta,\cdot)\|_{\mathbb{Q},2}^2 + 4 \, |x-y|^2 \\
&\leq \eta^2 \, \|\xi + w_{r,n}\|_{\mathbb{Q},2}^2.
\end{aligned}$$

Hence, setting $\mathbb{F}^{\overline{\Theta}} \triangleq \{G(\theta,\cdot) \mid \theta \in \overline{\Theta}\}$,

$$\begin{aligned}
&N\big(\eta \|\xi + w_{r,n}\|_{\mathbb{Q},2}, \overline{\mathbb{F}}^{r,n}, L^2(\mathbb{Q})\big) \\
&\leq N\big(\eta \|\xi\|_{\mathbb{Q},2}/2, \mathbb{F}^{\overline{\Theta}}, L^2(\mathbb{Q})\big) \cdot N\big(\eta \; w_{r,n}/2, I, |\cdot|\big) \\
&\leq N\big(\eta \|\xi\|_{\mathbb{Q},2}/4, \mathbb{F}^{\Theta}, L^2(\mathbb{Q})\big) \cdot N\big(\eta \; w_{r,n}/4, [-w_{r,n}, w_{r,n}], |\cdot|\big),
\end{aligned}$$

where for nonvoid bounded $J \subseteq \mathbb{R}$ and $\overline{\eta} > 0$, we denote by $N\big(\overline{\eta}, J, |\cdot|\big)$ the minimal number to cover J by intervals of the form $[x - \overline{\eta}, x + \overline{\eta}]$ with $x \in J$. Note that

$$N\big(\eta \; w_{r,n}/4, [-w_{r,n}, w_{r,n}], |\cdot|\big) \leq 8/\eta$$

holds. Then using change of variable formula we may conclude for any $\delta \in (0, 4]$

$$\begin{aligned}
&J(\overline{\mathbb{F}}_I^{r,n}, \xi + (w_{r,n} + B), \delta) \\
&= \int_0^\delta \sup_{\mathbb{Q} \in \mathcal{M}_{\text{fin}}} \sqrt{\log\big(2 \; N(\eta \; \|\xi + (w_{r,n} + B)\|_{\mathbb{Q}}, \overline{\mathbb{F}}_I^{r,n}, L^2(\mathbb{Q}))\big)} \mathrm{d}\eta \\
&\leq \int_0^\delta \sup_{\mathbb{Q} \in \mathcal{M}_{\text{fin}}} \sqrt{\log\big(2N(\eta \|\xi\|_{\mathbb{Q},2}/4, \mathbb{F}^{\Theta}, L^2(\mathbb{Q})) + \log(8/\eta)\big)} \mathrm{d}\eta \\
&\leq \int_0^\delta \sup_{\mathbb{Q} \in \mathcal{M}_{\text{fin}}} \sqrt{\log\big(2N(\eta \|\xi\|_{\mathbb{Q},2}/4, \mathbb{F}^{\Theta}, L^2(\mathbb{Q}))\big)} \mathrm{d}\eta + \int_0^\delta \sqrt{\log(8/\eta)} \mathrm{d}\eta \\
&= 4 \; J(\mathbb{F}^{\Theta}, \xi, \delta/4) + 4 \int_0^{\delta/4} \sqrt{\log(2/u)} \; du.
\end{aligned}$$

Furthermore, we may use a change of variable formula along with integration by parts to

obtain

$$\int_0^{\delta/4} \sqrt{\log(2/\epsilon)}\mathrm{d}\epsilon \leq \int_0^1 \sqrt{\log(e/\epsilon)}\mathrm{d}\epsilon = \int_0^\infty \sqrt{1+t}e^{-t}\mathrm{dt}$$
$$\leq \int_0^\infty (1+t/2)e^{-t}dt = 3/2.$$

This shows statement 2) and completes the proof. □

3.7.5 Proofs of Results from Section 3.4

We start with the derivation of Lemma 3.16.

Proof of Lemma 3.16. By definition, we have

$$\rho_h(X) = \int_0^\infty h\left(\mathbb{P}[X > t]\right) dt + \int_{-\infty}^0 \left(h\left(\mathbb{P}[X > t]\right) - 1\right) dt.$$

Let $F^{\leftarrow}$ be the quantile function of X, and p_0 the separation point between the concave and convex part of h. We note that for any $t > F^{\leftarrow}(1-p_0)$, we have that $\mathbb{P}(X > t) \leq p_0$. Suppose first that $F^{\leftarrow}(1-p_0) \geq 0$. Then, we have that

$$\begin{aligned}
&\int_0^\infty h\left(\mathbb{P}[X > t]\right) dt \\
&= \int_0^{F^{\leftarrow}(1-p_0)} h\left(\mathbb{P}[X > t]\right) dt + \int_{F^{\leftarrow}(1-p_0)}^\infty h\left(\mathbb{P}[X > t]\right) dt \\
&= \int_0^{F^{\leftarrow}(1-p_0)} 1 - \bar{h}\left(\mathbb{P}[X \leq t]\right) dt + \int_{F^{\leftarrow}(1-p_0)}^\infty h_0\left(\mathbb{P}[X > t]\right) dt \\
&= \int_0^{F^{\leftarrow}(1-p_0)} 1 - h(p_0) - \bar{h}\left(\mathbb{P}[X \leq t]\right) dt + \int_0^\infty h_0\left(\mathbb{P}[X > t]\right) dt.
\end{aligned}$$

We also note that $t \geq -F^{\leftarrow}(1-p_0) \Leftrightarrow \mathbb{P}(-X \geq t) \leq 1-p_0$. Hence, the first integral is also equal to

$$\begin{aligned}
&\int_0^{F^{\leftarrow}(1-p_0)} 1 - h(p_0) - \bar{h}\left(\mathbb{P}[-X \geq -t]\right) dt \\
&= \int_{-F^{\leftarrow}(1-p_0)}^0 1 - h(p_0) - \bar{h}\left(\mathbb{P}[-X \geq s]\right) ds \\
&= \int_{-\infty}^0 1 - h(p_0) - \bar{h}_0\left(\mathbb{P}[-X \geq s]\right) ds \\
&= \int_{-\infty}^0 1 - h(p_0) - \bar{h}_0\left(\mathbb{P}[-X > s]\right) ds.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\int_{-\infty}^{0} \left(h\left(\mathbb{P}[X > t]\right) - 1\right) dt &= -\int_{-\infty}^{0} \bar{h}(\mathbb{P}(X \leq t))dt \\ &= -\int_{0}^{\infty} \bar{h}(\mathbb{P}(-X \geq s))ds \\ &= -\int_{0}^{\infty} \bar{h}_0(\mathbb{P}(-X > s))ds.\end{aligned}$$

Moreover,

$$\int_{-\infty}^{0} h_0(\mathbb{P}(X > t)) - h(p_0)dt = 0,$$

since for $t \leq F^{\leftarrow}(1-p_0)$, $h_0(\mathbb{P}(X > t)) = h(p_0)$. Therefore, assembling all equalities yields that

$$\begin{aligned}\rho_h(X) &= \int_{0}^{\infty} h_0\left(\mathbb{P}[X > t]\right) dt + \int_{-\infty}^{0} h_0(\mathbb{P}(X > t)) - h(p_0)dt \\ &\quad - \left(\int_{0}^{\infty} \bar{h}_0(\mathbb{P}(-X > s))ds + \int_{-\infty}^{0} \bar{h}_0\left(\mathbb{P}[-X > s]\right) - (1 - h(p_0))ds\right) \\ &= h(p_0)\rho_{h_0/h(p_0)}(X) - (1 - h(p_0))\rho_{\bar{h}_0/(1-h(p_0))}(-X).\end{aligned}$$

Now, suppose $F^{\leftarrow}(1-p_0) < 0$. Then,

$$\begin{aligned}&\int_{-\infty}^{0} \left(h\left(\mathbb{P}[X > t]\right) - 1\right) dt \\ &= \int_{-\infty}^{F^{\leftarrow}(1-p_0)} \left(h\left(\mathbb{P}[X > t]\right) - 1\right) dt + \int_{F^{\leftarrow}(1-p_0)}^{0} \left(h\left(\mathbb{P}[X > t]\right) - 1\right) dt \\ &= -\int_{-\infty}^{F^{\leftarrow}(1-p_0)} \bar{h}\left(\mathbb{P}[X \leq t]\right) dt + \int_{F^{\leftarrow}(1-p_0)}^{0} \left(h_0\left(\mathbb{P}[X > t]\right) - 1\right) dt \\ &= -\int_{-F^{\leftarrow}(1-p_0)}^{\infty} \bar{h}_0\left(\mathbb{P}[-X \geq s]\right) ds + \int_{F^{\leftarrow}(1-p_0)}^{0} \left(h_0\left(\mathbb{P}[X > t]\right) - 1\right) dt.\end{aligned}$$

We note that

$$\begin{aligned}&\int_{-F^{\leftarrow}(1-p_0)}^{\infty} \bar{h}_0\left(\mathbb{P}[-X \geq s]\right) ds \\ &= \int_{0}^{\infty} \bar{h}_0\left(\mathbb{P}[-X \geq s]\right) ds + (1 - h(p_0))F^{\leftarrow}(1-p_0),\end{aligned}$$

and

$$\int_{F^{\leftarrow}(1-p_0)}^{0} \left(h_0\left(\mathbb{P}[X > t]\right) - 1\right) dt$$
$$= \int_{-\infty}^{0} \left(h_0\left(\mathbb{P}[X > t]\right) - h(p_0)\right) dt + (1 - h(p_0))F^{\leftarrow}(1 - p_0).$$

Hence,

$$\int_{-\infty}^{0} \left(h\left(\mathbb{P}[X > t]\right) - 1\right) dt$$
$$= \int_{-\infty}^{0} \left(h_0\left(\mathbb{P}[X > t]\right) - h(p_0)\right) dt - \int_{0}^{\infty} \bar{h}_0\left(\mathbb{P}[-X \geq s]\right) ds.$$

Moreover,

$$\int_0^\infty h(\mathbb{P}(X > t))dt = \int_0^\infty h_0(\mathbb{P}(X > t))dt,$$

and

$$\int_{-\infty}^{0} \bar{h}_0(\mathbb{P}(-X > s)) - (1 - h(p_0))ds = 0.$$

Hence, we again have

$$\rho_h(X) = \int_0^\infty h_0(\mathbb{P}(X > t))dt + \int_{-\infty}^{0} \left(h_0\left(\mathbb{P}[X > t]\right) - h(p_0)\right) dt$$
$$- \int_0^\infty \bar{h}_0\left(\mathbb{P}[-X \geq s]\right) ds - \int_{-\infty}^{0} \bar{h}_0(\mathbb{P}(-X > s)) - (1 - h(p_0))ds.$$

□

Let us turn over to the proof of Theorem 3.17

Proof of Theorem 3.17. Let $\theta \in \Theta$ and $F_\theta^-, \widehat{F}_{n,\theta}^-$ denote respectively the true and empirical distribution functions of $-G(\theta, Z)$. Then, we have that

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left|\rho_h(\widehat{F}_{n,\theta}) - \rho_h(F_\theta)\right| > \epsilon\right)$$
$$\leq \mathbb{P}\left(\sup_{\theta\in\Theta}\left|\rho_{h_{cc}}(\widehat{F}_{n,\theta}) - \rho_{h_{cc}}(F_\theta)\right| \geq \epsilon/(2h(p_0))\right)$$
$$+ \mathbb{P}\left(\sup_{\theta\in\Theta}\left|\rho_{h_{cv}}(\widehat{F}_{n,\theta}^-) - \rho_{h_{cv}}(F_\theta^-)\right| \geq \epsilon/(2(1 - h(p_0)))\right).$$

Note that the class $\{-G(\theta, Z) : \theta \in \Theta\}$ is also uniformly bounded by B and has the same covering numbers as the class $\mathbb{F}^\Theta$. Hence, applying Theorem 3.22 to each separate term

gives the statement. □

3.7.6 Proof of Theorem 3.20 and Theorem 3.22

To begin the proof, let us first introduce the following notation:

$$L(n,q,r,\delta) \triangleq \left(\frac{q}{r\ (2n)^{(r-q)/r}}\right)^{1/q} 2\ (3\overline{M}^{\xi}_{r,q}+\delta)\ \|\xi\|_{\mathbb{P}^Z,r},\ (\delta>0) \tag{3.28}$$

$$\tilde{L}(n,q,r) \triangleq 2\overline{N}_q\|\xi\|_{\mathbb{P}^Z,r}\frac{q}{r-q}\frac{1}{n^{1/q-1/r}}. \tag{3.29}$$

The following proposition allows us to reduce the unbounded problem to a bounded one.

Proposition 3.25. *If $r \in [2,\infty)\cap(q,\infty)$, then the inequalities*

$$\begin{aligned}
&\mathbb{P}\Big(\Big\{\big|\inf_{\theta\in\Theta}\mathcal{R}_\rho\big(\widehat{F}_{n,\theta}\big)-\inf_{\theta\in\Theta}\mathcal{R}_\rho\big(F_\theta\big)\big|\geq\varepsilon\Big\}\ \cap\ B^{\xi}_{n,r}\Big)\\
&\leq \mathbb{P}^*\Big(\Big\{\sup_{\theta\in\Theta}\big|\mathcal{R}_\rho\big(\widehat{F}_{n,\theta}\big)-\mathcal{R}_\rho\big(F_\theta\big)\big|\geq\varepsilon\Big\}\ \cap\ B^{\xi}_{n,r}\Big)\\
&\leq \mathbb{P}^*\Big(\Big\{\sup_{\theta\in\Theta}\big|\mathcal{R}_\rho\big(\widehat{F}^{\phi_{r,n}}_{n,\theta}\big)-\mathcal{R}_\rho\big(F^{\phi_{r,n}}_{\theta}\big)\big|\geq\epsilon-L(n,r,q,\delta)\Big\}\ \cap\ B^{\xi}_{n,r}\Big)
\end{aligned}$$

hold for any $\delta,\varepsilon>0$ and every $n\in\mathbb{N}$. Here $\mathbb{P}^*$ denotes the outer probability w.r.t. $\mathbb{P}$. Moreover, if ρ is coherent, then one may replace $L(n,r,q,\delta)$ with $\tilde{L}(n,q,r)$.

The proof is relegated to Section 3.7.7.

In the next step, we want to bound the penalty function in Kusuoka representation (3.8) for $\mathcal{R}_\rho(\widehat{F}^{\phi_{r,n}}_{n,\theta})$ and $\mathcal{R}_\rho(F^{r,n}_\theta)$ on the auxiliary event $B^{\xi}_{n,r}$.

Proposition 3.26. *Under assumptions (A 1), (A 2) and (A 6) with q,r as well as $\overline{M}^{\xi}_{q,r}$ as in (A 6), we obtain the following Kusuoka representation for $\theta\in\Theta, n\in\mathbb{N}$ and $\delta>0$:*

$$\begin{aligned}
\mathcal{R}_\rho(\widehat{F}^{\phi_{r,n}}_{n,\theta|_\omega}) &= \sup_{\substack{\mu\in\mathrm{dom}(\beta_\rho)\\ \beta_\rho(\mu)\leq 2\overline{M}^{\xi}_{r,q}+\delta}}\Big(\int_{[0,1)}\mathcal{R}_\alpha(\widehat{F}^{\phi,r}_{n,\theta|_\omega})\ \mu(d\alpha)-\beta_\rho(\mu)\Big)\quad \text{if } \omega\in B^{\xi}_{n,r},\\
\mathcal{R}_\rho(F^{\phi_{r,n}}_{\theta}) &= \sup_{\substack{\mu\in\mathrm{dom}(\beta_\rho)\\ \beta_\rho(\mu)\leq 2\overline{M}^{\xi}_{r,q}+\delta}}\Big(\int_{[0,1)}\mathcal{R}_\alpha(F^{\phi_{r,n}}_{\theta})\ \mu(d\alpha)-\beta_\rho(\mu)\Big),
\end{aligned}$$

where $\mathcal{R}_\alpha$ stands for the functional associated with $AV@R_\alpha$.

Proof. Let $\theta\in\Theta, n\in\mathbb{N}$ and $\delta>0$. We fix any $\omega\in B^{\xi}_{n,r}$ and denote by $\widehat{F}_{n,\xi}(\cdot,\omega)$ the empirical distribution function based on $\xi\big(Z_1(\omega)\big),\dots,\xi\big(Z_n(\omega)\big)$.

Since $(\Omega,\mathcal{F},\mathbb{P})$ is atomless, we may find random variables $X^{(\omega)},Y^{(\omega)}$ on $(\Omega,\mathcal{F},\mathbb{P})$ which have respectively $\widehat{F}_{n,\xi}(\cdot,\omega)$ and $\widehat{F}^{\phi_{r,n}}_{n,\theta|_\omega}$ as distribution functions. Both random variables are $\mathbb{P}$-essentially bounded so that they belong to $\mathcal{X}$. Furthermore, the distribution functions

$F_{-X^{(\omega)}}, F_{Y^{(\omega)}}, F_{X^{(\omega)}}$ of $-X^{(\omega)}, Y^{(\omega)}$ and $X^{(\omega)}$ respectively, satisfy pointwise the inequalities $F_{X^{(\omega)}} \leq F_{Y^{(\omega)}} \leq F_{-X^{(\omega)}}$ due to (A 2). Moreover, by (A 2) again, we also have $-\xi(Z_1) \leq \phi_{r,n}(G(\theta, Z_1)) \leq \xi(Z_1)$. Hence, by Proposition 3.A.2 (see Appendix 3.A)

$$\mathcal{R}_\rho(\widehat{F}^{\phi_{r,n}}_{n,\theta|_\omega}) = \sup_{\substack{\mu\in\mathrm{dom}(\beta_\rho)\\ \beta_\rho(\mu)\leq 2\rho(2X^{(\omega)})+\delta}} \Big(\int_{[0,1)} \mathcal{R}_\alpha(\widehat{F}^{\phi_{r,n}}_{n,\theta|_\omega})\ \mu(d\alpha) - \beta_\rho(\mu)\Big) \quad \text{if } \omega \in B^\xi_{n,r},$$

$$\mathcal{R}_\rho(F^{\phi_{r,n}}_{\theta}) = \sup_{\substack{\mu\in\mathrm{dom}(\beta_\rho)\\ \beta_\rho(\mu)\leq 2\rho(2\xi(Z_1))+\delta}} \Big(\int_{[0,1)} \mathcal{R}_\alpha(F_\theta)\ \mu(d\alpha) - \beta_\rho(\mu)\Big).$$

Furthermore, by Lemma 4.3 in Bartl and Tangpi [2022], and recalling $r \geq q$ as well as $\omega \in B^\xi_{n,r}$,

$$\begin{aligned} AV@R_\alpha(2X^{(\omega)}) \leq \frac{2\left(\frac{1}{n}\sum_{j=1}^n \xi_j\big(Z_j(\omega)\big)^q\right)^{1/q}}{(1-\alpha)^{1/q}} &\leq \frac{2\left(\frac{1}{n}\sum_{j=1}^n \xi_j\big(Z_j(\omega)\big)^r\right)^{1/r}}{(1-\alpha)^{1/q}} \\ &\leq \frac{2^{1+1/r}\|\xi\|_{\mathbb{P}^Z,r}}{(1-\alpha)^{1/q}}. \end{aligned}$$

Also, by Lemma 4.3 in Bartl and Tangpi [2022],

$$AV@R_\alpha(2\xi) \leq \frac{2\|\xi\|_{\mathbb{P}^Z,q}}{(1-\alpha)^{1/q}} \leq \frac{2\|\xi\|_{\mathbb{P}^Z,r}}{(1-\alpha)^{1/q}}.$$

So we may conclude $2\rho(2X^{(\omega)}) \vee 2\rho(2\xi) \leq 2\overline{M}^\xi_{r,q}$ which completes the proof. □

Now we are ready to show Theorem 3.20 and 3.22.

Proof of Theorem 3.20. Let us fix $n \in \mathbb{N}$ and $\delta, \varepsilon > 0$, and let I be any countable dense subset of $[-w_{r,n}, w_{r,n}]$, whereas $\overline{\Theta}$ is the at most countable subset of Θ from assumption (A 3).

Clearly, the inequality

$$\mathbb{P}\left(\Big\{\Big|\inf_{\theta\in\Theta}\mathcal{R}_\rho(F_\theta) - \inf_{\theta\in\Theta}\mathcal{R}_\rho(\widehat{F}_{n,\theta})\Big| \geq \varepsilon\Big\}\right) \leq \mathbb{P}^*\left(\Big\{\sup_{\theta\in\Theta}\Big|\mathcal{R}_\rho(F_\theta) - \mathcal{R}_\rho(\widehat{F}_{n,\theta})\Big| \geq \varepsilon\Big\}\right)$$

is valid. Then, since ρ is coherent, Proposition 3.25 implies that

$$\begin{aligned} &\mathbb{P}^*\left(\Big\{\sup_{\theta\in\Theta}\Big|\mathcal{R}_\rho(F_\theta) - \mathcal{R}_\rho(\widehat{F}_{n,\theta})\Big| \geq \varepsilon\Big\} \cap B^\xi_{n,r}\right) \\ &\leq \mathbb{P}^*\Big(\Big\{\sup_{\theta\in\Theta}\big|\mathcal{R}_\rho\big(\widehat{F}^{\phi_{r,n}}_{n,\theta}\big) - \mathcal{R}_\rho\big(F^{\phi_{r,n}}_{\theta}\big)\big| \geq \epsilon - \tilde{L}(n,r,q)\Big\}\ \cap\ B^\xi_{n,r}\Big). \end{aligned}$$

By the Kusuoka representation (3.8), we have that

$$\begin{aligned}\left|\mathcal{R}_\rho\big(\widehat{F}_{n,\theta}^{\phi_{r,n}}\big) - \mathcal{R}_\rho\big(F_\theta^{\phi_{r,n}}\big)\right| &\leq \sup_{\mu:\beta_\rho(\mu)=0} \int_{[0,1)} \left|\mathcal{R}_\alpha\big(\widehat{F}_{n,\theta}^{\phi_{r,n}}\big) - \mathcal{R}_\alpha\big(F_\theta^{\phi_{r,n}}\big)\right| \mu(d\alpha) \\ &\leq \sup_{\mu:\beta_\rho(\mu)=0} \int_{[0,1)} \frac{\Delta_n}{1-\alpha} \wedge (2w_{r,n})\mu(d\alpha),\end{aligned}$$

where $\Delta_n = \sup_{f\in\mathbb{F}_{w_{r,n}}^{\Theta,\phi_{r,n}}} |(\mathbb{P}_n - \mathbb{P})(f)|$ with $\mathbb{F}_{w_{r,n}}^{\Theta,\phi_{r,n}}$ as defined in (3.27). Then, following the proof of Theorem 3.10, we obtain that

$$\sup_{\theta\in\Theta}\left|\mathcal{R}_\alpha\big(\widehat{F}_{n,\theta}^{\phi_{r,n}}\big) - \mathcal{R}_\alpha\big(F_\theta^{\phi_{r,n}}\big)\right| \leq \overline{N}_q 2^{(2q-1)/q} w_{r,n}^{1-1/q} \Delta_n^{1/q}.$$

Hence, we obtain,

$$\begin{aligned}&\mathbb{P}^*\Big(\Big\{\sup_{\theta\in\Theta}\big|\mathcal{R}_\rho\big(\widehat{F}_{n,\theta}^{\phi_{r,n}}\big) - \mathcal{R}_\rho\big(F_\theta^{\phi_{r,n}}\big)\big| \geq \epsilon - \tilde{L}(n,r,q)\Big\} \cap B_{n,r}^{\xi}\Big) \\ &\leq \mathbb{P}^*\left(\left\{\sup_{f\in\overline{\mathbb{F}}_I^{r,n}}\big|(\mathbb{P}_n-\mathbb{P})(f)\big| \geq \frac{[\varepsilon - \tilde{L}(n,q,r)]^q}{\overline{N}_q^q 2^{2q-1} w_{r,n}^{(q-1)}}\right\}\right),\end{aligned}$$

where

$$\overline{\mathbb{F}}_I^{r,n} \triangleq \big\{\big(\phi_{r,n}\circ G(\theta,\cdot) - x\big)^+ \mid (\theta,x)\in\overline{\Theta}\times I\big\}, \quad \tilde{L}(n,q,r) = 2\overline{N}_q \|\xi\|_{\mathbb{P}^Z,r} \frac{q}{r-q}\frac{1}{n^{1/q-1/r}}.$$

To further estimate the deviation probability, we want to apply the concentration inequality (Theorem in Appendix 3.B.4) to $\overline{\mathbb{F}}_I^{r,n}$.

Firstly, every member of this class is Borel measurable, and $\xi + w_{r,n}$ is a positive envelope. Since in addition $\overline{\mathbb{F}}_I^{r,n}$ is at most countable, we may apply Theorem 2.1 from Krätschmer [2024a] to conclude

$$\mathbb{E}\left[\sup_{f\in\overline{\mathbb{F}}_I^{r,n}}\big|(\mathbb{P}_n-\mathbb{P})(f)\big|\right] \leq \frac{16\sqrt{2}\,\|\xi+w_{r,n}\|_{\mathbb{P}^Z,2}}{\sqrt{n}}\, J\big(\overline{\mathbb{F}}_I^{r,n}, \xi+w_{r,n}, 1/2\big).$$

Therefore, by Lemma 3.24,

$$\mathbb{E}\left[\sup_{f\in\overline{\mathbb{F}}_I^{r,n}}\big|(\mathbb{P}_n-\mathbb{P})(f)\big|\right] \leq \frac{256\,\|\xi\|_{\mathbb{P}^Z,r}}{n^{1/2-1/r}}\,\big[J\big(\mathbb{F}^\Theta,\xi,1/8\big)+3/2\big]. \tag{3.30}$$

For abbreviation, we set

$$\begin{aligned}\overline{\varepsilon} &\triangleq \frac{\varepsilon}{\overline{N}_q 2^{(2q-1)/q} w_{r,n}^{(q-1)/q}} \\ \overline{L} &\triangleq \frac{\tilde{L}(n,q,r)}{\overline{N}_q 2^{(2q-1)/q} w_{r,n}^{(q-1)/q}}.\end{aligned}$$

Now, if ε satisfies the lower bound in the statement, which is

$$\epsilon \geq 2\tilde{L}(n,q,r) + \frac{((t+1)256\|\xi\|_{\mathbb{P}^Z,r}(J(\mathbb{F}^\Theta,\xi,1/8)+3/2))^{1/q}\overline{N}_q 2^{(2q-1)/q} w_{r,n}^{(q-1)/q}}{n^{1/(2q)-1/(qr)}},$$

then inequality (3.30) implies

$$(\overline{\varepsilon}-\overline{L})^q > \frac{\overline{\varepsilon}^q}{2^q} \vee \mathbb{E}\left[(t+1)\sup_{f\in\overline{\mathbb{F}}_I^{r,n}}\left|(\mathbb{P}_n-\mathbb{P})(f)\right|\right].$$

Hence

$$(\overline{\varepsilon}-\overline{L})^q = \frac{t}{t+1}\,(\overline{\varepsilon}-\overline{L})^q + \frac{(\overline{\varepsilon}-\overline{L})^q}{t+1} > \frac{t}{t+1}\,\frac{\overline{\varepsilon}^q}{2^q} + \mathbb{E}\left[\sup_{f\in\overline{\mathbb{F}}_I^{r,n}}\left|(\mathbb{P}_n-\mathbb{P})(f)\right|\right].$$

Finally, note that the mappings from $\overline{\mathbb{F}}_I^{r,n}$ are uniformly bounded by $2w_{r,n}$. Thus, applying Theorem 3.B.4 (see Appendix 3.B) to $\overline{\mathbb{F}}_I^{r,n}$ yields

$$\begin{aligned}
&\mathbb{P}\left(\left\{\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(F_\theta)-\mathcal{R}_\rho(\widehat{F}_{n,\theta})\right|\geq\varepsilon\right\}\cap B_{n,r}^\xi\right)\\
&\leq \mathbb{P}\left(\left\{\sup_{f\in\overline{\mathbb{F}}_I^{r,n}}\left|(\mathbb{P}_n-\mathbb{P})(f)\right|\geq(\overline{\varepsilon}-\overline{L})^q\right\}\right)\\
&\leq \mathbb{P}\left(\left\{\sup_{f\in\overline{\mathbb{F}}_I^{r,n}}\left|(\mathbb{P}_n-\mathbb{P})(f)\right|\geq\frac{t}{t+1}\,\frac{\overline{\varepsilon}^q}{2^q}+\mathbb{E}\left[\sup_{f\in\overline{\mathbb{F}}_I^{r,n}}\left|(\mathbb{P}_n-\mathbb{P})(f)\right|\right]\right\}\right)\\
&\leq \exp\left(-\frac{n\;t^2\;\overline{\varepsilon}^{2q}}{2^{2q+3}\;(t+1)^2 w_{r,n}^2}\right)\\
&= \exp\left(-\frac{n^{1-2q/r}\;t^2\;\varepsilon^{2q}}{2^{4q+3}\;(t+1)^2\overline{N}_q^{2q}\|\xi\|_{\mathbb{P}^Z,r}^{2q}}\right).
\end{aligned}$$

□

Proof of Theorem 3.22. We modify the proof of Theorem 3.20. Denote $b_{r,\delta} \triangleq 2\overline{M}_{r.q}^{\xi}+\delta$ and $\mathcal{M}_{b_{r,\delta}} \triangleq \{\mu\in\mathcal{M} \mid \beta_\rho(\mu)\leq b_{r,\delta}\}$. Then, Proposition 3.25 implies that for general convex risk measure ρ,

$$\begin{aligned}
&\mathbb{P}^*\left(\left\{\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(F_\theta)-\mathcal{R}_\rho(\widehat{F}_{n,\theta})\right|\geq\varepsilon\right\}\cap B_{n,r}^\xi\right)\\
&\leq \mathbb{P}^*\Big(\Big\{\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho\big(\widehat{F}_{n,\theta}^{\phi_{r,n}}\big)-\mathcal{R}_\rho\big(F_\theta^{\phi_{r,n}}\big)\right|\geq\epsilon-L(n,r,q,\delta)\Big\}\;\cap\;B_{n,r}^\xi\Big),
\end{aligned}$$

where $L(n,q,r,\delta)$ is as defined in (3.28). By the Kusuoka representation (3.8), we have

that

$$\begin{aligned}\left|\mathcal{R}_\rho\big(\widehat{F}_{n,\theta}^{\phi_{r,n}}\big) - \mathcal{R}_\rho\big(F_\theta^{\phi_{r,n}}\big)\right| &\le \sup_{\mu\in\mathcal{M}_{b_{r,\delta}}} \int_{[0,1)} \left|\mathcal{R}_\alpha\big(\widehat{F}_{n,\theta}^{\phi_{r,n}}\big) - \mathcal{R}_\alpha\big(F_\theta^{\phi_{r,n}}\big)\right| \mu(d\alpha) \\ &\le \sup_{\mu\in\mathcal{M}_{b_{r,\delta}}} \int_{[0,1)} \frac{\Delta_n}{1-\alpha} \wedge 2w_{r,n}\mu(d\alpha),\end{aligned}$$

where $\Delta_n = \sup_{f\in\mathbb{F}_{w_{r,n}}^{\Theta,\phi_{r,n}}} |(\mathbb{P}_n - \mathbb{P})(f)|$ with $\mathbb{F}_{w_{r,n}}^{\Theta,\phi_{r,n}}$ as defined in (3.27). Then, proceed the same as in the proof of Theorem 3.20 to obtain that

$$\begin{aligned}&\mathbb{P}^*\Big(\Big\{\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho\big(\widehat{F}_{n,\theta}^{\phi_{r,n}}\big) - \mathcal{R}_\rho\big(F_\theta^{\phi_{r,n}}\big)\right| \ge \epsilon - L(n,r,q,\delta)\Big\} \cap B_{n,r}^{\xi}\Big) \\ &\le \mathbb{P}^*\left(\left\{\sup_{f\in\overline{\mathbb{F}}_I^{r,n}} \big|(\mathbb{P}_n - \mathbb{P})(f)\big| \ge \frac{[\varepsilon - L(n,q,r,\delta)]^q}{N_{q,b_{r,\delta}}^q 2^{2q-1} w_{r,n}^{(q-1)}}\right\}\right),\end{aligned}$$

where $N_{q,b_{r,\delta}}$ is the constant defined in assumption (A 5), which is finite by assumption (A 6). Next, set

$$\begin{aligned}\overline{\varepsilon} &\triangleq \frac{\varepsilon}{N_{q,b_{r,\delta}} 2^{(2q-1)/q} w_{r,n}^{(q-1)/q}} \\ \overline{L} &\triangleq \frac{L(n,q,r,\delta)}{N_{q,b_{r,\delta}} 2^{(2q-1)/q} w_{r,n}^{(q-1)/q}}.\end{aligned}$$

Following the proof of Theorem 3.20 further, if ε satisfies the lower bound in the statement, which means

$$\epsilon \ge 2L(n,q,r,\delta) + \frac{((t+1)256\|\xi\|_{\mathbb{P}^Z,r}(J\big(\mathbb{F}^\Theta,\xi,1/8\big)+3/2))^{1/q} N_{q,b_{r,\delta}} 2^{(2q-1)/q} w_{r,n}^{(q-1)/q}}{n^{1/(2q)-1/(qr)}},$$

then we obtain by applying Theorem 3.B.4 (see Appendix 3.B) to $\overline{\mathbb{F}}_I^{r,n}$ that

$$\begin{aligned}&\mathbb{P}^*\left(\left\{\sup_{\theta\in\Theta}\left|\mathcal{R}_\rho(F_\theta) - \mathcal{R}_\rho(\widehat{F}_{n,\theta})\right| \ge \varepsilon\right\} \cap B_{n,r}^{\xi}\right) \\ &\le \mathbb{P}^*\left(\left\{\sup_{f\in\overline{\mathbb{F}}_I^{r,n}} \big|(\mathbb{P}_n - \mathbb{P})(f)\big| \ge (\overline{\varepsilon} - \overline{L})^q\right\}\right) \\ &\le \mathbb{P}^*\left(\left\{\sup_{f\in\overline{\mathbb{F}}_I^{r,n}} \big|(\mathbb{P}_n - \mathbb{P})(f)\big| \ge \frac{t}{t+1}\,\frac{\overline{\varepsilon}^q}{2^q} + \mathbb{E}\left[\sup_{f\in\overline{\mathbb{F}}_I^{r,n}} \big|(\mathbb{P}_n - \mathbb{P})(f)\big|\right]\right\}\right) \\ &\le \exp\left(-\frac{n\ t^2\ \overline{\varepsilon}^{2q}}{2^{2q+1}\ (t+1)^2 w_{r,n}^2}\right) \\ &= \exp\left(-\frac{n^{1-2q/r}\ t^2\ \varepsilon^{2q}}{2^{6q+1}\ (t+1)^2 N_{q,b_{r,\delta}}^{2q} \|\xi\|_{\mathbb{P}^Z,r}^{2q}}\right).\end{aligned}$$

□

3.7.7 Proof of Proposition 3.25

The starting point is the following auxiliary result.

Lemma 3.27. *Let ξ from (A 2) be $\mathbb{P}^Z$-integrable of order r for some $r \geq 2$. Then,*

$$\begin{aligned}
&\Big\{ \big| \inf_{\theta\in\Theta} \mathcal{R}_\rho\big(\widehat{F}_{n,\theta}\big) - \inf_{\theta\in\Theta} \mathcal{R}_\rho\big(F_\theta\big) \big| \geq \epsilon \Big\} \ \cap \ B^\xi_{n,r} \\
&\subseteq \Big\{ \sup_{\theta\in\Theta} \big| \mathcal{R}_\rho\big(\widehat{F}_{n,\theta}\big) - \mathcal{R}_\rho\big(F_\theta\big) \big| \geq \epsilon \Big\} \ \cap \ B^\xi_{n,r} \\
&\subseteq \Big\{ \sup_{\theta\in\Theta} \big| \mathcal{R}_\rho\big(\widehat{F}^{\phi_{r,n}}_{n,\theta}\big) - \mathcal{R}_\rho\big(F^{\phi_{r,n}}_{\theta}\big) \big| + \sup_{\theta\in\Theta} \big| \mathcal{R}_\rho\big(F^{\phi_{r,n}}_{\theta}\big) - \mathcal{R}_\rho\big(F_\theta\big) \big| \geq \epsilon \Big\} \ \cap \ B^\xi_{n,r},
\end{aligned}$$

for $\epsilon > 0$ and $n \in \mathbb{N}$.

Proof. This follows from the fact that on $B^\xi_{n,r}$, we have that

$$\frac{\xi(Z_j)^r}{n} \leq \frac{1}{n} \sum_{j=1}^n \xi(Z_j)^r \leq 2\mathbb{E}[\xi(Z_1)^r],$$

which implies that $\xi(Z_j) \leq (2n)^{1/r} \|\xi\|_{\mathbb{P}^Z,r}$. In particular, this means that $\widehat{F}^{\phi_{r,n}}_{n,\theta} = \widehat{F}_{n,\theta}$ on the event $B^\xi_{n,r}$. □

It remains to provide an upper estimate for

$$\sup_{\theta\in\Theta} \big| \mathcal{R}_\rho\big(F^{\phi_{r,n}}_{\theta}\big) - \mathcal{R}_\rho\big(F_\theta\big) \big|, \quad (n \in \mathbb{N}).$$

Proposition 3.28. *Let (A 1), (A 2) be fulfilled with ξ from (A 2), and let $r, q \in [1, \infty)$ and $\overline{M}^\xi_{r,q}$ be as in (A 6). If $r \in [2, \infty) \cap (q, \infty)$, then*

$$\sup_{\theta\in\Theta} \big| \mathcal{R}_\rho\big(F^{\phi_{r,n}}_{\theta}\big) - \mathcal{R}_\rho\big(F_\theta\big) \big| \leq L(n, r, q, \delta), \quad \text{for } \delta > 0, \ n \in \mathbb{N}.$$

If ρ is coherent that satisfies assumption (A 4), then one may replace $L(n, r, q, \delta)$ with $\tilde{L}(n, r, q)$ as defined in (3.29).

Proof. Let $n \in \mathbb{N}$ and fix $\delta > 0$. We note that the following inequalities

$$-\xi(z) \leq G(\theta, z), \phi_{r,n}\big(G(\theta, z)\big) \leq \xi(z),$$

are valid for $\theta \in \Theta$ and $z \in \mathbb{R}^d$. Hence, by Proposition 3.A.2 in Appendix 3.A,

$$\begin{aligned}
&\big| \mathcal{R}_\rho\big(F^{\phi_{r,n}}_{\theta}\big) - \mathcal{R}_\rho\big(F_\theta\big) \big| \\
&\leq \sup_{\substack{\mu\in\mathrm{dom}(\beta) \\ \beta_\rho(\mu)\leq b_\delta}} \int_{[0,1)} \Big| AV@R_\alpha\Big(\phi_{r,n}\big(G(\theta, Z_1)\big)\Big) - AV@R_\alpha\big(G(\theta, Z_1)\big) \Big| \, \mu(d\alpha), \quad \text{for } \theta \in \Theta,
\end{aligned}$$

where $b_\delta \triangleq 2\rho\big(2\xi(Z_1)\big) + \delta$. Furthermore, by monotonicity and subadditivity of the Average Value at Risk, we obtain

$$\begin{aligned}
&\Big|AV@R_\alpha\Big(\phi_{r,n}\big(G(\theta,Z_1)\big)\Big) - AV@R_\alpha\big(G(\theta,Z_1)\big)\Big| \\
&\leq AV@R_\alpha\Big(\Big|\phi_{r,n}\big(G(\theta,Z_1)\big) - G(\theta,Z_1)\Big|\Big) \\
&= AV@R_\alpha\Big(\big(G(\theta,Z_1) + w_{r,n}\big)^- + \big(G(\theta,Z_1) - w_{r,n}\big)^+\Big) \\
&\leq AV@R_\alpha\Big(2\,\big(\xi(Z_1) - w_{r,n}\big)^+\Big),
\end{aligned}$$

for $\alpha \in [0,1)$. In addition, $2\,\big(\xi(Z_1) - w_{r,n}\big)^+$ is integrable of order $q < r$ by assumption on ξ, so that by the proof of Lemma 4.3 in Bartl and Tangpi [2022]

$$AV@R_\alpha\Big(2\,\big(\xi(Z_1) - w_{r,n}\big)^+\Big) \leq \frac{2\,\big\|\big(\xi(Z_1) - w_{r,n}\big)^+\big\|_{\mathbb{P}^Z,q}}{\big(1-\alpha\big)^{1/q}}, \quad \text{for } \alpha \in [0,1).$$

Therefore, we may conclude

$$\begin{aligned}
&\sup_{\theta\in\Theta}\big|\mathcal{R}_\rho\big(F_\theta^{\phi_{r,n}}\big) - \mathcal{R}_\rho\big(F_\theta\big)\big| \\
&\leq \sup_{\substack{\mu\in\mathrm{dom}(\beta_\rho)\\ \beta_\rho(\mu)\leq b_\delta+\delta}} \int_{[0,1)} \frac{2\,\big\|\big(\xi(Z_1) - w_{r,n}\big)^+\big\|_{\mathbb{P}^Z,q}}{\big(1-\alpha\big)^{1/q}}\,\mu(d\alpha). \qquad (3.31)
\end{aligned}$$

Since ξ is $\mathbb{P}^Z$-integrable of order $r > q$, and using in the first step the change of variable formula, we may further observe

$$\begin{aligned}
\big\|\big(\xi - w_{r,n}\big)^+\big\|^q_{\mathbb{P}^Z,q} &= q\int_{w_{r,n}}^\infty (y - w_{r,n})^{q-1}\,\mathbb{P}\big(\big\{\xi(Z_1) > y\big\}\big)\,dy \\
&= \frac{q}{w_{r,n}^{r-q}}\int_{w_{r,n}}^\infty w_{r,n}^{r-q}(y - w_{r,n})^{q-1}\,\mathbb{P}\big(\big\{\xi(Z_1) > y\big\}\big)\,dy \\
&\leq \frac{q}{w_{r,n}^{r-q}}\int_{w_{r,n}}^\infty y^{r-1}\,\mathbb{P}\big(\big\{\xi(Z_1) > y\big\}\big)\,dy \\
&\leq \frac{q}{r\;w_{r,n}^{r-q}}\,\|\xi\|^r_{\mathbb{P}^Z,r} = \frac{q}{r\;(2n)^{(r-q)/r}}\,\|\xi\|^q_{\mathbb{P}^Z,r}. \qquad (3.32)
\end{aligned}$$

Finally,

$$\int_{[0,1)} \frac{1}{(1-\alpha)^{1/q}}\,\mu(d\alpha) \leq \int_{[0,1)} \frac{(4\|\xi\|_{\mathbb{P}^Z,q}) \vee 1}{(1-\alpha)^{1/q}}\,\mu(d\alpha) \leq \overline{M}^\xi_{r,q} + \beta(\mu), \quad \text{for } \mu \in \mathrm{dom}(\beta).$$

Hence,

$$\sup_{\substack{\mu\in\mathrm{dom}(\beta_\rho)\\ \beta_\rho(\mu)\le b_\delta+\delta}} \int_{[0,1)} \frac{1}{\big(1-\alpha\big)^{1/q}}\mu(d\alpha) \le 3\overline{M}^{\xi}_{r,q} + \delta.$$

If ρ is coherent, then the estimation may be modified. Indeed, Since $AV@R_\alpha$ is also a distortion risk measure with $h(p) = \min\{p/(1-\alpha), 1\}$, we have

$$AV@R_\alpha\left(2\left(\xi(Z_1) - w_{r,n}\right)^+\right) = 2\int_{w_{r,n}}^{\infty} \frac{(1-F_\xi(y)) \wedge (1-\alpha)}{1-\alpha} dy,$$

where $F_\xi(y)$ is the distribution function of ξ. Then, we have

$$\begin{aligned}
&\sup_{\theta\in\Theta} \left|\mathcal{R}_\rho\big(F_\theta^{\phi_{r,n}}\big) - \mathcal{R}_\rho\big(F_\theta\big)\right| \\
&\le \sup_{\substack{\mu\in\mathrm{dom}(\beta)\\ \beta_\rho(\mu)\le 0}} \int_{[0,1)} 2\int_{w_{r,n}}^{\infty} \frac{(1-F_\xi(y)) \wedge (1-\alpha)}{1-\alpha} dy\ \mu(d\alpha) \\
&= \sup_{\substack{\mu\in\mathrm{dom}(\beta)\\ \beta_\rho(\mu)\le 0}} 2\int_{w_{r,n}}^{\infty} \int_{[0,1)} \frac{(1-F_\xi(y)) \wedge (1-\alpha)}{1-\alpha} dy\ \mu(d\alpha) \\
&\le 2\int_{w_{r,n}}^{\infty} (1-F_\xi(y))^{1/q}\overline{N}_q dy \\
&\le 2\overline{N}_q \int_{w_{r,n}}^{\infty} \mathbb{E}[\xi(Z_1)^r]^{1/q} y^{-r/q} dy \\
&= 2\overline{N}_q \|\xi\|^{r/q}_{\mathbb{P}^Z,r} \frac{q}{r-q} w_{r,n}^{(q-r)/q} \\
&\le 2\overline{N}_q \|\xi\|_{\mathbb{P}^Z,r} \frac{q}{r-q} \frac{1}{n^{1/q-1/r}} \\
&= \tilde{L}(n,r,q).
\end{aligned}$$

□

Now, Proposition 3.25 is an immediate consequence of Lemma 3.27 together with Proposition 3.28. □

3.8 Additional Details for Distortion Functions

3.8.1 Some Details on Table 3.1

- If $h(p) = (1-(1-p)^k)^{1/k}$, then

$$\lim_{p\downarrow 0} \frac{h(p)}{p^{1/k}} = \lim_{p\downarrow 0} \left(\frac{1-(1-p)^k}{p}\right)^{1/k} = \left(\lim_{p\downarrow 0} \frac{1-(1-p)^k}{p}\right)^{1/k} = k^{1/k}.$$

Hence, $\overline{N}_q < \infty$ for $q \geq k$. Moreover, $\overline{N}_q \leq k^{1/k}$, for $q \geq k$, since $\overline{N}_k = k^{1/k}$, due to the fact that $\frac{1-(1-p)^k}{p}$ is decreasing in p (since it is the slope of the concave function $h(p) = 1 - (1-p)^k$ between $[0, p]$).

- If $h(p) = p^a(1 - \log(p^a))$ for $a \in (0, 1)$, then $\lim_{p \downarrow 0} p^{a-1/q}(1 - \log(p^a)) < \infty$ if and only if $q > 1/a$. To calculate $\overline{N}_q$, we examine the first order condition:

$$\begin{aligned} \frac{d}{dp} p^{a-1/q}(1 - \log(p^a)) &= 0 \\ p^{a-1/q-1}(a - 1/q)(1 - \log(p^a)) - a p^{a-1/q} p^{-a+a-1} &= 0 \\ 1 - \frac{a}{a - 1/q} &= a \log(p) \\ p &= \exp\left(\frac{-1}{a^2 q - a}\right). \end{aligned}$$

  Hence,

$$\begin{aligned} \overline{N}_q &= \exp\left(\frac{-a + 1/q}{a^2 q - a}\right)\left(1 + \frac{1}{aq - 1}\right) \\ &= \exp\left(\frac{-a + 1/q}{a^2 q - a}\right)\left(\frac{aq}{aq - 1}\right). \end{aligned}$$

3.8.2 Extra Details on Kahneman-Tversky's Distortion Function

Recall the Kahneman-Tversky's function

$$h(p) = 1 - \frac{(1-p)^\beta}{\left((1-p)^\beta + p^\beta\right)^{1/\beta}},\ 0 < \beta < 1.$$

Its dual function is given by

$$\bar{h}(p) = \frac{p^\beta}{\left(p^\beta + (1-p)^\beta\right)^{1/\beta}},\ 0 < \beta < 1.$$

To show $\overline{N}_{1/\beta, h_{cv}} < \infty$, it is sufficient to notice that

$$\lim_{p \downarrow 0} \frac{\bar{h}(p)}{p^\beta} = \lim_{p \downarrow 0} \frac{1}{\left(p^\beta + (1-p)^\beta\right)^{1/\beta}} = 1.$$

Now, to show $\overline{N}_{1/\beta,h_{cc}} < \infty$, we need to show $\lim_{p\downarrow 0} h(p)/p^\beta < \infty$. This is equal to

$$\begin{aligned}\lim_{p\downarrow 0} \frac{1-\frac{(1-p)^\beta}{\left((1-p)^\beta+p^\beta\right)^{1/\beta}}}{p^\beta} &= \lim_{p\downarrow 0} \frac{\left((1-p)^\beta+p^\beta\right)^{1/\beta}-(1-p)^\beta}{p^\beta} \cdot \frac{1}{\left((1-p)^\beta+p^\beta\right)^{1/\beta}} \\ &= \lim_{p\downarrow 0} \frac{\left((1-p)^\beta+p^\beta\right)^{1/\beta}-(1-p)^\beta}{p^\beta} \cdot \lim_{p\downarrow 0} \frac{1}{\left((1-p)^\beta+p^\beta\right)^{1/\beta}} \\ &= \lim_{p\downarrow 0} \frac{\left((1-p)^\beta+p^\beta\right)^{1/\beta}-(1-p)^\beta}{p^\beta},\end{aligned}$$

provided the last limit exists. Hence, to examine it, we use L'Hopital's rule and differentiate both the numerator and the denominator. This gives

$$\begin{aligned}&\lim_{p\downarrow 0} \frac{\frac{1}{\beta}\left((1-p)^\beta+p^\beta\right)^{1/\beta-1}\left(\beta p^{\beta-1}-\beta(1-p)^{\beta-1}\right)+\beta(1-p)^{\beta-1}}{\beta p^{\beta-1}} \\ &= \lim_{p\downarrow 0} \frac{1}{\beta}\left((1-p)^\beta+p^\beta\right)^{1/\beta-1}\left(1-(1-p)^{\beta-1}p^{1-\beta}\right)+(1-p)^{\beta-1}p^{1-\beta} \\ &= \frac{1}{\beta}.\end{aligned}$$

Appendix

3.A Kusuoka representation

Let $\rho : \mathcal{X} \to \mathbb{R}$ be the law-invariant convex risk measure on the $\mathbb{R}$-vector space $\mathcal{X}$ as introduced in Section 3.2. Furthermore, let β_ρ be the function on the set $\mathcal{M}\big([0,1)\big)$ of all Borel probability measures on $[0,1)$ which has been introduced by (3.9). We want to verify the Kusuoka representation (3.8) for ρ.

Proposition 3.A.1. *If ρ satisfies the Lebesgue property, then under the assumptions on $\mathcal{X}$ made in Section 3.2, then*

$$\rho(X) = \sup_{\mu\in\mathcal{M}([0,1))} \left(\int_{[0,1)} AV@R_\alpha(X)\ \mu(d\alpha) - \beta_\rho(\mu)\right)$$

holds for every $X \in \mathcal{X}$. Moreover, if ρ is coherent, then $\beta_\rho(\mu) = 0$ for any μ from the effective domain of β_ρ.

Proof. First of all, β_ρ vanishes on its effective domain if ρ is coherent due to positive homogeneity.

Since ρ fulfills the Lebesgue property, its restriction to $L^\infty(\Omega,\mathcal{F},\mathbb{P})$ satisfies the fol-

lowing property

$$\rho(X_k) \nearrow \rho(X) \text{ whenever } X_k \nearrow X \text{ and } X_k, X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P}). \tag{3.33}$$

Under this property, we may draw on Theorem 4.62 from Föllmer and Schied [2016] to conclude

$$\rho(X) = \sup_{\mu \in \mathcal{M}([0,1))} \left(\int_{[0,1)} AV@R_\alpha(X) \; \mu(d\alpha) - \beta_\rho(\mu) \right), \quad \text{for } X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P}). \tag{3.34}$$

Denoting the effective domain of β_ρ by $\mathrm{dom}(\beta_\rho)$, we may introduce via

$$\overline{\rho}(X) = \sup_{\mu \in \mathrm{dom}(\beta_\rho)} \left(\int_{[0,1)} AV@R_\alpha(X) \; \mu(d\alpha) - \beta_\rho(\mu) \right)$$

the mapping $\overline{\rho} : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$. Recall that the mappings from $\mathcal{X}$ are assumed to be $\mathbb{P}$-integrable. We have

$$|AV@R_\alpha(X)| \leq AV@R_\alpha(|X|), \quad \text{for } \alpha \in (0,1) \text{ and } X \in \mathcal{X}, \tag{3.35}$$

because the Average Value at Risk satisfies monotonicity and sublinearity. Furthermore, the Average Value at Risk is continuous w.r.t. the L^1-norm (see Ruszczynski and Shapiro, 2006, Proposition 3.1). This means

$$\lim_{k \to \infty} AV@R_\alpha(|X| \wedge k) = AV@R_\alpha(|X|), \quad \text{for } \alpha \in (0,1), X \in \mathcal{X}. \tag{3.36}$$

Combining (3.35) and (3.36), we may conclude by monotone convergence that

$$\begin{aligned}
&\int_{[0,1)} |AV@R_\alpha(X)| \; \mu(d\alpha) - \beta_\rho(\mu) \\
&= |\mathbb{E}[X]| \; \mu(\{0\}) + \int_{(0,1)} |AV@R_\alpha(X)| \; \mu(d\alpha) - \beta_\rho(\mu) \\
&\leq \mathbb{E}[|X|] \; \mu(\{0\}) + \int_{(0,1)} AV@R_\alpha(|X|) \; \mu(d\alpha) - \beta_\rho(\mu) \\
&= \lim_{k \to \infty} \left(\mathbb{E}[|X| \wedge k]] + \int_{(0,1)} AV@R_\alpha(|X| \wedge k) \; \mu(d\alpha) - \beta_\rho(\mu) \right)
\end{aligned}$$

holds for any $\mu \in \mathrm{dom}(\beta)$ and every $X \in \mathcal{X}$. Moreover, as a law-invariant convex risk measure on a Banach lattice, ρ is norm-continuous (e.g., [Ruszczynski and Shapiro, 2006, Proposition 3.1]), and $\big\| |X| - |X| \wedge k \big\|_{\mathcal{X}} \to 0$ for $k \to \infty$, if $X \in \mathcal{X}$ due to (3.6). Hence,

in view of (3.34), we end up with

$$|\mathbb{E}[X]\ \mu(\{0\})| + \int_{(0,1)} |AV@R_\alpha(X)|\ \mu(d\alpha) - \beta_\rho(\mu) \leq \lim_{k\to\infty} \rho(|X| \wedge k) = \rho(|X|),$$

for $\mu \in \mathrm{dom}(\beta)$ and $X \in \mathcal{X}$. In particular,

$$\overline{\rho}(X) \leq \rho(|X|) < \infty, \quad \text{for } X \in \mathcal{X}.$$

Then $\overline{\rho}$ may be verified easily as a law-invariant convex risk measure so that it is also continuous w.r.t. $\|\cdot\|_{\mathcal{X}}$. Since ρ and $\overline{\rho}$ coincide on $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ by (3.34), and since $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ is a dense subset of $\mathcal{X}$ w.r.t. $\|\cdot\|_{\mathcal{X}}$ by (3.6), the norm continuity of both mappings yields that they also coincide on the entire space $\mathcal{X}$. This completes the proof. □

For any $X \in \mathcal{X}$, we denote its distribution function by F_X. If we restrict ρ to subsets $\{X \in \mathcal{X} \mid F_Y \leq F_X \leq F_{-Y} \text{ pointwise}\}$ for some nonnegative $Y \in \mathcal{X}$, we may bound the penalty function β_ρ in the Kusuoka representation.

Proposition 3.A.2. *Let $Y \in \mathcal{X}$ with $Y \geq 0$ $\mathbb{P}$-a.s.. Then, with the Kusuoka representation from Proposition 3.A.1, we may find some $b \in \mathbb{R}$ such that*

$$\rho(X) = \sup_{\substack{\mu \in \mathrm{dom}(\beta_\rho) \\ \beta_\rho(\mu) \leq b+\delta}} \left(\int_{[0,1)} AV@R_\alpha(X)\ \mu(d\alpha) - \beta_\rho(\mu) \right)$$

holds for any $\delta > 0$, and arbitrary $X \in \mathcal{X}$ satisfying $F_Y \leq F_X \leq F_{-Y}$ pointwise. Here $\mathrm{dom}(\beta)$ denotes the effective domain of β. Sufficient choices are $b = \rho(2\ Y) - 2\ \rho(-Y)$ or $b = 2\rho(2Y) - 3\rho(0)$.

Proof. Fix $\delta > 0$ and $X \in \mathcal{X}$ with $F_Y \leq F_X \leq F_{-Y}$ pointwise. This implies $F_{2Y} \leq F_{2X}$. We shall denote the expectation by $AV@R_0$. In view of the Kusuoka representation derived in Proposition 3.A.1, we may restrict the supremum to all $\mu \in \mathcal{M}$ such that

$$\rho(X) - \delta/2 < \int_{[0,1)} AV@R_\alpha(X)\ \mu(d\alpha) - \beta_\rho(\mu).$$

Then, by positive homogeneity of the expectation and Average Value at Risk along with the Kusuoka representation from Proposition 3.A,

$$\begin{aligned}
&\beta_\rho(\mu) - \delta/2 \\
&< \left[\int_{[0,1)} AV@R_\alpha(2X)\ \mu(d\alpha) - \beta_\rho(\mu) \right] - \rho(X) - \left[\int_{[0,1)} AV@R_\alpha(X)\ \mu(d\alpha) - \beta_\rho(\mu) \right] \\
&\leq \rho(2X) - \rho(X) - \rho(X) + \delta/2 = \rho(2X) - 2\rho(X) + \delta/2.
\end{aligned}$$

Since the expectation as well as the Average Value at Risk are monotone, ρ satisfies this property as well by the Kusuoka representation. This means, in particular, that $\rho(-Y) \leq \rho(X)$ due to $F_X \leq F_{-Y}$ pointwise, and $AV@R_\alpha(2X) \leq AV@R_\alpha(2Y)$ for every $\alpha \in [0,1)$ because $F_Y \leq F_X$ pointwise and thus $F_{2Y} \leq F_{2X}$ pointwise. Hence,

$$\beta_\rho(\mu) \leq \rho(2Y) - 2\rho(-Y) + \delta,$$

and the first statement follows with $b \triangleq \rho(2Y) - 2\rho(-Y)$. Moreover, convexity of ρ implies $-\rho(-Y) \leq \rho(Y) - 2\rho(0)$. In addition, by the Kusuoka representation $\rho(0) + \beta_\rho(\mu) \geq 0$ holds for $\mu \in \operatorname{dom}(\beta_\rho)$ so that by positive homogeneity of the Average Value at Risk we may conclude

$$\begin{aligned} 2\rho(Y) &= \sup_{\mu \in \operatorname{dom}(\beta_\rho)} \left(\int_{[0,1)} AV@R_\alpha(2Y)\ \mu(d\alpha) - 2\beta_\rho(\mu) \right) \\ &\leq \sup_{\mu \in \operatorname{dom}(\beta_\rho)} \left(\int_{[0,1)} AV@R_\alpha(2Y)\ \mu(d\alpha) - \beta_\rho(\mu) \right) + \rho(0) = \rho(2Y) + \rho(0). \end{aligned}$$

Thus, we end up with $b \leq 2\rho(2Y) - 3\rho(0)$, which completes the proof. □

As a consequence of the representation result in Theorem 3.A.1, we may verify the following type of Fatou property.

Corollary 3.A.3. *Let ρ fulfill the Lebesgue property, and let $\{X_k, X, Y \mid k \in \mathbb{N}\} \subset \mathcal{X}$ such that $\liminf_{k\to\infty} X_k \geq X$ $\mathbb{P}$-a.s. and $\sup_{k\in\mathbb{N}} |X_k| \leq Y$ $\mathbb{P}$-a.s. Then,*

$$\liminf_{k\to\infty} \rho(X_k) \geq \rho(X).$$

Proof. By assumption $\overline{X}_k \triangleq \inf_{l\in\mathbb{N}, l\geq k} X_l$ and $\overline{X} \triangleq \sup_{k\in\mathbb{N}} \overline{X}_k$ are well-defined random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ satisfying $\overline{X}_k \leq Y$ $\mathbb{P}$-a.s. for $k \in \mathbb{N}$ and $\overline{X} \leq Y$ $\mathbb{P}$-a.s. These random variables belong to $\mathcal{X}$ because $\mathcal{X}$ is solid. Furthermore, we may find by Proposition 3.A.2 a set $\mathcal{M}$ of Borel probability measures on $[0,1)$ such that β_ρ is bounded on $\mathcal{M}$, and

$$\rho(\overline{Y}) = \sup_{\mu\in\mathcal{M}} \left(\int_{[0,1)} AV@R_\alpha(\overline{Y})\ \mu(d\alpha) - \beta_\rho(\mu) \right), \tag{3.37}$$

for $\overline{Y} \in \{X_k, \overline{X}, X, \overline{X}, Y \mid k \in \mathbb{N}\}$, where $AV@R_0$ stands for the expectation. By the dominated convergence theorem the sequence $(\overline{X}_k)_{k\in\mathbb{N}}$ converges to $\overline{X}$ w.r.t. the L^1-norm. Hence, by L^1-norm continuity of the expectation and the Average Value at Risk, we obtain $AV@R_\alpha(\overline{X}_k) \to AV@R_\alpha(\overline{X})$. Moreover, $X_k \geq \overline{X}_k$ $\mathbb{P}$-a.s. for $k \in \mathbb{N}$, and $\overline{X} \geq X$ $\mathbb{P}$-a.s. Then, by monotonicity of the expectation and the Average Value at Risk

$$\liminf_{k\to\infty} AV@R_\alpha(X_k) \geq \liminf_{k\to\infty} AV@R_\alpha(\overline{X}_k) \geq AV@R_\alpha(\overline{X}) \geq AV@R_\alpha(X),$$

$\alpha \in [0,1)$. Finally, observe $AV@R_\alpha(X_k) \geq AV@R_\alpha(-Y)$ for $\alpha \in [0,1)$ due to the monotonicity of expectation and Average Value at Risk again. Then the application of the Fatou lemma yields

$$\liminf_{k\to\infty} \int_{[0,1)} AV@R_\alpha(X_k)\ \mu(d\alpha) \geq \int_{[0,1)} AV@R_\alpha(X)\ \mu(d\alpha).$$

Now, the statement of Corollary 3.A.3 may be derived easily from the representation 3.37. □

3.B A concentration inequality

Let $(Z_j)_{j\in\mathbb{N}}$ be a sequence of independent, identically distributed d-dimensional random vectors on $(\Omega, \mathcal{F}, \mathbb{P})$. The main tool for the proof of Theorem 3.22 is the following concentration inequality.

Theorem 3.B.4. *Let $\mathbb{F}$ be some at most countable set of real-valued Borel measurable mappings on $\mathbb{R}^d$ that are uniformly bounded by some positive constant b. Then for every $n \in \mathbb{N}$ and any $\varepsilon > 0$, we have*

$$\mathbb{P}\big(\big\{S_n \geq \mathbb{E}[S_n] + \varepsilon\big\}\big) \leq \exp\left(\frac{-\varepsilon^2}{2nb^2}\right),$$

where $S_n \triangleq \sup_{f\in\mathbb{F}} \big|\sum_{j=1}^n f(Z_j) - n\ \mathbb{E}\big[f(Z_1)\big]\big|$.

Proof. Since the mappings from $\mathbb{F}$ are uniformly bounded, and since $\mathbb{F}$ is at most countable, the function

$$h : \mathbb{R}^{dn} \to \mathbb{R},\ (z_1, \ldots, z_n) \mapsto \sup_{f\in\mathbb{F}} \left|\sum_{j=1}^n f(z_j) - n\ \mathbb{E}\big[f(Z_1)\big]\right|$$

is Borel measurable and bounded. Furthermore, it satisfies for $z_1, \ldots, z_d, z \in \mathbb{R}^d$ and any $i \in \{1, \ldots, n\}$ the inequality

$$|h(z_1, \ldots, z_n) - h(z_1^i, \ldots, z_n^i)| \leq \sup_{f\in\mathbb{F}} \big|f(z_i) - f(z_i^i)\big| \leq 2b,$$

where $z_i^i \triangleq z$ and $z_j^i \triangleq z_j$ if $j \neq i$. Hence, we may conclude immediately the statement from McDiarmid's bounded differences inequality (see e.g., Gine and Nickl, 2016, Theorem 3.3.14). □

Chapter 4

A New Distributionally Robust Optimization Model Based on Maxiance Regularization

4.1 Introduction

A large part of the data-driven optimization literature has been concentrated on finding optimal solutions for the following problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})], \tag{4.1}$$

where $l(\mathbf{x},\boldsymbol{\xi})$ is some real-valued loss function that evaluates a decision variable $\mathbf{x}$ contained in a compact convex set $\mathcal{X}\subset\mathbb{R}^d$ and an uncertain parameter $\boldsymbol{\xi}\in\mathbb{R}^k$ following the true, but often unknown, distribution $\mathbb{P}_0$. Problem (4.1) encompasses a broad class of decision problems in operations research, statistics and economics, with examples including regression/classification problems, portfolio optimization and inventory management problems.

A very direct approach to solve (4.1) is to replace the true expectation with its sample average approximation (SAA), formulated as

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x},\boldsymbol{\xi})], \tag{4.2}$$

where $\hat{\mathbb{P}}_n$ is the empirical distribution constructed from an i.i.d. sample of $\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n$. However, SAA is known to be sensitive to overfitting, see for example Smith and Winkler [2006] that discusses the shortcomings of SAA in decision analysis. To address this, various distributionally robust optimization (DRO) models have been proposed as an alternative for SAA, where the expected loss value is minimized with respect to a set of distributions. One popular example is DRO based on ϕ-divergence ambiguity sets, which has emerged in

many applications, see e.g., Duchi and Namkoong [2019], Duchi et al. [2021], Lam [2019], Ben-Tal et al. [2013], Postek et al. [2016]. It is formulated as the following min-max optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbb{Q}:I_\phi(\mathbb{Q},\hat{\mathbb{P}}_n)\leq r} \mathbb{E}_{\mathbb{Q}}[l(\mathbf{x},\boldsymbol{\xi})], \tag{4.3}$$

where the ϕ-divergence $I_\phi(\mathbb{Q},\mathbb{P}) = \int_\Omega \phi\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\right)\mathrm{d}\mathbb{P}$, for a convex function $\phi : [0,\infty) \to [0,\infty)$ with $\phi(1) = 0$, serves as a variational measure between two probability measures $\mathbb{Q}$ and $\mathbb{P}$.

From a decision-theoretic perspective, the optimization problem (4.3) can be motivated from variational divergence preferences (Maccheroni et al. 2006). Another main reason that makes ϕ-divergence DRO (4.3) an attractive model is its equivalence to mean-variance regularization, when $r \to 0$. This has been recognized in Gotoh et al. [2018, 2021], Duchi and Namkoong [2019], Duchi et al. [2021], Ben-Tal and Teboulle [2007]. In particular, (4.3) provides a tractable methodology for minimizing both the bias and variance. When there are near-optimal solutions of (4.1) that have loss distributions with much larger variance than that of the true optimum, ϕ-divergence DRO is able to exclude those solutions, whereas SAA is unable to make such a distinction.

However, the use of the variance as a measure of variability has important limitations. For example, it is only justified when the distribution is symmetric. Otherwise, a large variance does not necessarily imply a high degree of variability among a sample of realized values (for instance, consider the sample $(1, 1, 1, 1, 6)$, which has a large variance, but not necessarily much variability). By definition, variance is the squared difference between a realized value and its mean. Since this measures the variability with respect to one reference point, it makes variance sensitive to outliers, which is one of the main causes for overfitting. By contrast, the Gini's mean difference (GMD), introduced by Corrado Gini in 1912, which is a well-known measure among economists who study income inequality (see e.g., Gini 1912, 1921), is based on relative differences, and defined as:

$$\frac{1}{2}\mathbb{E}|X^{(1)} - X^{(2)}|, \tag{4.4}$$

where, for a random variable X, $X^{(1)}, X^{(2)}$ denote two i.i.d. copies. The advantages of GMD as a substitute for variance have been demonstrated in many examples in statistics and economics, such as regression models and portfolio theory. We refer the readers to notable works by Yitzhaki [1982], Shalit and Yitzhaki [1984], Olkin and Yitzhaki [1992], and the textbook by Yitzhaki and Schechtman [2012]. In decision theory and the economics of risk, the GMD, which is referred to as *maxiance* in Eeckhoudt and Laeven [2021], has emerged as the key ingredient for constructing a local index of absolute risk aversion in non-expected utility theory,[1] similar to the appearance of variance in expected

[1]See Yaari [1987], Quiggin [1982], Schmeidler [1986, 1989] for more background on non-expected utility

utility theory. To emphasize the dual relationship between variance and GMD that is also highlighted in this paper, we will henceforth adopt the maxiance terminology as suggested by Eeckhoudt and Laeven [2021].

When outliers are present in the data, using a variance regularization model such as ϕ-divergence DRO to address overfitting in SAA can, in fact, produce counter-effective results. To illustrate this, consider a simple decision problem that only involves choosing between a linear combination of two loss distributions X and Y according to their expected loss. This problem is formulated as (4.1), where $\boldsymbol{\xi} = (X, Y)$, $\mathcal{X} = [0, 1]$ and $l(x, \boldsymbol{\xi}) = xX + (1 - x)Y$. The two loss distributions are defined as $X = \text{Unif}[5, 7] + 25\epsilon$, which is a sum of a uniform distribution on $[5, 7]$, with a Bernoulli variable ϵ that takes the values 1 with probability 0.05 and zero otherwise (ϵ may be interpreted as some exogenous noise, such as measurement errors). On the other hand, $Y = \text{Unif}[1.85, 12.85]$. By construction, $\mathbb{E}[X] = 7.25 < 7.35 = \mathbb{E}[Y]$. Therefore, X is the optimal choice with the minimal expected loss (corresponding to the optimal solution $x = 1$ of (4.1)). Suppose we do not know the true distributions but only observe 1,000 samples of both X and Y, as represented in Figure 4.1. It can be observed that, except for the outliers, the samples of Y exhibit a larger variability than X. Since their true expected values are also very similar, the empirical mean of Y can sometimes be less than that of X due to sampling errors, leading to the wrong conclusion that Y is the optimal solution for (4.1) (for example, in the particular realizations given in Figure 4.1, we have the sample mean $\hat{\mathbb{E}}[X] = 7.47 > 7.45 = \hat{\mathbb{E}}[Y]$). If one uses ϕ-divergence DRO to address overfitting, then the bias is even more towards Y, since the empirical standard deviation estimated from the 1,000 data in Figure 4.1 yields $\hat{\sigma}_X = 5.96 > 3.2 = \hat{\sigma}_Y$, due to the presence of outliers. On the other hand, the maxiance is less affected by the outliers of X: it outputs an estimate of 1.70 for X and a value of 1.85 for Y. Therefore, in this situation, we see that the maxiance provides a more appropriate, outlier-robust representation of variability, and is a more reasonable choice than the variance.

Motivated by this example, we aim to construct a new DRO model, where not the variance but the maxiance is regularized. It turns out, as we formally establish in the paper, that such construction is completely dual to the standard DRO procedure, in the sense that it considers deviations in a different dimension: while the traditional DRO such as (4.3) makes a distributional shift in the probability plane, our new DRO does that in the outcome plane. To emphasize this duality, we will call our new DRO model *dual* DRO, and refer to the ϕ-divergence DRO in (4.3) as *primal* DRO.

More precisely, given a positive loss random variable $X = [x_1; p_1, x_2; p_2, \ldots, x_n; p_n]$ with outcomes x_i that take probabilities p_i, we define the dual DRO by minimizing the expected loss, under the shifted outcome values that are penalized from the nominal

theory.

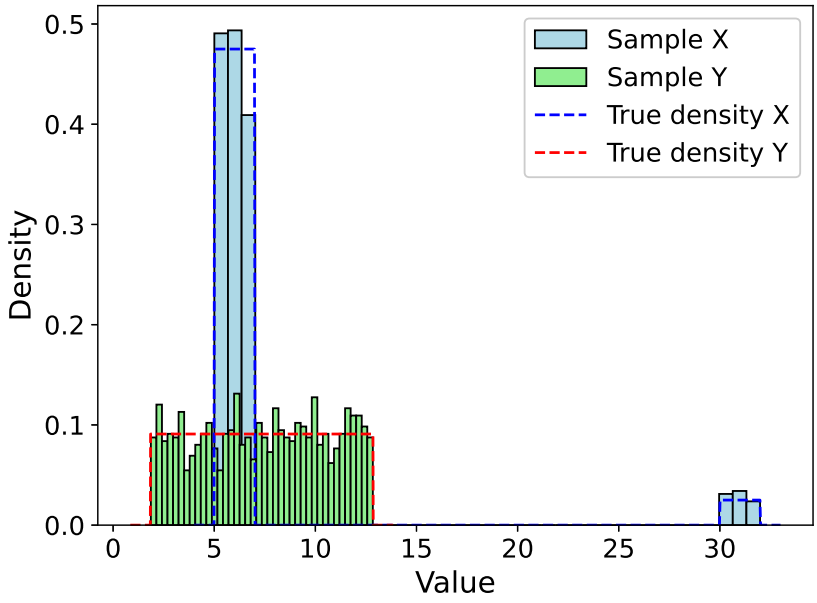


Figure 4.1: This figure displays 1,000 samples drawn for two distributions: $X = \text{Unif}[5,7] + 25\epsilon$ and $Y = \text{Unif}[1.85, 12.85]$, where $\text{Unif}[a,b]$ is the uniform distribution on $[a,b]$ and $\epsilon \sim \text{Ber}(0.05)$ is a Bernoulli random variable with success probability 0.05.

outcomes $(x_1, \ldots, x_n)$, measured using a ϕ-divergence with a penalization constant $\delta > 0$:

$$\rho^{\phi}_{\delta,\mathbf{p}}(X) = \inf_{\Delta\mathbf{y}\in\mathbb{R}^n_+} \sum_{i=1}^{n} \overline{F}_i \Delta y_i + \frac{1}{\delta} I_\phi(\Delta\mathbf{y}, \Delta\mathbf{x}). \tag{4.5}$$

Here, we define $\Delta\mathbf{x} = (\Delta x_1, \ldots, \Delta x_n)$ with $\Delta x_i = x_i - x_{i-1}$ for $i = 2, \ldots, n$ and $\Delta x_1 = x_1$, and similarly for $\Delta\mathbf{y}$. We assume that the outcomes are ordered: $x_1 > 0, \Delta x_i \geq 0$, for all $i \geq 2$. We let $\overline{F}_i = \sum_{j=i}^{n} p_j$ for $i = 1, \ldots, n$ denote the decumulative probabilities, so that for any $\Delta\mathbf{y} \in \mathbb{R}^+_n = \{\mathbf{a} \in \mathbb{R}^n : a_i \geq 0\}$, we have $\sum_{i=1}^{n} \overline{F}_i \Delta y_i = \sum_{i=1}^{n} p_i y_i$. Moreover, we define the divergence $I_\phi(\Delta\mathbf{y}, \Delta\mathbf{x}) = \sum_{i=1}^{n} \Delta x_i \phi\left(\Delta y_i / \Delta x_i\right)$. Let us comment more on (4.5). We note that the dual DRO is designed to minimize both the expected loss and a measure of relative differences. If X admits a distribution that has large relative differences, then $I_\phi(\Delta\mathbf{y}, \Delta\mathbf{x})$ is a large penalization term since it is a sum of the variations Δx_i with non-negative coefficients $\phi(\Delta y_i / \Delta x_i) \geq 0$. Therefore, when solving a dual DRO optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \ \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})), \tag{4.6}$$

where we minimize the dual DRO risk measure of the loss function under the empirical distribution $\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \iota_{\boldsymbol{\xi}_i}$ constructed from i.i.d. data $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$ (where $\iota_{\boldsymbol{\xi}_i}$ denotes the Dirac delta distribution), we are effectively searching for a decision $\mathbf{x}$ that minimizes the expected loss and relative differences of its distribution $l(\mathbf{x}, \boldsymbol{\xi}) \sim \hat{\mathbb{P}}_n$.

Our main contributions may be summarized as follows:

- We prove the asymptotic equivalence between dual DRO (4.5) and mean-maxiance regularization, as $\delta \to 0$. This equivalence is also shown to hold uniformly across all decision variables, for the optimization problem (4.6). Furthermore, we extend this asymptotic equivalence to a primal-dual DRO model, where we consider both distributional and outcomes shifts penalized by two ϕ-divergences. This leads to a more general mean-maxiance-variance regularization model, encompassing mean-variance regulatization and mean-maxiance regularization as special cases.

- We show that the minimization problem (4.6) can be computed by solving a convex optimization problem that provides tight lower and upper bounds on (4.6). We then provide numerical examples where the solution of a dual DRO optimization problem can be more attractive than the ones obtained from the primal DRO and the empirical risk minimization, in terms of finding an optimal solution for the true minimization problem $\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}, \boldsymbol{\xi})]$.

- Finally, we establish the asymptotic convergence rate of the dual DRO minimization problem (4.6) as an estimator and a lower bound on the true optimal value $\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}, \boldsymbol{\xi})]$. We show that a similar dual DRO upper bound can also be obtained.

Finally, we give a brief review of other related literature. Besides the equivalence between ϕ-divergence DRO and mean-variance optimization that are established in Gotoh et al. [2018], Duchi and Namkoong [2019], Duchi et al. [2021], regularization effects of other DRO models have also been established in for example Gotoh et al. [2020]. In particular, it is known that Wasserstein DRO regularizes the gradient norm of the decision function (see, e.g., Gao et al. 2022). This has led to the application of Wasserstein DRO in machine learning, such as improving the robustness of regression models against outliers, see e.g., Chen and Paschalidis [2018], and other learning tasks, see Kuhn et al. [2019]. Recent papers by Bartl and Mendelson [2022] and van Parys and Zwart [2025] have also examined robust estimation procedures for the optimal expected value problem (4.1) in the setting of heavy-tailed distributions. Bartl and Mendelson [2022] proposed a novel procedure for estimating the solution and objective value of (4.1) that attains the optimal Gaussian rate in finite samples. However, their desired statistical property comes at the expense of computational tractability. Van Parys and Zwart [2025] examined the rate at which the probability of over- and underestimating the true expected value decays for various data-driven formulations. In particular, they showed that the mean-variance model (with a $O(1/\sqrt{n})$ penalization coefficient for the standard deviation) has a large overestimation probability for the expected value when the distributions are heavy-tailed. On the other hand, the dual DRO model based on mean-maxiance regularization that we propose in this paper is computationally tractable, and can serve as an attractive new data-driven

formulation for heavy-tailed distributions, as our numerical simulations suggest that the maxiance is (much) more robust against outliers than the variance.

The remaining parts of this paper are organized as follows: we first introduce the maxiance and its relation to distortion risk measures in Section 4.2. We then prove our main results on asymptotic equivalence in Section 4.3. In Section 4.4, we study the minimization problem (4.6), and its statistical properties are established in Section 4.5. Finally, a concluding remark is provided in Section 4.6.

4.2 Preliminaries

4.2.1 ϕ-Divergences

We start by recalling the definition of a Csiszar ϕ-divergence. Let $\phi : \mathbb{R} \to [0, \infty)$ be a convex function such that $\phi(1) = 0$ and $\mathrm{dom}(\phi) \subset [0, \infty)$. Then, the ϕ-divergence between any two non-negative vectors $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n_+$ is defined as:

$$I_\phi(\mathbf{q}, \mathbf{p}) \triangleq \sum_{i=1}^{n} p_i \phi\left(\frac{q_i}{p_i}\right).$$

We note that $I_\phi(\mathbf{q}, \mathbf{p})$ is jointly convex in $\mathbf{q}$ and $\mathbf{p}$, and that $I_\phi(\mathbf{q}, \mathbf{p}) \geq 0$, $I_\phi(\mathbf{p}, \mathbf{p}) = 0$. Furthermore, we adopt the convention $0\phi\,(0/0) = 0$ and $0\phi\,(t/0) = \lim_{z\to\infty} \phi(z)/z$. Throughout this paper, we assume $\lim_{z\to\infty} \phi(z)/z = \infty$ and that ϕ is sufficiently smooth at 1 (i.e., infinitely differentiable at 1). We note that these assumptions are not restrictive, since they hold for many canonical examples of ϕ-divergences (see Section 1.2 of Pardo 2006).

For any convex function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $f^*(\mathbf{y}) = \sup_{\mathbf{x}\in\mathrm{dom}(f)} \mathbf{y}^T\mathbf{x} - f(\mathbf{y})$ as the convex conjugate of f. Furthermore, for any concave function $g : \mathbb{R}^d \to \mathbb{R}$, we denote by $g_*(\mathbf{y}) = \inf_{\mathbf{x}\in\mathrm{dom}(g)} \mathbf{y}^T\mathbf{x} - g(\mathbf{x})$ the concave conjugate of g. We note that a convex conjugate is always convex, and a concave conjugate is always concave. In particular, we have that $\phi^*(y) = \sup_{x\geq 0} yx - \phi(x)$ and $(-\phi)_*(y) = \inf_{x\geq 0} yx + \phi(x)$. By definition, the following relation holds:

$$(-\phi)_*(y) = -\phi^*(-y), \ \forall y \in \mathbb{R}. \tag{4.7}$$

In the following proposition, we summarize some properties of ϕ^* that are frequently used in this paper.

Proposition 4.1. *Let ϕ be a Csiszar ϕ-divergence function. Then*

1. *ϕ^* is non-decreasing.*
2. *$\phi^*(0) = 0$ and $\phi^*(y) \geq y$, for all $y \in \mathbb{R}$.*

If ϕ is also three times continuously differentiable at 1 with $\phi''(1) > 0$, then ϕ^* is twice continuously differentiable at 0, with $(\phi^*)'(0) = 1$, $(\phi^*)''(0) = 1/\phi''(1)$, and $(\phi^*)'''(0) = -\phi'''(1)/(\phi''(1))^3$.

4.2.2 Maxiance and Distortion Risk Measures

For a random variable X, the maxiance is defined as

$$\overline{\mathtt{m}}_2(X) \triangleq \mathbb{E}[\max\{X^{(1)}, X^{(2)}\}] - \mathbb{E}[X], \tag{4.8}$$

where $X^{(1)}, X^{(2)}$ are i.i.d. draws of X. We note that the maxiance can also be expressed as $\overline{\mathtt{m}}_2(X) = \mathbb{E}[X] - \mathbb{E}[\min\{X^{(1)}, X^{(2)}\}]$.[2] Throughout this paper, we also refer to the quantity $\mathbb{E}[\min\{X^{(1)}, X^{(2)}\}]$ as the second *dual* moment (terminology adopted from Eeckhoudt et al. 2020). As mentioned previously, the maxiance is also equal to the Gini's mean difference that measures the average of the absolute difference between two i.i.d. random variables. Indeed, one has $\overline{\mathtt{m}}_2(X) = \mathbb{E}|X^{(1)} - X^{(2)}|/2$; see Section 5 of Eeckhoudt and Laeven [2021].

An important tool that allows us to study the dual DRO model (4.5) is by relating it to the class of distortion risk measures. This class of risk measures is characterized by a distortion function $h : [0,1] \to [0,1]$ that is non-decreasing and satisfies $h(0) = 0$, $h(1) = 1$ (see Section 4.6 of Föllmer and Schied 2016 for more detail). Given a distortion function h, the corresponding distortion risk measure of a loss variable X is defined as the Choquet integral

$$\rho_h(X) \triangleq \int_0^\infty h\left(\mathbb{P}[X > t]\right) \mathrm{d}t + \int_{-\infty}^0 \left(h\left(\mathbb{P}[X > t]\right) - 1\right) \mathrm{d}t. \tag{4.9}$$

If $X = [x_1; p_1, x_2; p_2, \ldots, x_n; p_n]$ is a discrete random variable with realizations x_i's that take probabilities p_i's, such that $x_1 \leq x_2 \leq \ldots \leq x_n$, then $\rho_h(X)$ becomes a rank-dependent sum

$$\rho_h(X) = \sum_{i=1}^n h\left(\overline{F}_i\right)(x_i - x_{i-1}), \tag{4.10}$$

where $\overline{F}_i = \sum_{j=i}^n p_j$ for $i = 1, \ldots, n$, are the decumulative probabilities and we denote $x_0 \triangleq 0$. The class of distortion risk measures contains many interesting examples. This includes the Conditional Value-at-Risk/Expected Shortfall (when $h(p) = \min\{p/\alpha, 1\}$), and the Value-at-Risk (when $h(p) = \mathbf{1}_{[\alpha,1]}(p)$). If $h(p) = 2p - p^2$, then the corresponding distortion risk measure is also equal to a mean plus maxiance evaluation.

Further Notations. Throughout this paper, we adopt the landau O notation, where $a_n = O(b_n)$ if $\limsup_{n\to\infty} |a_n/b_n| < \infty$, and $a_n = o(b_n)$ if $\lim_{n\to\infty} |a_n/b_n| = 0$. Similar

[2] This can be verified by noting that $\mathbb{E}[\max\{X^{(1)}, X^{(2)}\}] + \mathbb{E}[\min\{X^{(1)}, X^{(2)}\}] = 2\mathbb{E}[X]$.

notations for stochastic convergence are adopted from van der Vaart [1998]. We also write $X_n \overset{P^*}{\to} X$ for convergence in outer probability, as defined in van der Vaart and Wellner [2023] (page 52). Moreover, for any integer n, we let $[n] := \{1, \ldots, n\}$.

4.3 Asymptotic Equivalence of Dual DRO and Mean-Maxiance Optimization

In this section, we establish the asymptotic equivalence between dual DRO and mean-maxiance optimization. We study the dual DRO model in a discrete probability setting, since data-driven optimization often relies on the empirical distribution. Let $(\Omega, \mathcal{F})$ be a sample space and $X : \Omega \to \mathbb{R}$ a positive random variable that is distributed on n support points $0 < x_1 \leq x_2 \leq \ldots \leq x_n$ with probabilities $p_1, p_2, \ldots, p_n$. We assume that $\min_i p_i > 0$ (otherwise, discard those x_i's where $p_i = 0$).

Let ϕ be a divergence function. We recall the definition of the dual DRO model, where instead of with respect to the probabilities, we optimize with respect to the payoffs dimension $\Delta\mathbf{y}$, controlled by a ϕ-divergence penalization term:

$$\rho^{\phi}_{\delta,\mathbf{p}}(X) = \inf_{\Delta\mathbf{y}\in\mathbb{R}^n_+} \sum_{i=1}^{n} \overline{F}_i \Delta y_i + \frac{1}{\delta} I_\phi(\Delta\mathbf{y}, \Delta\mathbf{x}), \quad \delta > 0, \tag{4.11}$$

where $I_\phi(\Delta\mathbf{y}, \Delta\mathbf{x}) = \sum_{i=1}^{n} \Delta x_i \phi\left(\Delta y_i / \Delta x_i\right)$.

Our first main result is the asymptotic equivalence between the dual DRO model and mean-maxiance regularization as $\delta \to 0$, which is stated in the next theorem.

Theorem 4.2. *Let ϕ be a ϕ-divergence function that is twice continuously differentiable at 1. Then, we have that*

$$\rho^{\phi}_{\delta,\mathbf{p}}(X) = \left(1 - \frac{\delta}{2\phi''(1)}\right) \mathbb{E}_{\mathbf{p}}[X] + \frac{\delta}{2\phi''(1)} \overline{\mathbb{m}}_{2,\mathbf{p}}(X) + o(\delta). \tag{4.12}$$

We note that the dual DRO in (4.11) is a minimization problem that can be solved explicitly, and is (modulo a positive constant) equal to a distortion risk measure. This result dates back to Ben-Tal et al. [1991], which shows that many classes of risk measures admit a robust representation with a ϕ-divergence penalization. As shown in the proof of Theorem 4.2, the dual DRO model in (4.11) is proportional to the following distortion risk measure:

$$\begin{aligned} \rho^{\phi}_{\delta,\mathbf{p}}(X) &= \frac{(-\phi)_*(\delta)}{\delta} \mathcal{R}^{\phi}_{\delta,\mathbf{p}}(X) \\ &\triangleq \frac{(-\phi)_*(\delta)}{\delta} \sum_{i=1}^{n} h_\delta(\overline{F}_i) \Delta x_i, \end{aligned} \tag{4.13}$$

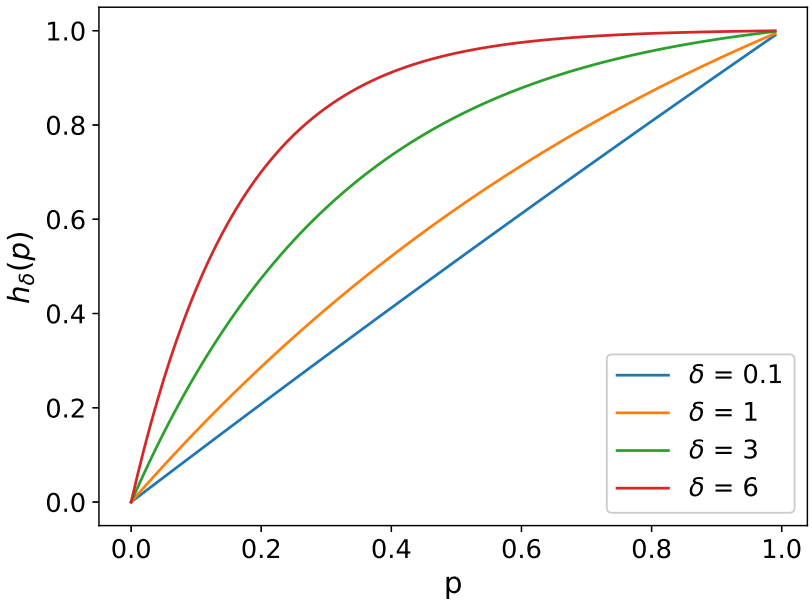


Figure 4.2: This figure displays the shape of the distortion function $h_\delta(p) = \frac{(-\phi)_*(\delta p)}{(-\phi)_*(\delta)}$ with varying δ, where ϕ is the Kullback-Leibler divergence given by $\phi(t) = t \log t - t + 1$.

where the distortion function $h_\delta : [0,1] \to [0,1]$ is given by $h_\delta(p) = (-\phi)_*(\delta p)/(-\phi)_*(\delta)$. This gives the dual DRO an interpretation from decision theory, which implies that decision-making with the dual DRO model is within the axioms of Yaari's dual theory (see Yaari 1987). The parameter δ, which controls the degree of penalization in the dual DRO model (4.11), shapes the concavity of the distortion function h_δ. As shown in Figure 4.2, a larger δ gives more concavity to h_δ, which puts larger probability weights on the worst outcomes of the loss random variable X.

As we can see in (4.12), the asymptotic analysis of $\rho^\phi_{\delta,\mathbf{p}}(X)$ shows that there is still a factor $-\delta/(2\phi''(1))$ in front of the expectation, since $\rho^\phi_{\delta,\mathbf{p}}(X)$ is by definition a lower bound on the expectation. This is not entirely analogous to the primal ϕ-divergence DRO case, where only the variance is penalized by a "first-order" factor. It turns out that the expansion of the distortion risk measure $\mathcal{R}^\phi_{\delta,\mathbf{p}}(X)$ excludes the delta factor at the expectation, and the maxiance emerges as the only first-order penalized factor.

Proposition 4.3. *Let ϕ be a ϕ-divergence that is twice continuously differentiable at 1. We have that*

$$\mathcal{R}^\phi_{\delta,\mathbf{p}}(X) = \mathbb{E}_{\mathbf{p}}[X] + \frac{\delta}{2\phi''(1)} \overline{\mathbb{m}}_{2,\mathbf{p}}(X) + o(\delta).$$

As is apparent from (4.13), $\rho^\phi_{\delta,\mathbf{p}}(X)$ and $\mathcal{R}^\phi_{\delta,\mathbf{p}}(X)$ differ only by a multiplicative factor. Therefore, minimizing $\rho^\phi_{\delta,\mathbf{p}}$ as in (4.6) yields the same optimal solution as minimizing $\mathcal{R}^\phi_{\delta,\mathbf{p}}$. The only difference is that in the former case, one obtains a lower bound on the expected value, while the latter yields an upper bound. Finally, we can extend the primal DRO and dual DRO to a *primal-dual* DRO model, where we consider shifts in both the outcome

space and the probability space, measured by two divergences ϕ and ψ. More precisely, we define

$$\rho_{\delta_1,\delta_2,xp}(X) = \sup_{\mathbf{q}\in\mathbb{P}^n} \inf_{\Delta\mathbf{y}\in\mathbb{R}^n_+} \sum_{i=1}^n q_i y_i + \frac{1}{\delta_1} I_\phi(\Delta\mathbf{y}, \Delta\mathbf{x}) - \frac{1}{\delta_2} I_\psi(\mathbf{q}, \mathbf{p}). \tag{4.14}$$

The next result states the asymptotic equivalence of a primal-dual DRO model to mean-variance-maxiance regularization.

Theorem 4.4. *Let ϕ, ψ be two Csiszar ϕ-divergence functions that are both twice continuously differentiable with $\phi''(1) > 0$ and $\psi''(1) > 0$. Assume $p_i > 0$ for all $i = 1, \ldots, n$. Then, we have that*

$$\begin{aligned}\rho_{\delta_1,\delta_2,xp}(X) = &\left(1 - \frac{\delta_1}{2\phi''(1)}\right)\mathbb{E}_{\mathbf{p}}[X] + \frac{\delta_2}{2\psi''(1)}\mathrm{Var}_{\mathbf{p}}(X) \\ &+ \frac{\delta_1}{2\phi''(1)}\overline{\mathbb{m}}_{2,\mathbf{p}}(X) + o(\delta_1) + o(\delta_2) + O(\delta_1\delta_2 n).\end{aligned} \tag{4.15}$$

4.4 Optimization of Dual DRO Models

In this section, we first discuss how to solve a dual DRO minimization problem in a data-driven setting, Then we numerically investigate the performance of dual DRO.

4.4.1 Solving Dual DRO

Let $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ an i.i.d. sample of data of $\boldsymbol{\xi}$. Then, we consider the following dual DRO minimization problem with respect to the empirical distribution $\hat{\mathbb{P}}_n = \sum_{i=1}^n \iota_{\boldsymbol{\xi}_i}/n$:

$$\min_{\mathbf{x}\in\mathcal{X}} \ \rho^\phi_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})), \tag{4.16}$$

where we assume that the positive loss function $l(\mathbf{x}, \boldsymbol{\xi})$ is convex and continuous in $\mathbf{x}$, for each random vector $\boldsymbol{\xi}$ in $\mathbb{R}^k$. We recall that the dual DRO evaluation $\rho^\phi_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi}))$ is a rank-dependent evaluation for each decision $\mathbf{x} \in \mathcal{X}$. Indeed, for any fixed $\mathbf{x} \in \mathcal{X}$, we must first rank the outcomes by indices $i(\mathbf{x})$ such that $l(\mathbf{x}, \boldsymbol{\xi}_{1(\mathbf{x})}) \leq \ldots \leq l(\mathbf{x}, \boldsymbol{\xi}_{n(\mathbf{x})})$. Then, according to (4.11), we have that

$$\rho^\phi_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})) = \inf_{\Delta\mathbf{y}\in\mathbb{R}^n_+} \sum_{i=1}^n \frac{y_{i(\mathbf{x})}}{n} + \frac{1}{\delta(n)}\Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})})\phi\left(\frac{\Delta y_{i(\mathbf{x})}}{\Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})})}\right),$$

where $\Delta l(\mathbf{x}, \boldsymbol{\xi}_{1(\mathbf{x})}) = l(\mathbf{x}, \boldsymbol{\xi}_{1(\mathbf{x})})$, and $\Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}) = l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}) - l(\mathbf{x}, \boldsymbol{\xi}_{(i-1)(\mathbf{x})})$ for $i = 2, \ldots, n$. Hence, we see that optimization of dual DRO models is fundamentally different from its primal counterpart (4.3), due to its rank-dependent nature on the decision

variable. To circumvent this, we use (4.13), which gives

$$\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) = \frac{1}{\delta(n)}\sum_{i=1}^{n}(-\phi)_*\left(\delta(n)\frac{n-i+1}{n}\right)\Delta l(\mathbf{x},\boldsymbol{\xi}_{i(\mathbf{x})}).$$

Therefore, the dual DRO minimization problem (4.16) is equivalent to a distortion risk measure minimization problem, where the distortion function is

$$h_{\delta(n)}(p) = (-\phi)_*(\delta(n)p)/(-\phi)_*(\delta(n)).$$

The idea now is to invoke tools from risk theory (see Proposition 10.3 of Denneberg 1994), which is summarized in the following proposition that states an equivalence between distortion risk measures and robust optimization.

Proposition 4.5. *Let $h : [0,1] \to [0,1]$ be a distortion function that is concave, non-decreasing, and satisfies the boundary conditions $h(0) = 0$ and $h(1) = 1$. If $X = [x_1; p_1, x_2; p_2, \ldots, x_n; p_n]$ is a discrete random variable with realizations x_i's that take probabilities p_i's, such that $x_1 \leq x_2 \leq \ldots \leq x_n$, then we have that*

$$\rho_h(X) = \sup_{\mathbf{q}\in M^h(\mathbf{p})}\sum_{i=1}^{n} q_i x_i,$$

where $\rho_h(X)$ is the rank-dependent sum as defined in (4.10), and $M^h(\mathbf{p})$ is the set

$$M^h(\mathbf{p}) = \left\{\mathbf{q}\in\mathbb{R}^n \;\middle|\; \mathbf{q}\geq\mathbf{0},\ \sum_{i=1}^{n} q_i = 1,\ \sum_{j\in J} q_j \leq h\left(\sum_{j\in J} p_j\right), \forall J\subset[n]\right\}. \tag{4.17}$$

Hence, by (4.13), it immediately follows from Theorem 4.5 that problem (4.16) is equivalent to a min-max problem, where the uncertainty set is given by (4.17), for $h = h_{\delta(n)}$. This leads to the following lemma.

Lemma 4.6. *We have for any $\delta(n) > 0$,*

$$\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) = \frac{(-\phi)_*(\delta(n))}{\delta(n)}\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta(n)}}\sum_{i=1}^{n} q_i l(\mathbf{x},\boldsymbol{\xi}_i). \tag{4.18}$$

where

$$\mathcal{M}^{\phi}_{\delta(n)} = \left\{\mathbf{q}\in\mathbb{R}^n \;\middle|\; \mathbf{q}\geq\mathbf{0},\ \sum_{i=1}^{n} q_i = 1,\ \sum_{j\in J} q_j \leq h_{\delta(n)}\left(\sum_{j\in J}\frac{1}{n}\right), \forall J\subset[n]\right\}. \tag{4.19}$$

We note that although the right-hand side of (4.18) is a primal DRO problem, the ambiguity set $\mathcal{M}^{\phi}_{\delta(n)}$ is fundamentally different from the canonical examples that are considered in the standard DRO literature, namely sets that are based on statistical

distances or imposed moment conditions. Instead, $\mathcal{M}^{\phi}_{\delta(n)}$ uses probability weighting, and it contains 2^n number of constraints: $\mathbf{q} \in \mathcal{M}^{\phi}_{\delta(n)}$, if and only if the probability of any event under the distribution $\mathbf{q}$ is bounded by the probability of that event under the distorted empirical distribution. Due to this complexity, the right-hand side of (4.18) cannot be solved using the standard reformulation technique as developed in Ben-Tal et al. [2013]. Fortunately, optimization of distortion risk measure with discrete probabilities has been studied in previous work by Jin et al. [2025b]. One way to efficiently compute (4.18) is to use a piecewise linear approximation of the concave function $h_{\delta(n)}$, both from below and from above. This gives an upper and a lower bound on the optimal value (4.18), both of which converge as the approximation error of $h_{\delta(n)}$ tends to zero. More precisely, for any concave distortion function h, one may approximate h from below with a piecewise-linear function $h_L = \min_{j=1,\ldots,K} h_j$, where $h_j(p) = l_j \cdot p + b_j$ are affine functions such that the slopes $l_1 > \ldots > l_K$ are decreasing, and the intercepts $b_1 < \ldots < b_K$ are increasing. The affine functions are defined on a set of K support points $0 = s_0 < s_1 < \ldots < s_K = 1$, such that $h_L(p) = h_j(p)$, if $p \in [s_{j-1}, s_j]$, for all $j = 1, \ldots, K$. Moreover, we may impose $b_1 = 0$ and $l_K + b_K = 1$, so that $h_L(0) = 0$ and $h_L(1) = 1$. Therefore, if one replaces $h_{\delta(n)}$ by its lower piecewise-linear approximation $h_L \leq h_{\delta(n)}$, then the ambiguity set $\mathcal{M}^{\phi}_{\delta(n)}$ is approximated by the smaller subset $M^{h_L}(\mathbf{1}/n)$ (see definition in (4.17)), where $\mathbf{1}/n = (1/n, \ldots, 1/n) \in \mathbb{R}^n$. Hence, (4.18) can be approximated with the lower bound $\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{q}\in M^{h_L}(\mathbf{1}/n)} \sum_{i=1}^n q_i l(\mathbf{x}, \boldsymbol{\xi}_i)$. As shown by Jin et al. [2025b], this lower bound can be computed as the following convex optimization problem with $O(n \cdot K)$ number of constraints, as summarized in the following theorem.

Theorem 4.7. *Let $h_L = \min_{1\leq j\leq K} h_j$ be a piecewise-linear concave distortion function. Then, we have that $\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{q}\in M^{h_L}(\mathbf{1}/n)} \sum_{i=1}^n q_i l(\mathbf{x}, \boldsymbol{\xi}_i)$ can be computed by solving the following optimization problem:*

$$\begin{aligned} \min_{\substack{\mathbf{x}\in\mathcal{X} \\ \beta, \lambda_{ij}, \nu_j}} \ & \beta + \sum_{j=1}^K \nu_j b_j + \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^K \lambda_{ij} l_j \\ \text{s.t. } & l(\mathbf{x}, \boldsymbol{\xi}_i) - \beta - \sum_{j=1}^K \lambda_{ij} \leq 0, \ \forall i \in [n] \\ & \lambda_{ij} \leq \nu_j, \ \forall i \in [n], \ \forall j \in [K] \\ & \lambda_{ij}, \nu_j \geq 0, \ \forall i \in [n], \ \forall j \in [K]. \end{aligned} \tag{4.20}$$

We further note that as outlined by Jin et al. [2025b], for any constant $\epsilon > 0$, one can find a piecewise-linear approximation $h_L \leq h$, with the smallest number of K pieces, that satisfies the error bound $\sup_{p\in[0,1]} |h_L(p) - h(p)| \leq \epsilon$. This allows us to minimize the number of constraints in (4.20). Furthermore, for any approximation function h_L with error bound ϵ, one can also obtain an upper approximation of $\tilde{h}_L \geq h_{\delta(n)}$ by defining the

function

$$\tilde{h}_L(p) = \begin{cases} 0 & p = 0 \\ \min\{h_L(p) + \epsilon, 1\} & 0 < p \leq 1. \end{cases}$$

Then, the ambiguity set $\mathcal{M}^{\phi}_{\delta(n)}$ is contained in the larger set $M^{\tilde{h}_L}(\mathbf{1}/n)$. This allows us to also compute an upper bound on (4.18), by solving $\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{q}\in M^{\tilde{h}_L}(\mathbf{1}/n)} \sum_{i=1}^{n} q_i l(\mathbf{x}, \boldsymbol{\xi}_i)$, which again can be computed using (4.20), by simply replacing the constants b_j with $b_j + \epsilon$. As further shown by Jin et al. [2025b], if $\inf_{\mathbf{x}\in\mathcal{X}, i=1,\ldots,n} l(\mathbf{x}, \boldsymbol{\xi}_i) > -\infty$ holds, then both bounds converge to the exact value (4.18) as $\epsilon \to 0$. Since our loss function is assumed to be positive, this condition is automatically satisfied.

4.4.2 A Numerical Investigation of the Advantages of Dual DRO

We investigate numerically the difference between dual DRO (4.5), primal DRO (4.3), and SAA (4.2), as data-driven methods for obtaining solutions of the true nominal problem $\inf_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}, \boldsymbol{\xi})]$. In the following two simulation studies, we show that the dual DRO model can provide solutions that are closer to the true optimum of the nominal problem than the primal DRO and the SAA, when the true optimal solution has a small "variability", but a large variance due to the presence of outliers. The Python codes for each experiment are provided on https://github.com/GuanJinNL/Dual_DRO.

Example 1: Toy Example

We consider a linear stochastic optimization problem with two loss random variables $\boldsymbol{\xi} = (\xi_1, \xi_2)$ that follow a similar distribution as the example presented in Figure 4.1. We let $\xi_1 = \text{Unif}[6, 8] + 26\epsilon_0$, where ϵ_0 is a Bernoulli random variable that takes the value 1 with probability 0.05. Let $\xi_2 = \text{Unif}[0.4, 16.4]$. Then, $\mathbb{E}[\xi_1] = 8.3 < 8.4 = \mathbb{E}[\xi_2]$. Hence, in the linear optimization problem where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^2 \mid x_1 + x_2 = 1, x_1 \geq 0, x_2 \geq 0\}$ and $l(\mathbf{x}, \boldsymbol{\xi}) = \mathbf{x}^T\boldsymbol{\xi}$, it is clear that $\mathbf{x} = (1, 0)$ should be the true optimal solution, i.e., ξ_1 is the optimal choice among the linear combinations of loss distributions $x_1\xi_1 + x_2\xi_2$.

To compare the performance between primal DRO, dual DRO, and SAA, we generate 100 samples of $\boldsymbol{\xi}$ of size $n \in \{50, 100, 200, 500, 1{,}000, 2{,}000\}$ and record the average (among the 100) weight $\overline{\mathbf{x}}^*_{1,\text{model}}$ which is defined as the projection value on ξ_1 of the optimal solution obtained from each model $\in \{\text{saa}, \text{p.dro}, \text{d.dro}\}$.[3] For both primal and dual DRO models, we choose the KL-divergence $\phi(t) = t\log t - t + 1$. Then, for each sample of size n that we draw from $\boldsymbol{\xi}$, we choose for the primal DRO model the radius $r(n) = \chi^2_{0.95,1}/(2n)$, and for the dual DRO model, we choose $\delta(n) = \sqrt{2\chi^2_{0.95,1}/n}$. According to Proposition 4.3,

[3] We solve the dual DRO using (4.20), and a piecewise-linear approximation of $h_{\delta(n)}$ with uniform approximation error $\epsilon = 0.0001$.

identity (4.13) and Duchi et al. [2021], we have that for each $\mathbf{x} \in \{\mathbf{x}_1, \mathbf{x}_2\}$,

$$\begin{aligned}
&\sup_{\mathbb{Q}: I_\phi(\mathbb{Q}, \hat{\mathbb{P}}_n) \leq r(n)} \mathbb{E}_{\mathbb{Q}}[l(\mathbf{x}, \boldsymbol{\xi})] \\
&\quad = \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x}, \boldsymbol{\xi})] + \sqrt{\frac{\chi^2_{0.95,1}}{2n}} \sqrt{\mathrm{Var}_{\hat{\mathbb{P}}_n}[l(\mathbf{x}, \boldsymbol{\xi})]} \\
&\qquad + o(1/n), \qquad \text{(p.DRO)}
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\delta(n)}{(-\phi)_*(\delta(n))} \rho^{\phi}_{\delta(n), \hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})) \\
&\quad = \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x}, \boldsymbol{\xi})] + \sqrt{\frac{\chi^2_{0.95,1}}{2n}} \cdot \overline{\mathrm{m}}_{2, \hat{\mathbb{P}}_n}[l(\mathbf{x}, \boldsymbol{\xi})] \\
&\qquad + o(1/n). \qquad \text{(d.DRO)}
\end{aligned}$$

Therefore, our choice of $r(n)$ and $\delta(n)$ ensures that the coefficients for expectation, standard deviation, and the maxiance are all equal, for a fair comparison.

The results are displayed in Table 4.1. As we can observe, the average optimal weight on ξ_1 for the dual DRO model is higher than those for SAA and primal DRO for all sample sizes n. This shows that dual DRO is typically better at identifying the optimal choice ξ_1, in cases where SAA underperforms due to its sensitivity to the large maxiance of ξ_2, and primal DRO underperforms due to its sensitivity to outliers of ξ_1. To further examine the difference between primal and dual DRO relative to the performance of SAA, we calculate the average optimal weight on ξ_1 for both DRO models conditionally upon the realizations $\hat{\mu}_1 > \hat{\mu}_2$ and $\hat{\mu}_1 < \hat{\mu}_2$, where $\hat{\mu}_1, \hat{\mu}_2$ are empirical estimates of $\mathbb{E}[\xi_1], \mathbb{E}[\xi_2]$. Indeed, the two cases $\hat{\mu}_1 < \hat{\mu}_2$ and $\hat{\mu}_1 > \hat{\mu}_2$ correspond to the situations where SAA makes either a perfectly correct or an incorrect decision, respectively. As we can observe from Table 4.1, when $\hat{\mu}_1 > \hat{\mu}_2$, which is the case when SAA outputs the correct optimal solution $(1, 0)$, the dual DRO also does that by putting almost all weights on ξ_1, while the primal DRO puts only half of the weight on ξ_1. On the contrary, when $\hat{\mu}_1 < \hat{\mu}_2$, which is the case when SAA is incorrect and outputs a solution $(0, 1)$, both the dual and primal DRO will make the correction by allocating weights on ξ_1, and on average the dual DRO model does that more than its primal counterpart.

Example 2: Optimal Insurance Portfolio

We consider an insurance portfolio optimization problem, where the insurer aims to design an insurance policy with the optimal risk exposure between two claim distributions. The

Table 4.1: This table displays the average optimal weight on ξ_1 obtained from solving the dual DRO, the primal DRO, and the SAA over 100 samples of $\boldsymbol{\xi}$ of size n. This average is also calculated conditionally upon sample realizations $\hat{\mu}_1 > \hat{\mu}_2$ and $\hat{\mu}_1 < \hat{\mu}_2$, where $\hat{\mu}_1, \hat{\mu}_2$ are empirical estimates of $\mathbb{E}[\xi_1], \mathbb{E}[\xi_2]$.

n	$\overline{\mathbf{x}}^*_{1,\text{saa}}$	$\overline{\mathbf{x}}^*_{1,\text{d.dro}}$	$\overline{\mathbf{x}}^*_{1,\text{p.dro}}$	$\overline{\mathbf{x}}^*_{1,\text{d.dro},\hat{\mu}_1>\hat{\mu}_2}$	$\overline{\mathbf{x}}^*_{1,\text{p.dro},\hat{\mu}_1>\hat{\mu}_2}$	$\overline{\mathbf{x}}^*_{1,\text{d.dro},\hat{\mu}_1<\hat{\mu}_2}$	$\overline{\mathbf{x}}^*_{1,\text{p.dro},\hat{\mu}_1<\hat{\mu}_2}$
50	0.570	0.652	0.393	0.988	0.544	0.206	0.192
100	0.550	0.647	0.346	0.979	0.465	0.240	0.201
200	0.510	0.673	0.380	0.985	0.513	0.348	0.242
500	0.610	0.679	0.375	0.986	0.479	0.198	0.211
1000	0.630	0.749	0.453	0.990	0.564	0.339	0.264
2000	0.790	0.852	0.499	0.992	0.558	0.328	0.280

problem is formulated as

$$\inf_{x \geq 0} \mathbb{E}_{\mathbb{P}_0}[(x\xi_1 + (1-x)\xi_2)^{1.5}], \tag{4.21}$$

where ξ_1, ξ_2 are two positive claim distributions and x is the fraction of coverage that the insurer is willing to provide. To account for risk-aversion, a convex loss function $l(x) = x^{1.5}$ is applied.

We let $\log(\xi_1) \sim \mu_1 + \sigma_1 Z_1 + b$ be a mixture of normal and Bernoulli distributions with $\mu_1 = 5$, $\sigma_1 = 0.2$, Z_1 standard normal and b Bernoulli distributed with parameter 0.05. Let $\log(\xi_2) \sim \mu_2 + \sigma_2 Z_2$ with parameter $\mu_2 = 5.1$, $\sigma_2 = 0.29$, and Z_2 standard normal. The parameters are chosen such that ξ_1 has a larger variance than ξ_2, but has a smaller maxiance on the other hand.

We investigate the performance of SAA, primal DRO, and dual DRO in this example using the same choices of ϕ, $r(n)$ and $\delta(n)$ as in the toy example given in Subsection 4.4.2. Using 100 samples of size 1,000 for (ξ_1, ξ_2), we repeatedly solve the SAA, primal DRO, and dual DRO formulations of (4.21) and record their optimal solutions. The performance of each solution $\mathbf{x}^*$ of the three models is evaluated by an empirical estimation of their expected loss $\mathbb{E}_{\mathbb{P}_0}[(x^*\xi_1 + (1-x^*)\xi_2)^{1.5}]$ using a second independent test sample of (ξ_1, ξ_2) of size 1,000,000, and then averaged over 100 repetitions.

The results are displayed in Table 4.2. As we can observe, the solutions obtained from dual DRO have the lowest average expected loss as compared to SAA and primal DRO, whereas primal DRO has the worst performance, due to the large variance of ξ_1.

Example 3: Median Estimation

We provide another example where the dual DRO model exhibits more robustness against outliers than its primal counterpart. We investigate the problem of estimating the median

	Average Expected Loss of Solutions
SAA	2214
Primal DRO	2222
Dual DRO	2213

Table 4.2: This table displays the average expected loss (estimated using an independent sample of (ξ_1, ξ_2) of size 1,000,000) of 100 solutions obtained from solving each data-driven model (SAA, primal DRO, dual DRO) using 1,000 samples of (ξ_1, ξ_2).

of a distribution, which is also the optimum of the following nominal problem:

$$\inf_{x \in \mathbb{R}} \mathbb{E}_{\mathbb{P}_0} |x - \xi|. \tag{4.22}$$

Using the example given by Duchi and Namkoong [2019], we let $\xi \in \{-1, 0, 1\}$ be a distribution that takes the value 0 with some probability δ_0, and the values -1 and 1 with probability $(1 - \delta_0)/2$. Clearly, the true median is equal to zero. Duchi and Namkoong [2019] has shown that for this particular distribution, the primal DRO model will provide a better estimation of the median than the SAA model, as $\delta_0 \to 0$. For this experiment, we also consider a perturbation $\xi' = \xi + e$, where e is an adversarial attack on the data of ξ such that it takes the value 0 with some probability $1 - \epsilon_0$, but the value 300 with the remaining small probability ϵ_0. We set $\delta_0 = 0.01$ and $\epsilon_0 = 0.009$. By construction, the true median of ξ' is still equal to 0.

We then solve the SAA, the primal DRO, and the dual DRO versions of problem (4.22) for a sample size of n = 1,000, over 100 repetitions. We again use the KL-divergence. For the primal DRO model, we choose the radius $r(n) = \chi^2_{0.999,1}/n$. The $\delta(n)$ in the dual DRO model is then set to be $\delta(n) = \sqrt{2\chi^2_{0.999,1}/n}$ to ensure equal penalization of both DRO models, similar to the previous example. To compare the performance of the three models, we count the number of accepted solutions (over 100 trials) of each model, where a solution is considered accepted if it has an absolute value within 0.01 of the true median 0. We do this both for the samples of ξ, and the samples of ξ', i.e., the perturbed samples.

The results are given in Table 4.3. As we can observe, in the case where the data ξ is not perturbed, the primal DRO model outperforms the SAA model as predicted by Duchi and Namkoong [2019], and the dual DRO model shows even better performance. However, when the perturbed data ξ' is used, the performance of the primal DRO model quickly deteriorates, whereas the dual DRO model is still yielding a similar number of accepted solutions. This again shows that the dual DRO model is more robust against outliers.

	No perturbation	With perturbation
SAA	26	22
Primal DRO	50	22
Dual DRO	86	84

Table 4.3: This table displays the number of accepted solutions obtained from each model, over 100 repetitions. A solution is accepted if its absolute value is within 0.01 of the true median value 0.

4.5 First-Order Asymptotics of Dual DRO Estimators

In this section, we establish the asymptotic convergence rate of the dual DRO model as an estimator of the true expectation minimization problem (4.1), by deriving the limiting distribution of the following quantity

$$\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}}\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))-\min_{\mathbf{x}\in\mathcal{X}}\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right). \tag{4.23}$$

In addition, as suggested by Proposition 4.3, an upper bound on (4.1) can also be obtained by simply minimizing the distortion risk measure as defined in (4.13). Hence, we are also interested in examining the limiting distribution of

$$\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}}\mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))-\min_{\mathbf{x}\in\mathcal{X}}\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right). \tag{4.24}$$

To study both (4.23) and (4.24), we use tools from empirical process theory, where we view the empirical measure as a random element in $l^\infty(\mathcal{H})$, the set of all bounded real-valued functions defined on a class $\mathcal{H}$.[4] In the context of our problem, we define the function class $\mathcal{H}=\{\mathbf{x}\in\mathcal{X}: l(\mathbf{x},\boldsymbol{\xi}):\Omega\to\mathbb{R}\}$. We assume that $\mathcal{H}$ has an envelope function $M_2:\mathbb{R}^k\to\mathbb{R}$, such that $\sup_{\mathbf{x}\in\mathcal{X}}|l(\mathbf{x},\tilde{\boldsymbol{\xi}})|\le M_2(\tilde{\boldsymbol{\xi}}),\ \forall\tilde{\boldsymbol{\xi}}\in\mathbb{R}^k$. We also introduce the following notation, where for each $\mathbf{x}\in\mathcal{X}$, we denote the second dual moment $\mathrm{dm}_{2,\mathbb{P}_0}(\mathbf{x})=\mathbb{E}_{\mathbb{P}_0}[\min\{l(\mathbf{x},\boldsymbol{\xi}^{(1)}),l(\mathbf{x},\boldsymbol{\xi}^{(2)})\}]$, $\sigma^2_{\mathbb{P}_0}(\mathbf{x})=\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x},\boldsymbol{\xi}))$ and with a slight abuse of notation $\overline{\mathrm{m}}_{2,\mathbb{P}_0}(\mathbf{x})=\overline{\mathrm{m}}_{2,\mathbb{P}_0}(l(\mathbf{x},\boldsymbol{\xi}))$. Furthermore, for an empirical distribution $\hat{\mathbb{P}}_n=\frac{1}{n}\sum_{i=1}^n\iota_{\boldsymbol{\xi}_i}$, we denote $\mathrm{dm}_{2,\hat{\mathbb{P}}_n}(\mathbf{x})=\mathbb{E}_{\hat{\mathbb{P}}_n\times\hat{\mathbb{P}}_n}[\min\{l(\mathbf{x},\boldsymbol{\xi}^{(1)}),l(\mathbf{x},\boldsymbol{\xi}^{(2)})\}]$, as an evaluation with respect to the product measure $\hat{\mathbb{P}}_n\times\hat{\mathbb{P}}_n$, and $\overline{\mathrm{m}}_{2,\hat{\mathbb{P}}_n}(\mathbf{x})=\mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x},\boldsymbol{\xi})]-\mathrm{dm}_{2,\hat{\mathbb{P}}_n}(\mathbf{x})$.

We first examine the measurability of (4.23) and (4.24) as functions of Ω to the outcome space $\mathbb{R}$.

Proposition 4.8. *The functions* $\min_{\mathbf{x}\in\mathcal{X}}\mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})):\Omega\to\mathbb{R}$ *and* $\min_{\mathbf{x}\in\mathcal{X}}\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})):\Omega\to\mathbb{R}$ *are measurable, if* $l(\mathbf{x},\tilde{\boldsymbol{\xi}})$ *is a Carathéodory function, i.e., continuous in* $\mathbf{x}$*, for all*

[4]Note that since $\mathcal{X}$ is assumed to be compact and $l(\mathbf{x},\tilde{\boldsymbol{\xi}})$ continuous in $\mathbf{x}$ for all $\tilde{\boldsymbol{\xi}}\in\mathbb{R}^k$, we have $\sup_{\mathbf{x}\in\mathcal{X}}\max_{i=1,\ldots,n}l(\mathbf{x},\boldsymbol{\xi}_i)<\infty$, and therefore (4.23) and (4.24) can also be viewed as elements in $l^\infty(\mathcal{H})$.

$\tilde{\boldsymbol{\xi}}$, and measurable in $\tilde{\boldsymbol{\xi}}$, for all $\mathbf{x}$.

The derivation of the limiting distributions for (4.23) and (4.24) requires the following theorem, which states that the maxiance regularization effect of dual DRO models also holds uniformly in the decision space.

Theorem 4.9. *Let ϕ be a ϕ-divergence function that is four times continuously differentiable at 1. Assume that the envelope function M_2 of the class $\mathcal{H}$ is integrable (i.e., $\mathbb{E}_{\mathbb{P}_0}|M_2(\boldsymbol{\xi})| < \infty$). Let $\delta(n) := \delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$ be a measurable function such that $\delta(\boldsymbol{\xi}_1(\omega), \ldots, \boldsymbol{\xi}_n(\omega)) > 0$ for all $\omega \in \Omega$ and $\delta(n) \xrightarrow{P} 0$. Then, we have that*

$$\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) = \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x},\boldsymbol{\xi})] - \frac{\delta(n)}{2\phi''(1)}\mathrm{dm}_{2,\hat{\mathbb{P}}_n}(\mathbf{x}) + \epsilon_n(\mathbf{x})$$
$$\mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) = \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x},\boldsymbol{\xi})] + \frac{\delta(n)}{2\phi''(1)}\overline{\mathrm{m}}_{2,\hat{\mathbb{P}}_n}(\mathbf{x}) + \tilde{\epsilon}_n(\mathbf{x}),$$

where $\sup_{\mathbf{x}\in\mathcal{X}} |\epsilon_n(\mathbf{x})|/\delta(n),\ \sup_{\mathbf{x}\in\mathcal{X}} |\tilde{\epsilon}_n(\mathbf{x})|/\delta(n) \xrightarrow{P} 0$.

We are now ready to state the following main theorem, which shows that if we choose $\delta(n) = \sqrt{r/n}$, for some $r > 0$, then (4.23) and (4.24) converge to a non-centered Gaussian process. This requires the assumption that the class $\mathcal{H}$ is $\mathbb{P}_0$-Donsker, which means that $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_0)$ converges weakly to a tight limit, as elements in $l^{\infty}(\mathcal{H})$. This is, for example, satisfied for function classes that are Hölder continuous (see Example 2.11.13 of van der Vaart and Wellner 2023 for more details).

Theorem 4.10. *Let $\mathcal{H}$ be a $\mathbb{P}_0$-Donsker class with a square integrable envelope function M_2. Assume ϕ is four times continuously differentiable in a neighborhood of 1. Denote $\mathcal{X}^*_{\mathbb{P}_0} = \mathrm{argmin}_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]$. Then, we have that for $\delta(n) = \sqrt{r/n}$,*

$$\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right)$$
$$\rightsquigarrow \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} G(\mathbf{x}) - \frac{\sqrt{r}}{2\phi''(1)}\mathrm{dm}_{2,\mathbb{P}_0}(\mathbf{x}),$$

and

$$\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right)$$
$$\rightsquigarrow \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} G(\mathbf{x}) + \frac{\sqrt{r}}{2\phi''(1)}\overline{\mathrm{m}}_{2,\mathbb{P}_0}(\mathbf{x}),$$

where $G(\mathbf{x})$ is a mean-zero Gaussian process with covariance

$$\mathrm{Cov}(\mathbf{x}_1,\mathbf{x}_2) = \mathrm{Cov}(l(\mathbf{x}_1,\boldsymbol{\xi}), l(\mathbf{x}_2,\boldsymbol{\xi})), \quad \mathbf{x}_1,\mathbf{x}_2 \in \mathcal{X}.$$

As shown in Theorem 4.10, at a rate of $1/\sqrt{n}$, the gap between the dual DRO upper and lower bounds with the true nominal optimal value converges in probability to zero. Moreover, the bias term is the evaluation of the Gaussian process (which is also present in SAA and primal DRO) and the extra maxiance term (or the second dual moment) on the set of optimal solutions of the true nominal problem (4.1). This is different from the bias term of the primal DRO, where the standard deviation is the extra term in the bias besides the Gaussian process (see Duchi et al. 2021). As mentioned in Yitzhaki and Schechtman [2012] (identity (2.21)), the maxiance is always dominated by the standard deviation.[5] This is especially true in the case of a heavy-tailed distribution, where the variance is much larger than the maxiance. Therefore, dual DRO typically has a smaller bias than primal DRO.

We note that as a consequence of Theorem 4.10, if $\mathcal{X}^*_{\mathbb{P}_0} = \{\mathbf{x}^*\}$, then we can approximate the asymptotic probability that $\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))$ and $\min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))$ constitute one-sided lower and upper bounds for $\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]$ with a Gaussian distribution. Indeed, let Φ denote the standard normal cumulative distribution function. Then, we have that

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) \le \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right) = \Phi\left(\frac{\sqrt{r}}{2\phi''(1)}\frac{\mathrm{dm}_{2,\mathbb{P}_0}(\mathbf{x}^*)}{\sigma_{\mathbb{P}_0}(\mathbf{x}^*)}\right) \tag{4.25}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] \le \min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))\right) = \Phi\left(\frac{\sqrt{r}}{2\phi''(1)}\frac{\overline{\mathrm{m}}_{2,\mathbb{P}_0}(\mathbf{x}^*)}{\sigma_{\mathbb{P}_0}(\mathbf{x}^*)}\right). \tag{4.26}$$

Equations (4.25) and (4.26) allow us to calibrate r (and thus $\delta(n)$) such that the quantities $\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))$ and $\min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))$ become one-sided lower and upper bounds for the optimal value $\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]$ under a certain confidence value α.

Since $\mathbf{x}^*$ is an unknown quantity, the calibration of r through (4.25) and (4.26) will require splitting the sample $(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)$ into two independent parts, where the first part is used to obtain a consistent estimator for $\mathbf{x}^*$ (for example, one can use the solution of SAA as the consistent estimator), and the second part is used to estimate $\sigma_{\mathbb{P}_0}(\mathbf{x}^*)$, $\mathrm{dm}_{2,\mathbb{P}_0}(\mathbf{x}^*)$ and $\overline{\mathrm{m}}_{2,\mathbb{P}_0}(\mathbf{x}^*)$. In Section 4.B, we provide more details on this calibration of r.

If $\mathcal{X}^*_{\mathbb{P}_0}$ contains more than one solution, then the non-centered Gaussian process in Theorem 4.10 becomes hard to evaluate. In this case, we note that the limiting probability on the left-hand side of (4.25) can be bounded from below by the right-hand side of (4.25). Namely, for any $\mathbf{x}^* \in \mathcal{X}^*_{\mathbb{P}_0}$, one has

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) \le \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right) \ge \Phi\left(\frac{\sqrt{r}}{2\phi''(1)}\frac{\mathrm{dm}_{2,\mathbb{P}_0}(\mathbf{x}^*)}{\sigma_{\mathbb{P}_0}(\mathbf{x}^*)}\right).$$

Therefore, one can still calibrate r in the same procedure as previously discussed to ensure

[5] Using identity (2.21) of Yitzhaki and Schechtman [2012], one can derive that $\overline{\mathrm{m}}_2(X) \le \frac{1}{\sqrt{3}}\sigma_X \le \sigma_X$, for any distribution X.

that $\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))$ becomes a lower bound on the optimal value with a certain confidence level α. This is, however, not possible with the dual DRO upper bound in (4.26). In that case, one may consider an alternative upper bound for the optimal value. For example, using a second independent sample, one can estimate $\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}_{\text{ddro}},\boldsymbol{\xi})]$ of the dual DRO solution $\mathbf{x}_{\text{ddro}}$ obtained from solving $\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))$, since for any feasible solution $\mathbf{x}\in\mathcal{X}$, $\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]$ is an upper bound for $\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]$.

4.6 Conclusion

We have developed the dual DRO model that is shown to be asymptotically equivalent to mean-maxiance regularization. As illustrated with numerical examples, dual DRO provides (much) more robustness against outliers than the ϕ-divergence primal DRO does due to its maxiance regularization effect. This makes dual DRO an attractive model when used to address overfitting in empirical optimization. In addition, the dual DRO model can also be generalized to a primal-dual DRO model, where both the variance and the maxiance are regularized. Furthermore, we show that optimization of dual DRO models enjoys similar tractability as ϕ-divergence primal DRO, and provides moreover an estimator with a smaller bias for the nominal optimal value $\inf_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}, \boldsymbol{\xi})]$.

Appendix

4.A Proofs

Proof of Proposition 4.1. Since $\mathrm{dom}(\phi) \subset [0, \infty)$, we have that ϕ^* is non-decreasing. We have that $\phi^*(0) = -\inf_{x\geq 0} \phi(x) = 0$ and that $\phi^*(y) = \sup_{x\geq 0} yx - \phi(x) \geq y - \phi(1) = y$. If ϕ is twice continuously differentiable at 1 with $\phi''(1) > 0$, then the first order condition $y - \phi'(x) = 0$ of $\phi^*(y)$ is satisfied for y in a neighborhood of $0 = \phi'(1)$. It then follows from the implicit function theorem that in a neighborhood of 0, there exists a continuously differentiable function $z(y)$, such that $z(0) = 1$ and $y - \phi'(z(y)) = 0$. Therefore, in a neighborhood of 0, we have that $\phi^*(y) = yz(y) - \phi(z(y))$, $(\phi^*)'(y) = z(y)$, $(\phi^*)''(y) = z'(y) = \frac{1}{\phi''(z(y))}$, and $(\phi^*)'''(y) = z''(y) = -\frac{\phi'''(z(y))}{(\phi''(z(y))^3}$. □

Proof of Theorem 4.2. As a preliminary step, following a similar argument as in The-

orem 6.1 in Ben-Tal et al. [1991], we have that

$$\begin{aligned}
\rho^{\phi}_{\delta,\mathbf{p}}(X) &= \inf_{\Delta\mathbf{y}\in\mathbb{R}^n_+}\left\{\sum_{i=1}^n \overline{F}_i\Delta y_i + \frac{1}{\delta}\sum_{i=1}^n \Delta x_i\phi\left(\frac{\Delta y_i}{\Delta x_i}\right)\right\}\\
&= \sum_{i=1}^n \inf_{\Delta y_i\geq 0} \overline{F}_i\Delta y_i + \frac{1}{\delta}\Delta x_i\phi\left(\frac{\Delta y_i}{\Delta x_i}\right)\\
&\overset{(*)}{=} \sum_{i:\Delta x_i>0} \inf_{\Delta y_i\geq 0} \overline{F}_i\Delta y_i + \frac{1}{\delta}\Delta x_i\phi\left(\frac{\Delta y_i}{\Delta x_i}\right)\\
&= \sum_{i:\Delta x_i>0} \inf_{\frac{\Delta y_i}{\Delta x_i}\in\mathbb{R}_+} \left\{\overline{F}_i\frac{\Delta y_i}{\Delta x_i} + \frac{1}{\delta}\phi\left(\frac{\Delta y_i}{\Delta x_i}\right)\right\}\Delta x_i\\
&= \sum_{i:\Delta x_i>0} \frac{1}{\delta}\inf_{t\in\mathbb{R}_+}\left\{\delta\overline{F}_i t - (-\phi)(t)\right\}\Delta x_i + \sum_{i:\Delta x_i=0}(-\phi)_*(\delta\overline{F}_i)\Delta x_i\\
&= \frac{1}{\delta}\sum_{i=1}^n (-\phi)_*(\delta\overline{F}_i)\Delta x_i,
\end{aligned}$$

where at $(*)$ we used that if $\Delta x_i = 0$, then $0\phi\left(\frac{\Delta y_i}{0}\right) = \infty$ for all $\Delta y_i \neq 0$, and thus the infimum is obtained at $\Delta y_i = \Delta x_i = 0$, where $0\phi\left(\frac{0}{0}\right) = 0$. By Proposition 4.1, we have that ϕ^* is twice continuously differentiable at 0, with $\phi^*(0) = 0$, $(\phi^*)'(0) = 1$ and $(\phi^*)''(0) = \frac{1}{\phi''(1)}$. Moreover, we have that $(-\phi)_*(-x) = -\phi^*(-x)$. Therefore, using a second-order Taylor expansion around 0, we obtain

$$\begin{aligned}
\rho^{\phi}_{\delta,\mathbf{p}}(X) &= \frac{1}{\delta}\sum_{i=1}^n (-\phi)_*(\delta\overline{F}_i)\Delta x_i\\
&= \frac{1}{\delta}\sum_{i=1}^n \left[(-\phi)_*(0) + (-\phi)'_*(0)\delta\overline{F}_i + \frac{1}{2}(-\phi)''_*(0)\delta^2(\overline{F}_i)^2 + o(\delta^2)(\overline{F}_i)^2\right]\Delta x_i\\
&= \sum_{i=1}^n \overline{F}_i\Delta x_i - \frac{\delta}{2\phi''(1)}\sum_{i=1}^n (\overline{F}_i)^2\Delta x_i + o(\delta)\cdot\sum_{i=1}^n (\overline{F}_i)^2\Delta x_i\\
&= \sum_{i=1}^n \overline{F}_i\Delta x_i - \frac{\delta}{2\phi''(1)}\sum_{i=1}^n (\overline{F}_i)^2\Delta x_i + o(\delta)\\
&= \mathbb{E}_{\mathbf{p}}[X] - \frac{\delta}{2\phi''(1)}\mathbb{E}_{\mathbf{p}}[\min\{X^{(1)}, X^{(2)}\}] + o(\delta)\\
&= \left(1 - \frac{\delta}{2\phi''(1)}\right)\mathbb{E}_{\mathbf{p}}[X] + \frac{\delta}{2\phi''(1)}\overline{\mathrm{m}}_{2,\mathbf{p}}(X) + o(\delta).
\end{aligned}$$

□

Proof of Proposition 4.3. We have that

$$\begin{aligned}
\mathcal{R}^{\phi}_{\delta,\mathbf{p}}(X) &= \frac{1}{(-\phi)_*(\delta)} \sum_{i=1}^{n} (-\phi)_*(\delta \overline{F}_i) \Delta x_i \\
&= \frac{1}{(-\phi)_*(\delta)} \sum_{i=1}^{n} \delta \overline{F}_i \Delta x_i - \frac{1}{2\phi''(1)} \delta^2 \overline{F}_i^2 \Delta x_i + o(\delta^2) \\
&= \frac{1}{\delta - \frac{1}{2\phi''(1)}\delta^2 + o(\delta^2)} \sum_{i=1}^{n} \delta \overline{F}_i \Delta x_i - \frac{1}{2\phi''(1)} \delta^2 \overline{F}_i^2 \Delta x_i + o(\delta^2) \\
&= \frac{1}{1 - \frac{1}{2\phi''(1)}\delta + o(\delta)} \sum_{i=1}^{n} \overline{F}_i \Delta x_i - \frac{1}{2\phi''(1)} \delta \overline{F}_i^2 \Delta x_i + o(\delta) \\
&= \sum_{i=1}^{n} \overline{F}_i \Delta x_i + \left(\frac{1}{1 - \frac{1}{2\phi''(1)}\delta + o(\delta)} - 1 \right) \sum_{i=1}^{n} \overline{F}_i \Delta x_i - \frac{\delta}{2\phi''(1)} \sum_{i=1}^{n} \overline{F}_i^2 \Delta x_i + o(\delta) \\
&= \sum_{i=1}^{n} \overline{F}_i \Delta x_i + \left(\frac{\frac{1}{2\phi''(1)}\delta + o(\delta)}{1 - \frac{1}{2\phi''(1)}\delta + o(\delta)} \right) \sum_{i=1}^{n} \overline{F}_i \Delta x_i - \frac{\delta}{2\phi''(1)} \sum_{i=1}^{n} \overline{F}_i^2 \Delta x_i + o(\delta) \\
&= \sum_{i=1}^{n} \overline{F}_i \Delta x_i + \left(\frac{1 + o(1)}{1 - \frac{1}{2\phi''(1)}\delta + o(\delta)} \right) \frac{\delta}{2\phi''(1)} \sum_{i=1}^{n} \overline{F}_i \Delta x_i - \frac{\delta}{2\phi''(1)} \sum_{i=1}^{n} \overline{F}_i^2 \Delta x_i + o(\delta) \\
&= \sum_{i=1}^{n} \overline{F}_i \Delta x_i + (1 + o(1)) \frac{\delta}{2\phi''(1)} \sum_{i=1}^{n} \overline{F}_i \Delta x_i - \frac{\delta}{2\phi''(1)} \sum_{i=1}^{n} \overline{F}_i^2 \Delta x_i + o(\delta) \\
&= \sum_{i=1}^{n} \overline{F}_i \Delta x_i + \frac{\delta}{2\phi''(1)} \left(\sum_{i=1}^{n} \overline{F}_i \Delta x_i - \overline{F}_i^2 \Delta x_i \right) + o(1) \cdot \frac{\delta}{2\phi''(1)} \sum_{i=1}^{n} \overline{F}_i \Delta x_i + o(\delta) \\
&= \sum_{i=1}^{n} \overline{F}_i \Delta x_i + \frac{\delta}{2\phi''(1)} \left(\sum_{i=1}^{n} \overline{F}_i \Delta x_i - \overline{F}_i^2 \Delta x_i \right) + o(\delta) \\
&= \mathbb{E}_p[X] + \frac{\delta}{2\phi''(1)} \overline{\mathbb{m}}_{2,\mathbf{p}}(X) + o(\delta).
\end{aligned}$$

□

Proof of Theorem 4.4. Since $\delta_1, \delta_2 \to 0$, we may assume, without loss of generality, that $\delta_1 \leq 1$. Following the proof of Theorem 4.2, we have that

$$\rho_{xp}(X) = \sup_{\mathbf{q} \in \mathbb{P}^n} \left\{ \sum_{i=1}^{n} q_i x_i - \frac{\delta_1}{2\phi''(1)} \left(\sum_{j=i}^{n} q_j \right)^2 \Delta x_i - \frac{1}{\delta_2} p_i \psi \left(\frac{q_i}{p_i} \right) \right\} + o(\delta_1).$$

We have

$$\begin{aligned}\sum_{i=1}^{n}\left(\sum_{j=i}^{n}q_j\right)^2 &= \sum_{i=1}^{n}\left(\sum_{j=i}^{n}p_j+\sum_{j=i}^{n}q_j-p_j\right)^2\\ &= \sum_{i=1}^{n}\left(\sum_{j=i}^{n}p_j\right)^2+2\sum_{j=i}^{n}p_j\sum_{k=i}^{n}q_k-p_k+\left(\sum_{j=i}^{n}q_j-p_j\right)^2 \qquad (4.27)\\ &= -\sum_{i=1}^{n}\left(\sum_{j=i}^{n}p_j\right)^2+2\sum_{j=i}^{n}p_j\sum_{k=i}^{n}q_k+\left(\sum_{j=i}^{n}q_j-p_j\right)^2.\end{aligned}$$

Hence, we have that

$$\begin{aligned}\rho_{xp}(X) = \sup_{\mathbf{q}\in\mathbb{P}^n}\sum_{i=1}^{n}q_ix_i-\frac{\delta_1}{\phi''(1)}\sum_{j=i}^{n}p_j\sum_{k=i}^{n}q_k\Delta x_i-\frac{1}{\delta_2}p_i\psi\left(\frac{q_i}{p_i}\right)-\frac{\delta_1}{2\phi''(1)}\left(\sum_{j=i}^{n}q_j-p_j\right)^2\\ \cdot\Delta x_i+\frac{\delta_1}{2\phi''(1)}\left(\sum_{j=i}^{n}p_j\right)^2\Delta x_i+o(\delta_1).\end{aligned}$$

We note that the last two terms above do not depend on $\mathbf{q}$. Hence, we examine only

$$\begin{aligned}&\sup_{\mathbf{q}\in\mathbb{P}^n}\sum_{i=1}^{n}q_ix_i-\frac{\delta_1}{\phi''(1)}\sum_{j=i}^{n}p_j\sum_{k=i}^{n}q_k\Delta x_i-\frac{1}{\delta_2}p_i\psi\left(\frac{q_i}{p_i}\right)-\frac{\delta_1}{2\phi''(1)}\left(\sum_{j=i}^{n}q_j-p_j\right)^2\Delta x_i\\ &\triangleq \sup_{\mathbf{q}\in\mathbb{P}^n}E(\mathbf{q}).\end{aligned}$$

We have that each $\mathbf{q}\in\mathbb{P}^n$, and $\delta_1\leq 1$:

$$\begin{aligned}E(\mathbf{q}) &\leq x_{\max}+\frac{\delta_1}{\phi''(1)}\sum_{i=1}^{n}\Delta x_i+\frac{2\delta_1}{\phi''(1)}\sum_{i=1}^{n}\Delta x_i-\frac{1}{\delta_2}\sum_{i=1}^{n}p_i\psi\left(\frac{q_i}{p_i}\right)\\ &\leq \left(1+\frac{3}{\phi''(1)}\right)x_{\max}-\frac{1}{\delta_2}\sum_{i=1}^{n}p_i\psi\left(\frac{q_i}{p_i}\right),\end{aligned}$$

since $\sum_{j=i}^{n}p_j\sum_{k=i}^{n}q_k\leq 1$ and $\left(\sum_{j=i}^{n}q_j-p_j\right)^2\leq 4$, for all $i=1,\ldots,n$. On the other hand, we also have $E(\mathbf{p})\geq\sum_{i=1}^{n}p_ix_i-\frac{1}{\phi''(1)}x_{\max}$. Therefore, we have that for all $\delta_2>0$, if $\mathbf{q}\in\mathbb{P}^n$ is any probability vector such that

$$\sum_{i=1}^{n}p_i\psi\left(\frac{q_i}{p_i}\right)>\delta_2K,$$

where $K > 0$ is any constant (independent of δ_1, δ_2), such that

$$K \geq \left(1 + \frac{4}{\phi''(1)}\right) x_{\max} - \sum_{i=1}^{n} p_i x_i. \tag{4.28}$$

Then, for all such $\mathbf{q}$, we would have $E(\mathbf{q}) \leq E(\mathbf{p})$. Therefore, we may choose any positive constant K that satisfies (4.28), such that for all $\delta_2 > 0$, we can restrict $\mathbf{q}$ in a ψ-divergence ball without changing the optimum:

$$\sup_{\mathbf{q}\in\mathbb{P}^n} E(\mathbf{q}) = \sup_{\mathbf{q}\in\mathbb{P}^n:\sum_{i=1}^n p_i\psi\left(\frac{q_i}{p_i}\right)\leq\delta_2 K} E(\mathbf{q}).$$

Our next step is to bound the quadratic term $\sum_{i=1}^n \left(\sum_{j=i}^n q_j - p_j\right)^2 \Delta x_i$ of $E(\mathbf{q})$ in the order of δ_2. From Lemma 4.A.1, it follows that for all $\mathbf{q} \in \mathbb{P}^n$ such that $\sum_{i=1}^n p_i\psi\left(\frac{q_i}{p_i}\right) \leq \delta_2 K$ with δ_2 sufficiently small, then the following inequality holds for all $i = 2, \ldots, n$

$$\left(\sum_{j=i}^{n} q_j - p_j\right)^2 \leq (n-i)\sum_{j=i}^{n}(q_j - p_j)^2 \leq n\sum_{j=1}^{n}(q_j - p_j)^2 \leq n\sum_{j=1}^{n}\frac{(q_j - p_j)^2}{p_j} \leq 4\psi''(1)n\delta_2 K,$$

where we used Cauchy-Schwarz for the first inequality. Hence, we also have

$$\begin{aligned}\left|\frac{\delta_1}{2\phi''(1)}\sum_{i=1}^{n}\left(\sum_{j=i}^{n} q_j - p_j\right)^2 \Delta x_i\right| &\leq \frac{\delta_1}{2\phi''(1)} 4\psi''(1)n\delta_2 K \sum_{i=1}^{n}\Delta x_i = 2\frac{\psi''(1)}{\phi''(1)} n\delta_1\delta_2 K x_{\max} \\ &= O(\delta_1\delta_2 \cdot n).\end{aligned}$$

Thus, for all $\mathbf{q} \in \mathbb{P}^n$ such that $\sum_{i=1}^n p_i\psi\left(\frac{q_i}{p_i}\right) \leq \delta_2 K$, the following holds

$$\begin{aligned}E(\mathbf{q}) &= \sum_{i=1}^{n} q_i x_i - \frac{\delta_1}{\phi''(1)}\sum_{j=i}^{n} p_j \sum_{k=i}^{n} q_k \Delta x_i - \frac{1}{\delta_2} p_i\psi\left(\frac{q_i}{p_i}\right) + O(\delta_1\delta_2 \cdot n) \\ &\triangleq E_l(\mathbf{q}) + O(\delta_1\delta_2 \cdot n).\end{aligned}$$

Hence, we have shown that

$$\sup_{\mathbf{q}\in\mathbb{P}^n:\sum_{i=1}^n p_i\psi\left(\frac{q_i}{p_i}\right)\leq\delta_2 K} E(\mathbf{q}) = \sup_{\mathbf{q}\in\mathbb{P}^n:\sum_{i=1}^n p_i\psi\left(\frac{q_i}{p_i}\right)\leq\delta_2 K} E_l(\mathbf{q}) + O(\delta_1\delta_2 \cdot n).$$

We now work back towards

$$\sup_{\mathbf{q}\in\mathbb{P}^n:\sum_{i=1}^n p_i\psi\left(\frac{q_i}{p_i}\right)\leq\delta_2 K} E_l(\mathbf{q}) = \sup_{\mathbf{q}\in\mathbb{P}^n} E_l(\mathbf{q}). \tag{4.29}$$

Indeed, using the same argument before, we also have that

$$E_l(\mathbf{q}) \leq x_{\max} + \frac{1}{\phi''(1)} x_{\max} - \frac{1}{\delta_2} \sum_{i=1}^{n} p_i \psi\left(\frac{q_i}{p_i}\right),$$

and $E_l(\mathbf{p}) \geq \sum_{i=1}^{n} p_i x_i - \frac{1}{\phi''(1)} x_{\max}$. Hence, our choice of K that satisfies (4.28) also ensures (4.29). We now examine:

$$\begin{aligned}
\sup_{\mathbf{q}\in\mathbb{P}^n} E_l(\mathbf{q}) &= \sup_{\mathbf{q}\in\mathbb{P}^n} \sum_{i=1}^{n} q_i x_i - \frac{\delta_1}{\phi''(1)} \sum_{j=i}^{n} p_j \sum_{k=i}^{n} q_k \Delta x_i - \frac{1}{\delta_2} p_i \psi\left(\frac{q_i}{p_i}\right) \\
&= \sup_{\mathbf{q}\in\mathbb{P}^n} \sum_{i=1}^{n} \left(x_i - \frac{\delta_1}{\phi''(1)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j \right) q_i - \frac{1}{\delta_2} p_i \psi\left(\frac{q_i}{p_i}\right) \\
&= \inf_{\lambda\in\mathbb{R}} \lambda + \sup_{\mathbf{q}\in\mathbb{R}^n_+} \sum_{i=1}^{n} \left(x_i - \frac{\delta_1}{\phi''(1)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j - \lambda \right) q_i - \frac{1}{\delta_2} p_i \psi\left(\frac{q_i}{p_i}\right) \\
&= \inf_{\lambda\in\mathbb{R}} \lambda + \frac{1}{\delta_2} \sum_{i=1}^{n} p_i \sup_{q_i\in\mathbb{R}_+} \left(\delta_2(x_i - \lambda) - \frac{\delta_1\delta_2}{\phi''(1)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j \right) \frac{q_i}{p_i} - \psi\left(\frac{q_i}{p_i}\right) \\
&= \inf_{\lambda\in\mathbb{R}} \lambda + \frac{1}{\delta_2} \sum_{i=1}^{n} p_i \sup_{t\in\mathbb{R}_+} \left(\delta_2(x_i - \lambda) - \frac{\delta_1\delta_2}{\phi''(1)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j \right) t - \psi(t) \\
&= \inf_{\lambda\in\mathbb{R}} \lambda + \frac{1}{\delta_2} \sum_{i=1}^{n} p_i \psi^*\left(\delta_2(x_i - \lambda) - \frac{\delta_1\delta_2}{\phi''(1)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j \right).
\end{aligned} \tag{4.30}$$

By proposition 4.1, we have that $\psi^*(0) = 0$, $(\psi^*)'(0) = 1$, $(\psi^*)''(0) = \frac{1}{\psi''(1)}$ and $(\psi^*)'$ is increasing around zero due to convexity. With these properties, we use a similar argument of Proposition 2.1 in Ben-Tal and Teboulle [2007] to show that we may restrict λ to any compact set $[-b, b]$ with b sufficiently large (see Lemma 4.A.2). This gives us

$$\inf_{\lambda\in[-b,b]} \lambda + \frac{1}{\delta_2} \sum_{i=1}^{n} p_i \psi^*\left(\delta_2 \left[x_i - \lambda - \frac{\delta_1}{\phi''(1)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j \right] \right).$$

Denote $B_i(\lambda) \triangleq x_i - \lambda - \frac{\delta_1}{\phi''(1)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j$. Note that $B_i(\lambda)$ is a continuous function of λ and is thus bounded on a compact set $\lambda \in [-b, b]$. Therefore, by using a Taylor expansion of ψ^* around zero, we obtain

$$\begin{aligned}
\inf_{\lambda\in[-b,b]} \lambda + \frac{1}{\delta_2} \sum_{i=1}^{n} p_i \psi^*\left(\delta_2 B_i(\lambda)\right) = \inf_{\lambda\in[-b,b]} \lambda + \frac{1}{\delta_2} \sum_{i=1}^{n} p_i \left(\delta_2 B_i(\lambda)\right) + p_i \frac{1}{2\psi''(1)} \delta_2^2 (B_i(\lambda))^2 \\
+ o(\delta_2^2).
\end{aligned}$$

The above expression can be further expanded to

$$\begin{aligned}
&\sum_{i=1}^{n} p_i x_i - \frac{\delta_1}{\phi''(1)} \sum_{i=1}^{n} \left(\sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j \right) p_i + \frac{1}{2\psi''(1)} \inf_{\lambda \in [-b,b]} \sum_{i=1}^{n} p_i \delta_2 (B_i(\lambda))^2 + o(\delta_2) \\
&= \sum_{i=1}^{n} p_i x_i - \frac{\delta_1}{\phi''(1)} \sum_{i=1}^{n} \left(\sum_{j=i}^{n} p_j \right)^2 \Delta x_i + \frac{1}{2\psi''(1)} \inf_{\lambda \in [-b,b]} \sum_{i=1}^{n} p_i \delta_2 (B_i(\lambda))^2 + o(\delta_2),
\end{aligned}$$

where

$$\sum_{i=1}^{n} \left(\sum_{j=i}^{n} p_j \right)^2 \Delta x_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j \right) p_i$$

follows from examining the derivations in (4.27) and (4.30). We expand $\delta_2 B_i(\lambda)^2$, which is

$$\begin{aligned}
\delta_2 B_i(\lambda)^2 &= \delta_2 \left(x_i - \lambda - \frac{\delta_1}{\phi''(1)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j \right)^2 \\
&= \delta_2 (x_i - \lambda)^2 - 2\delta_1 \delta_2 (x_i - \lambda) \left(\frac{1}{\phi''(1)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j \right) + \frac{\delta_2 \delta_1^2}{(\phi''(1))^2} \left(\sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j \right)^2 \\
&= \delta_2 (x_i - \lambda)^2 + O(\delta_1 \delta_2).
\end{aligned}$$

Therefore, we have that

$$\inf_{\lambda \in [-b,b]} \sum_{i=1}^{n} p_i \delta_2 (B_i(\lambda))^2 = \delta_2 \mathrm{Var}_{\mathbf{p}}(X) + O(\delta_1 \delta_2),$$

by taking b sufficiently large such that $\sum_{i=1}^{n} p_i x_i \in [-b, b]$. Hence, we obtain as a final expression

$$\begin{aligned}
\rho_{xp}(X) &= \mathbb{E}_{\mathbf{p}}[X] + \frac{\delta_2}{2\psi''(1)} \mathrm{Var}_{\mathbf{p}}(X) - \frac{\delta_1}{2\phi''(1)} \sum_{i=1}^{n} \left(\sum_{j=i}^{n} p_j \right)^2 \Delta x_i + o(\delta_1) + o(\delta_2) + O(\delta_1 \delta_2 \cdot n) \\
&= \mathbb{E}_{\mathbf{p}}[X] + \frac{\delta_2}{2\psi''(1)} \mathrm{Var}_{\mathbf{p}}(X) - \frac{\delta_1}{2\phi''(1)} \mathbb{E}_{\mathbf{p}} \left[\min\{X^{(1)}, X^{(2)}\} \right] + o(\delta_1) + o(\delta_2) + O(\delta_1 \delta_2 \cdot n) \\
&= \left(1 - \frac{\delta_1}{2\phi''(1)} \right) \mathbb{E}_{\mathbf{p}}[X] + \frac{\delta_2}{2\psi''(1)} \mathrm{Var}_{\mathbf{p}}(X) + \frac{\delta_1}{2\phi''(1)} \overline{\mathbb{m}}_{2,\mathbf{p}}(X) + o(\delta_1) + o(\delta_2) \\
&\quad + O(\delta_1 \delta_2 \cdot n).
\end{aligned}$$

□

Lemma 4.A.1. *Suppose that ψ is a Csiszar ϕ-divergence function such that ψ is twice continuously differentiable at 1 with $\psi''(1) > 0$. Then, for any constant $K > 0$, there*

exists a $\delta_0 > 0$, such that for all $\delta_2 \le \delta_0$, we have

$$\sum_{i=1}^{n} p_i \psi\left(\frac{q_i}{p_i}\right) \le \delta_2 K \Rightarrow \sum_{i=1}^{n} p_i \left(\frac{q_i}{p_i} - 1\right)^2 \le \delta_2 K \cdot 4\psi''(1),$$

where we assume $p_{\min} \triangleq \min_i p_i > 0$.

Proof. Since $\psi(1) = 0$ and $\psi'(1) = 0$, a Taylor expansion shows that

$$\sum_{i=1}^{n} p_i \psi\left(\frac{q_i}{p_i}\right) = \sum_{i=1}^{n} p_i \left[\frac{\psi''(1)}{2}\left(\frac{q_i}{p_i} - 1\right)^2 + \left(\frac{q_i}{p_i} - 1\right)^2 R_2\left(\frac{q_i}{p_i}\right)\right],$$

where $\lim_{t\to 1} R_2(t) = 0$. Then, for all $\mathbf{q} \in \mathbb{P}^n$ such that $\sum_{i=1}^n p_i \psi\left(\frac{q_i}{p_i}\right) \le \delta_2 K$, we have that for all $i = 1, \ldots, n$

$$p_i \psi\left(\frac{q_i}{p_i}\right) \le \sum_{i=1}^{n} p_i \psi\left(\frac{q_i}{p_i}\right) \le \delta_2 K.$$

Hence, for all i, we have

$$\psi\left(\frac{q_i}{p_i}\right) \le \delta_2 K / p_i \le \delta_2 K / p_{\min}.$$

Since ψ is twice continuously differentiable around 1 with $\psi''(1) > 0$, it follows from convexity that ψ is strictly decreasing on $[0, 1]$ and strictly increasing on $[1, \infty]$. Hence, we have for any $\epsilon > 0$, that $M_\psi(\epsilon) \triangleq \max_{t \in [1-\epsilon, 1+\epsilon]} \psi(t) > 0$. Therefore, for all $\delta_2 > 0$ such that $\delta_2 K / p_{\min} < M_\psi(\epsilon)$, we have that

$$\psi\left(\frac{q_i}{p_i}\right) \le \delta_2 K / p_{\min} \Rightarrow \left|\frac{q_i}{p_i} - 1\right| \le \epsilon, \forall i = 1, \ldots, n.$$

Since $\lim_{t\to 1} R_2(t) = 0$, there exists an $\epsilon > 0$, such that for all $\delta_2 < M_\psi(\epsilon) p_{\min} / K \triangleq \delta_0$, we have

$$\begin{aligned}\sum_{i=1}^{n} p_i \psi\left(\frac{q_i}{p_i}\right) \le \delta_2 K &\Rightarrow \left|\frac{q_i}{p_i} - 1\right| \le \epsilon, \forall i \\ &\Rightarrow \left|R_2\left(\frac{q_i}{p_i}\right)\right| \le \frac{\psi''(1)}{4}, \forall i.\end{aligned}$$

Hence, for all $\delta_2 \leq \delta_0$, we have

$$\begin{aligned}\delta_2 K \geq \sum_{i=1}^n p_i \psi\left(\frac{q_i}{p_i}\right) &= \sum_{i=1}^n p_i \left(\frac{q_i}{p_i} - 1\right)^2 \left[\frac{\psi''(1)}{2} + R_2\left(\frac{q_i}{p_i}\right)\right] \\ &\geq \left[\frac{\psi''(1)}{2} - \frac{\psi''(1)}{4}\right] \sum_{i=1}^n p_i \left(\frac{q_i}{p_i} - 1\right)^2 \\ &\geq \frac{\psi''(1)}{4} \sum_{i=1}^n p_i \left(\frac{q_i}{p_i} - 1\right)^2,\end{aligned}$$

which gives

$$\sum_{i=1}^n p_i \left(\frac{q_i}{p_i} - 1\right)^2 \leq \delta_2 K \cdot 4\psi''(1).$$

□

Lemma 4.A.2. *The following optimization problem*

$$\inf_{\lambda \in \mathbb{R}} \lambda + \frac{1}{\delta_2} \sum_{i=1}^n p_i \psi^*\left(\delta_2(x_i - \lambda) - \frac{\delta_1 \delta_2}{\psi''(0)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j\right),$$

can be restricted to a compact set of λ.

Proof. Examining the first-order condition of the above infimum gives that the optimal λ^* must satisfy the if and only if condition,

$$\sum_{i=1}^n p_i (\psi^*)'\left(\delta_2\left(x_i - \frac{\delta_1}{\psi''(0)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j - \lambda^*\right)\right) = 1. \tag{4.31}$$

Let b be any number such that $b > \max_{i \in [n]} |x_i - \frac{\delta_1}{\psi''(0)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j|$. Suppose that any optimal λ^* satisfies $\lambda^* \notin [-b, b]$. If $\lambda^* \geq 0$, then we have

$$x_i - \frac{\delta_1}{\psi''(0)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j - \lambda^* \leq x_i - \frac{\delta_1}{\psi''(0)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j - b, \ \forall i = 1, \ldots, n.$$

By the first order condition (4.31) and the monotonicity of $(\psi^*)'$, we thus also have

$$\begin{aligned}&\sum_{i=1}^n p_i (\psi^*)'\left(\delta_2\left(x_i - \frac{\delta_1}{\psi''(0)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j - b\right)\right) \\ &\geq \sum_{i=1}^n p_i (\psi^*)'\left(\delta_2\left(x_i - \frac{\delta_1}{\psi''(0)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j - \lambda^*\right)\right) = 1.\end{aligned}$$

On the other hand, we have by the definition of b, that

$$x_i - \frac{\delta_1}{\psi''(0)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j - b \leq 0,$$

which implies

$$\sum_{i=1}^{n} p_i (\psi^*)' \left(\delta_2 \left(x_i - \frac{\delta_1}{\psi''(0)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j - b \right) \right) \leq (\psi^*)'(0) = 1.$$

Hence, we have

$$\sum_{i=1}^{n} p_i (\psi^*)' \left(\delta_2 \left(x_i - \frac{\delta_1}{\psi''(0)} \sum_{j=1}^{i} \sum_{k=j}^{n} p_k \Delta x_j - b \right) \right) = 1,$$

which implies that b is optimal. Therefore, we may indeed restrict λ to a compact set $[-b, b]$ with b sufficiently large. □

Proof of Proposition 4.8. Since $\mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) = \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))\delta(n)/(-\phi)_*(\delta(n))$, it is sufficient to examine only the measurability of $\min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))$. For ease of notation, we denote $\delta_n(\omega) = \delta(\boldsymbol{\xi}_1(\omega), \ldots, \boldsymbol{\xi}_n(\omega))$. Let $\Pi(n)$ denote the set of all permutations of the set $\{1, \ldots, n\}$. For any constant $c \in \mathbb{R}$, we have,

$$\begin{aligned}
&\left\{ \omega : \inf_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta_n(\omega)}} \sum_{i=1}^{n} q_i l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) \leq c \right\} \\
&= \bigcap_{n\geq 1} \bigcup_{\mathbf{x}\in\mathcal{X}} \left\{ \omega : \sup_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta_n(\omega)}} \sum_{i=1}^{n} q_i l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) \leq c + \frac{1}{n} \right\} \\
&= \bigcap_{n\geq 1} \bigcup_{\mathbf{x}\in\mathcal{X}} \left\{ \omega : \sum_{i=1}^{n} \frac{(-\phi)_* \left(\delta(\omega) \cdot \frac{n-i+1}{n} \right)}{(-\phi)_*(\delta(\omega))} \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)) \leq c + \frac{1}{n} \right\} \\
&= \bigcap_{n\geq 1} \bigcup_{\sigma\in\Pi(n)} \bigcup_{\mathbf{x}\in\mathcal{X}} \left\{ \omega : \sum_{i=1}^{n} \frac{(-\phi)_* \left(\delta(\omega) \cdot \frac{n-i+1}{n} \right)}{(-\phi)_*(\delta(\omega))} \Delta l(\mathbf{x}, \boldsymbol{\xi}_{\sigma(i)}(\omega)) \leq c + \frac{1}{n} \right\} \\
&\qquad \cap \left\{ \omega : l(\mathbf{x}, \boldsymbol{\xi}_{\sigma(1)}(\omega)) \leq \ldots \leq l(\mathbf{x}, \boldsymbol{\xi}_{\sigma(n)}(\omega)) \right\},
\end{aligned}$$

where the set

$$\begin{aligned}
&\bigcup_{\mathbf{x}\in\mathcal{X}} \left\{ \omega : \sum_{i=1}^{n} \frac{(-\phi)_* \left(\delta(\omega) \cdot \frac{n-i+1}{n} \right)}{(-\phi)_*(\delta(\omega))} \Delta l(\mathbf{x}, \boldsymbol{\xi}_{\sigma(i)}(\omega)) \leq c + \frac{1}{n} \right\} \\
&\qquad \cap \left\{ \omega : l(\mathbf{x}, \boldsymbol{\xi}_{\sigma(1)}(\omega)) \leq \ldots \leq l(\mathbf{x}, \boldsymbol{\xi}_{\sigma(n)}(\omega)) \right\},
\end{aligned}$$

is measurable, since the function $l(x, \tilde{\boldsymbol{\xi}})$ is a Caratheodory function (hence its epigraph

is a measurable set-valued function) and $\mathcal{X}$ is a closed set (see Section 7.2.3 of Shapiro et al. 2009). □

Proof of Theorem 4.9. We show the identity for $\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}$, since the proof for $\mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}$ is similar. Fix $\mathbf{x} \in \mathcal{X}$ and let $l(\mathbf{x}, \boldsymbol{\xi}_{1(\mathbf{x})}) \leq \cdots \leq l(\mathbf{x}, \boldsymbol{\xi}_{n(\mathbf{x})})$. Following the proof of Theorem 4.2, we have that for all possible realizations $\omega \in \Omega$, that

$$\begin{aligned}
\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))(\omega) &= \frac{1}{\delta(n)(\omega)} \sum_{i=1}^{n} (-\phi)_* \left(\delta(n)(\omega)\frac{n-i+1}{n}\right) \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)) \\
&= \frac{1}{\delta(n)(\omega)} \sum_{i=1}^{n} \left(\delta(n)(\omega)\frac{n-i+1}{n} - \frac{1}{2\phi''(1)} \left(\delta(n)(\omega)\frac{n-i+1}{n}\right)^2 \right. \\
&\quad \left. + \frac{(\phi^*)'''(\tilde{\xi}_n)}{6} \left(\delta(n)(\omega)\frac{n-i+1}{n}\right)^3 \right) \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)) \\
&= \frac{1}{n}\sum_{i=1}^{n} l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) - \frac{\delta(n)(\omega)}{2\phi''(1)} \sum_{i=1}^{n} \left(\frac{n-i+1}{n}\right)^2 \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)) \\
&\quad + \sum_{i=1}^{n} \frac{(\phi^*)'''(\tilde{\xi}_n)(\omega)}{6} \delta^2(n)(\omega) \left(\frac{n-i+1}{n}\right)^3 \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)).
\end{aligned}$$

Hence, we have that $\epsilon_n(\mathbf{x})(\omega) = \sum_{i=1}^{n} \frac{(\phi^*)'''(\tilde{\xi}_n(\omega))}{6} \delta^2(n)(\omega) \left(\frac{n-i+1}{n}\right)^3 \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega))$, where $\tilde{\xi}_n(\omega) \in [0, \delta(n)(\omega)]$. Since ϕ^* is smooth around zero, its third derivative is bounded in a neighborhood of zero. This means that there exists an $\epsilon_0 > 0$, such that for all ω such that $\delta(n)(\omega) \leq \epsilon_0$, there exists a constant $C > 0$ such that $\frac{(\phi^*)'''(\tilde{\xi}_n(\omega))}{6} \leq C$. Hence, for all such ω, we have that

$$\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{\delta(n)(\omega)} |\epsilon_n(\mathbf{x})(\omega)| &\leq C \cdot \delta(n)(\omega) \sup_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^{n} \frac{n-i+1}{n} \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)) \\
&= C \cdot \delta(n)(\omega) \sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) \\
&\leq C \cdot \delta(n)(\omega) \frac{1}{n} \sum_{i=1}^{n} M_2(\boldsymbol{\xi}_i(\omega)),
\end{aligned}$$

where the latter inequalities follow from the assumption. Now given any $\tilde{\epsilon} > 0$, we have the following inequalities (note that $\epsilon_n(\mathbf{x})$ is measurable since it inherits the measurability of $\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}$):

$$\begin{aligned}
\mathbb{P}_0 \left(\sup_{\mathbf{x}\in f\mathcal{X}} \frac{1}{\delta(n)} |\epsilon_n(\mathbf{x})| > \tilde{\epsilon} \right) &\leq \mathbb{P}_0 \left(\left\{ \sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{\delta(n)} |\epsilon_n(\mathbf{x})| > \tilde{\epsilon} \right\} \cap \{\delta(n) \leq \epsilon_0\} \right) \\
&\quad + \mathbb{P}_0 \left(\left\{ \sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{\delta(n)} |\epsilon_n(\mathbf{x})| > \tilde{\epsilon} \right\} \cap \{\delta(n) > \epsilon_0\} \right).
\end{aligned}$$

Since $\delta(n) \xrightarrow{P} 0$, we have that the second term vanishes as $n \to \infty$. Hence, we only examine the first term, and we have that

$$\begin{aligned}\mathbb{P}_0^*\left(\left\{\sup_{\mathbf{x}\in\mathcal{X}}\frac{1}{\delta(n)}|\epsilon_n(\mathbf{x})| > \tilde{\epsilon}\right\}\cap\{\delta(n)\le\epsilon_0\}\right) &\le \mathbb{P}_0\left(\left\{\delta(n)\frac{1}{n}\sum_{i=1}^n M_2(\boldsymbol{\xi}_i) > \frac{\tilde{\epsilon}}{C}\right\}\cap\{\delta(n)\le\epsilon_0\}\right)\\ &\le \mathbb{P}_0\left(\delta(n)\frac{1}{n}\sum_{i=1}^n M_2(\boldsymbol{\xi}_i) > \frac{\tilde{\epsilon}}{C}\right).\end{aligned}$$

By assumption, $\mathbb{E}_{\mathbb{P}_0}[|M_2(\boldsymbol{\xi})| < \infty$. Hence, the law of large number and $\delta(n) \xrightarrow{P} 0$ imply the convergence in probability $\delta(n)\frac{1}{n}\sum_{i=1}^n M_2(\boldsymbol{\xi}_i) \xrightarrow{P} 0$, which then implies that $\sup_{\mathbf{x}\in\mathcal{X}}|\epsilon_n(\mathbf{x})|/\delta(n) \xrightarrow{P} 0$. □

Proof of Theorem 4.10. Let $\delta(n) = \sqrt{r/n}$. Using Theorem 4.9, we have,

$$\begin{aligned}&\sqrt{n}\left(\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right)\\ &= \sqrt{n}\left(\mathbb{E}_{\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right) - \frac{\sqrt{r}}{2\phi''(1)}\mathrm{dm}_{2,\hat{\mathbb{P}}_n}(\mathbf{x}) + \sqrt{n}\epsilon_n(\mathbf{x}),\end{aligned}$$

where $\sup_{\mathbf{x}\in\mathcal{X}}\sqrt{n}|\epsilon_n(\mathbf{x})| \xrightarrow{P^*} 0$. Since $\mathcal{H}$ is $\mathbb{P}_0$-Donsker, we have that together with Lemma 4.A.8, the following weak convergence of processes indexed by $\mathbf{x}\in\mathcal{X}$

$$\sqrt{n}\left(\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\cdot,\boldsymbol{\xi})) - \mathbb{E}_{\mathbb{P}_0}[l(\cdot,\boldsymbol{\xi})]\right) \rightsquigarrow G(\cdot) - \frac{\sqrt{r}}{2\phi''(1)}\mathrm{dm}_{2,\mathbb{P}_0}(\cdot), \text{ in } l^\infty(\mathcal{H}).$$

Similarly, it follows from Theorem 4.9 that we also have the weak convergence

$$\sqrt{n}\left(\mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\cdot,\boldsymbol{\xi})) - \mathbb{E}_{\mathbb{P}_0}[l(\cdot,\boldsymbol{\xi})]\right) \rightsquigarrow G(\cdot) + \frac{\sqrt{r}}{2\phi''(1)}\overline{\mathrm{m}}_{2,\mathbb{P}_0}(\cdot), \text{ in } l^\infty(\mathcal{H}).$$

Adopting the same notations as in Duchi et al. [2021], we let $T(\mathbb{P}) \triangleq \min_{\mathbf{x}\in\mathcal{X}}\mathbb{E}_{\mathbb{P}}[l(\mathbf{x},\boldsymbol{\xi})]$. Using Lemma 4.6, we have that

$$\begin{aligned}&\min_{\mathbf{x}\in\mathcal{X}}\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}}\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\\ &\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta(n)}}\frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}}\sum_{i=1}^n q_i l(\mathbf{x},\boldsymbol{\xi}_i) - \min_{\mathbf{x}\in\mathcal{X}}\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\\ &= \max_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta(n)}}\min_{\mathbf{x}\in\mathcal{X}}\frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}}\sum_{i=1}^n q_i l(\mathbf{x},\boldsymbol{\xi}_i) - \min_{\mathbf{x}\in\mathcal{X}}\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\\ &= \max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)}\frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}}T(\mathbb{Q}) - T(\mathbb{P}_0),\end{aligned}$$

where we interchanged the minimum and maximum by applying the minimax theorem,

which holds since we have a convex-concave function in $(\mathbf{x}, \mathbf{q})$ on compact feasible sets. Moreover, we identified the probability vector $\mathbf{q} \in \mathcal{M}_{\delta(n)}^{\phi}$ (as defined in (4.19)) with a measure $\mathbb{Q} = \sum_{i=1}^{n} q_i \iota_{\boldsymbol{\xi}_i}$ and we denote the set

$$\mathcal{M}_{\delta(n)}^{\phi}(\hat{\mathbb{P}}_n) \triangleq \left\{ \sum_{i=1}^{n} q_i \iota_{\boldsymbol{\xi}_i} \;\middle|\; \mathbf{q} \in \mathcal{M}_{\delta(n)}^{\phi} \right\},$$

where $\iota_{\boldsymbol{\xi}_i}$ denotes a Dirac measure.

Denote the influence function $IF(\mathbf{x}, \mathbb{P}_0) = l(\mathbf{x}, \boldsymbol{\xi}) - \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}, \boldsymbol{\xi})]$ (see Lemma 17 of Duchi et al. 2021). Following the proofs outlined by Duchi et al. [2021], we start by examining

$$\begin{aligned}
&\frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \left| \max_{\mathbb{Q} \in \mathcal{M}_{\delta(n)}^{\phi}(\hat{\mathbb{P}}_n)} (T(\mathbb{Q}) - T(\mathbb{P}_0)) - \max_{\mathbb{Q} \in \mathcal{M}_{\delta(n)}^{\phi}(\hat{\mathbb{P}}_n)} \inf_{\mathbf{x} \in \mathcal{X}_{\mathbb{P}_0}^*} \mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x}, \mathbb{P}_0)] \right| \\
&\leq \sup_{\mathbb{Q} \in \mathcal{M}_{\delta(n)}^{\phi}(\hat{\mathbb{P}}_n)} \left| \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \left(T(\mathbb{Q}) - T(\mathbb{P}_0) - \inf_{\mathbf{x} \in \mathcal{X}_{\mathbb{P}_0}^*} \mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x}, \mathbb{P}_0)] \right) \right| \\
&\triangleq \sup_{\mathbb{Q} \in \mathcal{M}_{\delta(n)}^{\phi}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})|.
\end{aligned}$$

Our first goal is to show that for any $\epsilon > 0$,

$$\limsup_{n \to \infty} \mathbb{P}_0^* \left(\sqrt{n} \sup_{\mathbb{Q} \in \mathcal{M}_{\delta(n)}^{\phi}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})| \geq \epsilon \right) = 0. \tag{4.32}$$

To do this, we choose for any $\delta' > 0$, a sequence of measurable selection (see Lemma 4.A.6) $\mathbb{Q}_n \in \mathcal{M}_{\delta(n)}^{\phi}(\hat{\mathbb{P}}_n)$, such that $|\kappa(\mathbb{Q}_n)| \geq (1 - \delta') \sup_{\mathbb{Q} \in \mathcal{M}_{\delta(n)}^{\phi}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})|$. This gives the bound,

$$\mathbb{P}_0^* \left(\sqrt{n} \sup_{\mathbb{Q} \in \mathcal{M}_{\delta(n)}^{\phi}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})| \geq \epsilon \right) \leq \mathbb{P}_0^* \left(\sqrt{n} |\kappa_n(\mathbb{Q}_n)| \geq (1 - \delta')\epsilon \right).$$

Therefore, it remains to show that $\sqrt{n}\kappa_n(\mathbb{Q}_n) \overset{P^*}{\to} 0$. Viewing $\mathbb{Q}_n$ as a mapping $\mathbb{Q}_n : \Omega \to l^\infty(\mathcal{H})$, we have that by Lemma 4.A.5, that all subsequences of $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)$ are asymptotically tight and measurable. It follows from Prohorov's theorem (van der Vaart and Wellner 2023) that every subsequence of $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)$ has a further subsequence that converges to a weak limit. Hence, we may choose a subsequence such that $\sqrt{n(m)}(\mathbb{Q}_{n(m)} - \mathbb{P}_0) \rightsquigarrow Z$ for a tight weak limit Z and that

$$\limsup_{n \to \infty} \mathbb{P}_0^* \left(\sqrt{n} |\kappa_n(\mathbb{Q}_n)| \geq (1 - \delta')\epsilon \right) = \lim_{m \to \infty} \mathbb{P}_0^* \left(\sqrt{n(m)} |\kappa_{n(m)}(\mathbb{Q}_{n(m)})| \geq (1 - \delta')\epsilon \right) = 0,$$

where the latter is obtained by applying the functional delta theorem (Theorem 1, Römisch

2006) to conclude that

$$\sqrt{n(m)}\left(T(\mathbb{Q}_{n(m)}) - T(\mathbb{P}_0) - \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \mathbb{E}_{\mathbb{Q}_{n(m)}}[IF(\mathbf{x},\mathbb{P}_0)]\right) \overset{P^*}{\to} 0.$$

Then, we have that

$$\begin{aligned}
&\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right)\\
&= \sqrt{n}\left(\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} T(\mathbb{Q}) - T(\mathbb{P}_0)\right)\\
&= \sqrt{n}\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \left(T(\mathbb{Q}) - T(\mathbb{P}_0)\right) + \sqrt{n}\left(\frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} - 1\right) T(\mathbb{P}_0)\\
&\overset{(*)}{=} \sqrt{n}\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x},\mathbb{P}_0)] - \frac{\sqrt{r}}{2\phi''(1)} T(\mathbb{P}_0) + o_p(1)\\
&\overset{(**)}{=} \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \sqrt{n}\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x},\mathbb{P}_0)] - \frac{\sqrt{r}}{2\phi''(1)} T(\mathbb{P}_0) + o_p(1)\\
&= \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \sqrt{n}\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \left(\mathbb{E}_{\mathbb{Q}}[l(\mathbf{x},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right) - \frac{\sqrt{r}}{2\phi''(1)} T(\mathbb{P}_0) + o_p(1)\\
&= \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \sqrt{n}\left(\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \mathbb{E}_{\mathbb{Q}}[l(\mathbf{x},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right)\\
&\qquad + \sqrt{n}\left(1 - \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}}\right) \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] - \frac{\sqrt{r}}{2\phi''(1)} T(\mathbb{P}_0) + o_p(1),
\end{aligned}$$

where for $(*)$ we used $\sqrt{n}\left(\frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} - 1\right) = -\frac{\sqrt{r}}{2\phi''(1)} + o(1)$ and (4.32), and for $(**)$ we used the minimax theorem to interchange the sup-inf. Since for any $\mathbf{x} \in \mathcal{X}^*_{\mathbb{P}_0}$, we have $\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] = T(\mathbb{P}_0)$, we have that the above is equal to

$$\begin{aligned}
&\inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \sqrt{n}\left(\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \mathbb{E}_{\mathbb{Q}}[l(\mathbf{x},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right) + \frac{\sqrt{r}}{2\phi''(1)} T(\mathbb{P}_0)\\
&\qquad - \frac{\sqrt{r}}{2\phi''(1)} T(\mathbb{P}_0) + o_p(1)\\
&\rightsquigarrow \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} G(\mathbf{x}) - \frac{\sqrt{r}}{2\phi''(1)} \mathrm{dm}_{2,\mathbb{P}_0}(\mathbf{x}),
\end{aligned}$$

where for the latter convergence we also used that infimum operator is continuous with respect to the supremum norm, and thus we may apply the continuous mapping theorem. Indeed, for any $f, g \in l^\infty(\mathcal{H})$ such that $\sup_{h\in\mathcal{H}} |(f-g)(h)| < \epsilon$, we have that $|\inf_{h\in\mathcal{H}^*} f(h) - \inf_{h\in\mathcal{H}^*} g(h)| \le \sup_{h\in\mathcal{H}} |(f-g)(h)| \le \epsilon$, where $\mathcal{H}^* = \{l(x,.) : x \in \mathcal{X}^*_{\mathbb{P}_0}\} \subset$

$\mathcal{H}$.

Similarly, we have

$$\begin{aligned}
&\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}}\mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))-\min_{\mathbf{x}\in\mathcal{X}}\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right)\\
&=\sqrt{n}\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)}T(\mathbb{Q})-T(\mathbb{P}_0)\\
&=\sqrt{n}\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)}\inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}}\mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x},\mathbb{P}_0)]+o_p(1)\\
&=\inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}}\sqrt{n}\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)}\mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x},\mathbb{P}_0)]+o_p(1)\\
&=\inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}}\sqrt{n}\left(\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)}\mathbb{E}_{\mathbb{Q}}[l(\mathbf{x},\boldsymbol{\xi})]-\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right)+o_p(1)\\
&\rightsquigarrow\inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}}G(\mathbf{x})+\frac{\sqrt{r}}{2\phi''(1)}\overline{\mathbb{m}}_{2,\mathbb{P}_0}(\mathbf{x}).
\end{aligned}$$

□

The proof of the above theorems is supplemented with a couple of technical Lemmas.

Lemma 4.A.3. *Let ϕ be four-times continuously differentiable around 1 and $\mathbf{q}\in\mathcal{M}^{\phi}_{\delta(n)}$, where $\mathcal{M}^{\phi}_{\delta(n)}$ is the set defined in (4.19). Suppose $\delta(n)=\sqrt{r/n}$. Then, we have that for any $i=,1\ldots,n$,*

$$n\sqrt{n}\left|q_i-\frac{1}{n}\right|\leq\frac{\sqrt{r}}{2\phi''(1)}+o(1).$$

Proof. Fix $i\in[n]$. By definition of $\mathcal{M}^{h_{\delta(n)}}(\hat{\mathbf{p}})$ (see definition in (4.17)), we have that

$$\begin{aligned}
q_i\leq h_{\delta(n)}(1/n)=\frac{(-\phi)_*(\delta(n)1/n)}{(-\phi)_*(\delta(n))}=\frac{(-\phi)_*\left(\sqrt{\frac{r}{n}}\frac{1}{n}\right)}{(-\phi)_*\left(\sqrt{\frac{r}{n}}\right)}&=\frac{\sqrt{\frac{r}{n}}\frac{1}{n}-\frac{1}{2\phi''(1)}\frac{r}{n}\frac{1}{n^2}+O(\frac{r\sqrt{r}}{n^4\sqrt{n}})}{\sqrt{\frac{r}{n}}-\frac{1}{2\phi''(1)}\frac{r}{n}+O(\frac{r}{n}\sqrt{\frac{r}{n}})}\\
&=\frac{\frac{1}{n}-\frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\frac{1}{n^2}+O(\frac{r}{n^4})}{1-\frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}+O(\frac{r}{n})}.
\end{aligned}$$

Hence,

$$\begin{aligned}
q_i-\frac{1}{n}&\leq\frac{1}{1-\frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}+O(\frac{r}{n})}\left(\frac{1}{2\phi''(1)}\frac{1}{n}\sqrt{\frac{r}{n}}\left(1-\frac{1}{n}\right)+O\left(\frac{r}{n^2}\right)\right)\\
&=\frac{\sqrt{r}}{2\phi''(1)}\frac{1}{n\sqrt{n}}+o\left(\frac{1}{n\sqrt{n}}\right).
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}
q_i = 1 - \sum_{j\neq i} q_j &\geq 1 - h_{\delta(n)}\left(\sum_{j\neq i} 1/n\right) \\
&= 1 - \frac{(-\phi)_*\left(\sqrt{\frac{r}{n}}\left(1-\frac{1}{n}\right)\right)}{(-\phi)_*\left(\sqrt{\frac{r}{n}}\right)} \\
&= 1 - \frac{\sqrt{\frac{r}{n}}\left(1-\frac{1}{n}\right) - \frac{1}{2\phi''(1)}\frac{r}{n}\left(1-\frac{1}{n}\right)^2 - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n}\sqrt{\frac{r}{n}}\left(1-\frac{1}{n}\right)^3 + o\left(\frac{r^2}{n^2}\right)}{\sqrt{\frac{r}{n}} - \frac{1}{2\phi''(1)}\frac{r}{n} - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n}\sqrt{\frac{r}{n}} + o\left(\frac{r^2}{n^2}\right)} \\
&= 1 - \frac{\left(1-\frac{1}{n}\right) - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\left(1-\frac{1}{n}\right)^2 - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n}\left(1-\frac{1}{n}\right)^3 + o\left(\frac{r^2}{n\sqrt{n}}\right)}{1 - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}} - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n} + o\left(\frac{r^2}{n\sqrt{n}}\right)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
q_i - \frac{1}{n} &\geq \left(1-\frac{1}{n}\right) - \frac{\left(1-\frac{1}{n}\right) - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\left(1-\frac{1}{n}\right)^2 - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n}\left(1-\frac{1}{n}\right)^3 + o\left(\frac{r^2}{n\sqrt{n}}\right)}{1 - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}} - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n} + o\left(\frac{r^2}{n\sqrt{n}}\right)} \\
&= \frac{1}{1+o(1)}\left(-\frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\frac{1}{n}\left(1-\frac{1}{n}\right) - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n}\left(1-\frac{1}{n}\right)\left(1-\left(1-\frac{1}{n}\right)^2\right)\right. \\
&\qquad \left. + o\left(\frac{r^2}{n\sqrt{n}}\right)\right) \\
&= (1+o(1))\left(-\frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\frac{1}{n}\left(1-\frac{1}{n}\right) - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n^2}\left(1-\frac{1}{n}\right)\left(2-\frac{1}{n}\right)\right. \\
&\qquad \left. + o\left(\frac{r^2}{n\sqrt{n}}\right)\right) \\
&= (1+o(1))\left(-\frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\frac{1}{n}\left(1-\frac{1}{n}\right) + o\left(\frac{1}{n\sqrt{n}}\right)\right) \\
&= -\frac{\sqrt{r}}{2\phi''(1)}\frac{1}{n\sqrt{n}} + o\left(\frac{1}{n\sqrt{n}}\right).
\end{aligned}$$

Hence, we have shown that

$$n\sqrt{n}\left|q_i - \frac{1}{n}\right| \leq \frac{\sqrt{r}}{2\phi''(1)} + o(1).$$

□

Lemma 4.A.4. *Let ϕ be four times continuously differentiable around 1. Suppose $\delta(n) = \sqrt{r(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)/n}$ where $r(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n) \xrightarrow{P} r_0$ for some $r_0 > 0$ and $r(\boldsymbol{\xi}_1(\omega),\ldots,\boldsymbol{\xi}_n(\omega)) >$*

$0, \forall \omega \in \Omega, \forall n \geq 1$. *Then, we have that any* $\mathbf{q} \in \mathcal{M}_{\delta(n)}^{\phi}$ *satisfies*

$$n\sqrt{n}\left|q_i - \frac{1}{n}\right| \leq \frac{\sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}}{2\phi''(1)} + o_p(1).$$

Proof. Following the proof of lemma 4.A.3, we have that

$$n\sqrt{n}\left(q_i - \frac{1}{n}\right) \leq \frac{1}{1+o_p(1)}\left(\frac{\sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}}{2\phi''(1)}\left(1-\frac{1}{n}\right) + O\left(\frac{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}{\sqrt{n}}\right)\right),$$

and that

$$\begin{aligned} & n\sqrt{n}\left(q_i - \frac{1}{n}\right) \\ & \geq \frac{1}{1+o_p\left(1\right)}\left(-\frac{\sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}}{2\phi''(1)}\left(1-\frac{1}{n}\right) - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}{\sqrt{n}}\left(1-\frac{1}{n}\right)\left(2-\frac{1}{n}\right)\right. \\ & \qquad \left. + o_p(1)\right). \end{aligned}$$

Hence, an application of continuous mapping theorem and Slutsky's theorem gives that

$$n\sqrt{n}\left|q_i - \frac{1}{n}\right| \leq \frac{\sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}}{2\phi''(1)} + o_p(1).$$

□

Recall that we defined the function class $\mathcal{H} = \{\mathbf{x} \in \mathcal{X} : l(\mathbf{x}, \boldsymbol{\xi}) : \Omega \to \mathbb{R}\}$. For any $h \in \mathcal{H}$, we denote $h(\boldsymbol{\xi}) = l(\mathbf{x}, \boldsymbol{\xi})$.

Lemma 4.A.5. *Let* $\mathcal{H}$ *be* $\mathbb{P}_0$*-Donsker with* L^2*-integrable envelope* $M_2 : \Omega \to \mathbb{R}$*, i.e.* $\mathbb{E}_{\mathbb{P}_0}[M_2^2(\boldsymbol{\xi})] < \infty$ *and* $|l(\mathbf{x}, \boldsymbol{\xi}(\omega))| \leq M_2(\boldsymbol{\xi}(\omega))$*, for all* $\omega \in \Omega$*. Let* $\delta(n) = \sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)/n}$ *where* $r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n) \xrightarrow{P} r_0$ *for some* $r_0 > 0$ *and* $r(\boldsymbol{\xi}_1(\omega), \ldots, \boldsymbol{\xi}_n(\omega)) > 0, \forall \omega \in \Omega, \forall n \geq 1$*. Then, for any sequence* $\mathbb{Q}_n \in \mathcal{M}_{\delta(n)}^{\phi}(\hat{\mathbb{P}}_n)$*, we have that the sequence of mapping* $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0) : \Omega \to l^\infty(\mathcal{H})$ *is asymptotically tight.*

Proof. By Theorem 1.5.7 of van der Vaart and Wellner [2023], we have that $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)$ is asymptotically tight if and only if (i). the marginal $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)h$ is asymptotically tight for all $h \in \mathcal{H}$, (ii). There exists a semi-metric $\|.\|$ on $\mathcal{H}$ such that $(\mathcal{H}, \|.\|)$ is totally bounded, and (iii) $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)$ is asymptotically uniformly equicontinuous in probability (with respect to the semi-metric $\|.\|$), which means that for all $\epsilon > 0$, we have (where $\mathbb{P}_0^*$ denotes the outer probability)

$$\limsup_{\delta \to 0, n \to \infty} \mathbb{P}_0^*\left(\sup_{\|h-h'\|<\delta} \left|\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)(h-h')\right| > \epsilon\right) = 0.$$

We note that for any semi-metric $\|.\|$ on $\mathcal{H}$, we have

$$\begin{aligned}
&\limsup_{\delta\to 0,n\to\infty} \mathbb{P}_0^*\left(\sup_{\|h-h'\|<\delta}\left|\sqrt{n}(\mathbb{Q}_n-\mathbb{P}_0)(h-h')\right|>\epsilon\right)\\
&\leq \limsup_{\delta\to 0,n\to\infty} \mathbb{P}_0^*\left(\sup_{\|h-h'\|<\delta}\left|\sqrt{n}(\mathbb{Q}_n-\hat{\mathbb{P}}_n)(h-h')\right|>\epsilon/2\right)\\
&\quad + \mathbb{P}_0^*\left(\sup_{\|h-h'\|<\delta}\left|\sqrt{n}(\hat{\mathbb{P}}_n-\mathbb{P}_0)(h-h')\right|>\epsilon/2\right).
\end{aligned}$$

By assumption, we have that $\mathcal{H}$ is $\mathbb{P}_0$-Donsker, which means that $\sqrt{n}(\hat{\mathbb{P}}_n-\mathbb{P}_0)\rightsquigarrow\mathbb{G}$, where $\mathbb{G}$ is a tight Gaussian limit. By Example 1.5.10 in van der Vaart and Wellner [2023], this implies that $\sqrt{n}(\hat{\mathbb{P}}_n-\mathbb{P}_0)$ is asymptotically uniformly equicontinuous in probability, with respect to the L_2 semi-metric $\|.\|_2$, and that $(\mathcal{H},\|.\|_2)$ is also totally bounded. Therefore, we may take the L_2 semi-metric and conclude that the second term above vanishes as $\delta\to 0, n\to\infty$. Therefore, we examine only the first term. We have that for any $\epsilon>0$,

$$\begin{aligned}
&\mathbb{P}_0^*\left(\sup_{\|h-h'\|<\delta}\left|\sqrt{n}(\mathbb{Q}_n-\hat{\mathbb{P}}_n)(h-h')\right|>\epsilon\right)\\
&=\mathbb{P}_0^*\left(\sup_{\|h-h'\|<\delta}\left|\sqrt{n}\sum_{i=1}^n(q_{i,n}-\frac{1}{n})(h(\boldsymbol{\xi}_i)-h'(\boldsymbol{\xi}_i))\right|>\epsilon\right)\\
&\leq\mathbb{P}_0^*\left(\sup_{\|h-h'\|<\delta}\left|n\sqrt{\sum_{i=1}^n(q_{i,n}-\frac{1}{n})^2}\sqrt{\frac{1}{n}\sum_{i=1}^n(h(\boldsymbol{\xi}_i)-h'(\boldsymbol{\xi}_i))^2}\right|>\epsilon\right).
\end{aligned}$$

Now, since $\mathbb{Q}_n\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)$, we have by Lemma 4.A.4 that for all $\omega\in\Omega$,

$$\begin{aligned}
\sum_{i=1}^n\left(q_{i,n}(\omega)-\frac{1}{n}\right)^2 &\leq \sum_{i=1}^n\frac{1}{n^3}\left(\frac{\sqrt{r(\boldsymbol{\xi}_1(\omega),\ldots,\boldsymbol{\xi}_n(\omega))}}{2\phi''(1)}+e_n(\omega)\right)^2\\
&=\frac{1}{n^2}\left(\frac{\sqrt{r(\boldsymbol{\xi}_1(\omega),\ldots,\boldsymbol{\xi}_n(\omega))}}{2\phi''(1)}+e_n(\omega)\right)^2\\
&=\frac{1}{n^2}\left(\frac{r(\boldsymbol{\xi}_1(\omega),\ldots,\boldsymbol{\xi}_n(\omega))}{4\phi''(1)^2}+\tilde{e}_n(\omega)\right),
\end{aligned}$$

where $e_n,\tilde{e}_n\xrightarrow{P}0$. This gives that

$$n\sqrt{\sum_{i=1}^n\left(q_{i,n}(\omega)-\frac{1}{n}\right)^2}\leq\sqrt{\frac{r(\boldsymbol{\xi}_1(\omega),\ldots,\boldsymbol{\xi}_n(\omega))}{4\phi''(1)^2}+\tilde{e}_n(\omega)}.$$

Since the right-hand side is a random variable that converges in probability to $\sqrt{r_0}/2\phi''(1)$, it is uniformly tight. Hence, for any $\tilde{\epsilon}>0$, there exists a constant $M>0$ such that for

all $n \geq 1$, we have

$$\mathbb{P}_0^*\left(n\sqrt{\sum_{i=1}^n\left(q_{i,n}-\frac{1}{n}\right)^2} \leq M\right) \geq \mathbb{P}_0\left(\sqrt{\frac{r(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}{4\phi''(1)^2}+\tilde{e}_n} \leq M\right) \geq 1-\tilde{\epsilon}.$$

Hence, we have that for any $\tilde{\epsilon} > 0$,

$$\begin{aligned}
&\mathbb{P}_0^*\left(\sup_{\|h-h'\|<\delta}\left|n\sqrt{\sum_{i=1}^n(q_{i,n}-\frac{1}{n})^2}\sqrt{\frac{1}{n}\sum_{i=1}^n(h(\boldsymbol{\xi}_i)-h'(\boldsymbol{\xi}_i))^2}\right| > \epsilon\right) \\
&\leq \mathbb{P}_0^*\left(\left\{\sup_{\|h-h'\|<\delta}\left|\sqrt{\frac{1}{n}\sum_{i=1}^n(h(\boldsymbol{\xi}_i)-h'(\boldsymbol{\xi}_i))^2}\right| > \epsilon/M\right\} \cap \left\{n\sqrt{\sum_{i=1}^n(q_{i,n}-\frac{1}{n})^2} \leq M\right\}\right) + \tilde{\epsilon} \\
&\leq \mathbb{P}_0^*\left(\sup_{\|h-h'\|<\delta}\left|\sqrt{\frac{1}{n}\sum_{i=1}^n(h(\boldsymbol{\xi}_i)-h'(\boldsymbol{\xi}_i))^2}\right| > \epsilon/M\right) + \tilde{\epsilon}.
\end{aligned}$$

Finally, the above outer probability vanishes as $n \to \infty, \delta \to 0$, due to the assumption that $\mathcal{H}$ is $\mathbb{P}_0$-Donsker and that we may take $\|.\|$ to be the L_2 semi-metric. This proves the equi-continuity.

Now let $h \in \mathcal{H}$. It only remains to show that the marginal $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)h$ is also asymptotically tight. Since $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)h$ is a real-valued sequence, it is sufficient to show that for any $\epsilon > 0$, there exists a constant $M_0 > 0$, such that $\liminf_{n\to\infty}(\mathbb{P}_0)_*(|\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)h| \leq M_0) \geq 1 - \epsilon$. We have that

$$\begin{aligned}
\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)h &= \sqrt{n}(\mathbb{Q}_n - \hat{\mathbb{P}}_n + \hat{\mathbb{P}}_n - \mathbb{P}_0)h \\
&= \sqrt{n}\sum_{i=1}^n\left(q_{i,n}-\frac{1}{n}\right)h(\boldsymbol{\xi}_i) + \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n h(\boldsymbol{\xi}_i) - \mathbb{E}_{\mathbb{P}_0}[h(\boldsymbol{\xi})]\right).
\end{aligned}$$

We examine the first sequence $\sqrt{n}\sum_{i=1}^n\left(q_{i,n}-\frac{1}{n}\right)h(\boldsymbol{\xi}_i)$. Invoking Lemma 4.A.4 again gives that for all $\omega \in \Omega$,

$$\left|\sqrt{n}\sum_{i=1}^n\left(q_{i,n}(\omega)-\frac{1}{n}\right)h(\boldsymbol{\xi}_i(\omega))\right| \leq \left(\frac{\sqrt{r(\boldsymbol{\xi}_1(\omega),\ldots,\boldsymbol{\xi}_n(\omega))}}{2\phi''(1)} + e_n(\omega)\right)\frac{1}{n}\sum_{i=1}^n|h(\boldsymbol{\xi}_i(\omega))|.$$

Hence, we have that for all ω:

$$\begin{aligned}
\sqrt{n}|(\mathbb{Q}_n(\omega) - \mathbb{P}_0)h| \leq &\left(\frac{\sqrt{r(\boldsymbol{\xi}_1(\omega),\ldots,\boldsymbol{\xi}_n(\omega))}}{2\phi''(1)} + e_n(\omega)\right)\frac{1}{n}\sum_{i=1}^n|h(\boldsymbol{\xi}_i(\omega))| \\
&+ \sqrt{n}\left|\frac{1}{n}\sum_{i=1}^n h(\boldsymbol{\xi}_i) - \mathbb{E}_{\mathbb{P}_0}[h(\boldsymbol{\xi})]\right|
\end{aligned}$$

Now, since $\mathbb{E}_{\mathbb{P}_0}|h(\boldsymbol{\xi})| < \infty$ due to $\mathbb{E}_{\mathbb{P}_0}h^2(\boldsymbol{\xi}) < \infty$, we have that

$$\left(\frac{\sqrt{r(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}}{2\phi''(1)} + e_n\right)\frac{1}{n}\sum_{i=1}^{n}|h(\boldsymbol{\xi}_i)| \xrightarrow{P} \frac{\sqrt{r_0}}{2\phi''(1)}\mathbb{E}_{\mathbb{P}_0}|h(\boldsymbol{\xi})|.$$

Furthermore, the central limit theorem and continuous mapping theorem imply that

$$\sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n}h(\boldsymbol{\xi}_i) - \mathbb{E}_{\mathbb{P}_0}[h(\boldsymbol{\xi})]\right| \rightsquigarrow |N(0,\mathrm{Var}_{\mathbb{P}_0}(h(\boldsymbol{\xi})))|.$$

Hence, Slutsky's theorem implies that

$$\begin{aligned}&\left(\frac{\sqrt{r(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}}{2\phi''(1)} + e_n\right)\frac{1}{n}\sum_{i=1}^{n}|h(\boldsymbol{\xi}_i)| + \sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n}h(\boldsymbol{\xi}_i) - \mathbb{E}_{\mathbb{P}_0}[h(\boldsymbol{\xi})]\right| \\ &\rightsquigarrow |N(0,\mathrm{Var}_{\mathbb{P}_0}(h(\boldsymbol{\xi})))| + \frac{\sqrt{r_0}}{2\phi''(1)}\mathbb{E}_{\mathbb{P}_0}|h(\boldsymbol{\xi})|.\end{aligned}$$

Hence, the above sequence of measurable functions is weakly convergent, and thus it is a uniformly tight sequence, which means that for any $\epsilon > 0$, there exists a constant $\tilde{M} > 0$, such that for all $n \geq 1$,

$$\begin{aligned}&(\mathbb{P}_0)_*\left(\left|\sqrt{n}\sum_{i=1}^{n}\left(q_{i,n} - \frac{1}{n}\right)h(\boldsymbol{\xi}_i)\right| \leq M_0\right) \\ &\geq \mathbb{P}_0\left(\left(\frac{\sqrt{r(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}}{2\phi''(1)} + e_n\right)\frac{1}{n}\sum_{i=1}^{n}|h(\boldsymbol{\xi}_i)| + \sqrt{n}\left|\frac{1}{n}\sum_{i=1}^{n}h(\boldsymbol{\xi}_i) - \mathbb{E}_{\mathbb{P}_0}[h(\boldsymbol{\xi})]\right| \leq M_0\right) \\ &\geq 1 - \epsilon.\end{aligned}$$

Hence, we may also conclude that $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)h$ is asymptotically tight. □

Lemma 4.A.6. *Let $\overline{\boldsymbol{\xi}}_n = (\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)$ be i.i.d. random vectors and let $\delta(\overline{\boldsymbol{\xi}}_n) : \Omega \to \mathbb{R}_{>0}$ be a measurable function. For any $\epsilon > 0$, there exists a measurable selection $\mathbf{q}(\omega) : \Omega \to \mathcal{M}^{\phi}_{\delta(\overline{\boldsymbol{\xi}}_n)}$, such that the associated random measure $\mathbb{Q}_n = \sum_{i=1}^{n} q_i(\omega)\iota_{\boldsymbol{\xi}_i}$ satisfies*

$$|\kappa(\mathbb{Q}_n)| \geq (1-\epsilon)\sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\overline{\boldsymbol{\xi}}_n)}(\hat{\mathbb{P}}_n)}|\kappa(\mathbb{Q})|,$$

for all $\omega \in \Omega$, where $\kappa(\mathbb{Q}) = T(\mathbb{Q}) - T(\mathbb{P}_0) - \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}}\mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x},\mathbb{P}_0)]$, $\mathcal{X}^*_{\mathbb{P}_0} = \mathrm{argmin}_{\mathbf{x}\in\mathcal{X}}\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]$.

Proof. We apply Theorem 4.10 in Rieder [1978]. By definition, we have that $\kappa(\mathbb{Q}) =$

$\inf_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^n q_i l(\mathbf{x}, \boldsymbol{\xi}_i) - \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \sum_{i=1}^n q_i l(\mathbf{x}, \boldsymbol{\xi}_i)$. Hence, we define a function

$$\begin{aligned} u : \Omega \times \mathbb{R}^n &\to \mathbb{R} \\ (\omega, \mathbf{q}) &\mapsto \left| \inf_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^n q_i l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) - \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \sum_{i=1}^n q_i l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) \right|. \end{aligned}$$

We define

$$\mathcal{D} = \left\{ (\omega, \mathbf{q}) \;\middle|\; \sum_{i=1}^n q_i = 1, q_i \geq 0, \sum_{j\in J} q_j \leq \frac{(-\phi)_* \left(\delta(\overline{\boldsymbol{\xi}}_n(\omega)) \cdot \frac{|J|}{n} \right)}{(-\phi)_*(\delta(\overline{\boldsymbol{\xi}}_n)(\omega))}, \forall J \subset [n] \right\}.$$

We note that $\mathcal{D} \in \mathcal{F} \times B(\mathbb{R}^n)$, since all constraint functions in $\mathcal{D}$ are (separable) measurable functions in the variables $(\omega, \mathbf{q})$. Moreover, the section set $\mathcal{D}(\omega) = \{\mathbf{q} \in \mathbb{R}^n : \mathbf{q} \in \mathcal{M}^{\phi}_{\delta(\overline{\boldsymbol{\xi}}_n)(\omega)}\}$ is a subset of $\mathbb{R}^n$ and thus has a countable dense subset, for all $\omega \in \Omega$. Finally, we note that u is measurable in ω for each $\mathbf{q}$, and continuous in $\mathbf{q}$ for each ω. Indeed, measurability follows from the fact that l is a Carathéodory function. To show continuity, it suffices to show that for each ω, the function $\mathbf{q} \mapsto \inf_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^n q_i l(\mathbf{x}, \boldsymbol{\xi}_i(\omega))$ is continuous (the case for $\mathbf{q} \mapsto \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \sum_{i=1}^n q_i l(\mathbf{x}, \boldsymbol{\xi}_i(\omega))$ is identical). Indeed,

$$\begin{aligned} &\left| \inf_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^n q_{1,i} l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) - \inf_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^n q_{2,i} l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) \right| \\ &\leq \sup_{\mathbf{x}\in\mathcal{X}} \left| \sum_{i=1}^n (q_{1,i} - q_{2,i}) l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) \right| \\ &\leq \max_{\mathbf{x}\in\mathcal{X}, i=1,\ldots,n} |l(\mathbf{x}, \boldsymbol{\xi}_i(\omega))| \cdot \|\mathbf{q_1} - \mathbf{q_2}\|_1. \end{aligned}$$

Since $\mathcal{X}$ is assumed to be compact and $l(., \boldsymbol{\xi}_i(\omega))$ is continuous for all i, the maximum exists. Therefore, u is a Carathéodory function and thus is measurable with respect to the product sigma-algebra $\mathcal{F} \times B(\mathbb{R}^n)$ (Theorem 7.36, Shapiro et al. 2009). Since $\mathcal{D}$ is a measurable set, it is also measurable with respect to the restricted sigma-algebra imposed on $\mathcal{D}$. Therefore, the conditions of Theorem 4.10 in Rieder [1978] are satisfied, and a measurable selection exists. □

Lemma 4.A.7. *Let $X_1, \ldots, X_n$ be i.i.d. samples of a measurable random variable X with $\mathbb{E}_{\mathbb{P}}[X^2] < \infty$. Denote the order statistics $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. Define $\Delta X_{(1)} = X_{(1)}$, $\Delta X_{(i)} = X_{(i)} - X_{(i-1)}$ for $i = 2, \ldots, n$. Then, we have that*

$$\sum_{i=1}^n \left(\frac{n-i+1}{n} \right)^2 \Delta X_{(i)} \xrightarrow{P} \mathbb{E}_{\mathbb{P}}[\min\{X_1, X_2\}].$$

Proof. We have that

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{n-i+1}{n}\right)^2\Delta X_{(i)} &= \frac{1}{n^2}\sum_{1\le i\le j\le n}\min\{X_i,X_j\}\\ &= \frac{1}{n^2}\sum_{i=1}^{n}X_i+\frac{2}{n^2}\sum_{i<j}\min\{X_i,X_j\}\\ &= \frac{1}{n^2}\sum_{i=1}^{n}X_i+\frac{n-1}{n}\frac{2}{n(n-1)}\sum_{i<j}\min\{X_i,X_j\}.\end{aligned}$$

We note that $\frac{1}{n^2}\sum_{i=1}^{n}X_i \xrightarrow{P} 0$ and $\frac{2}{n(n-1)}\sum_{i<j}\min\{X_i,X_j\} \xrightarrow{P} \mathbb{E}_{\mathbb{P}}[\min\{X_1,X_2\}]$ by Theorem 12.3 of van der Vaart [1998] for U-statistics. Hence, the statement follows. □

We recall the notation of the dual second moment that is evaluated using two i.i.d. random variables $\boldsymbol{\xi}^{(1)},\boldsymbol{\xi}^{(2)}$: $\mathrm{dm}_{2,\mathbb{P}}(\mathbf{x}) = \mathbb{E}_{\mathbb{P}\times\mathbb{P}}[\min\{l(\mathbf{x},\boldsymbol{\xi}^{(1)},l(\mathbf{x},\boldsymbol{\xi}^{(2)}\}]$.

Lemma 4.A.8. *Let $\mathcal{H} = \{l(\mathbf{x},.) \mid \mathbf{x}\in\mathcal{X}\}$ be a Donsker class with a square integrable envelope M_2. Then, by viewing the second dual moment $\mathrm{dm}_{2,\mathbb{P}_0}(.), \mathrm{dm}_{2,\hat{\mathbb{P}}_n}(.)$ as a mapping in $l^\infty(\mathcal{H})$, we have that*

$$\mathrm{dm}_{2,\hat{\mathbb{P}}_n}(.) \rightsquigarrow \mathrm{dm}_{2,\mathbb{P}_0}(.), \text{ in } l^\infty(\mathcal{H}). \tag{4.33}$$

Proof. Since $\mathcal{H}$ is Donsker, we have that by viewing the empirical measure $\hat{\mathbb{P}}_n$ as a mapping in $l^\infty(\mathcal{H})$, that

$$\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_0) \rightsquigarrow \mathbb{G},$$

where $\mathbb{G}$ is a tight, measurable Gaussian process in $l^\infty(\mathcal{H})$. By Example 1.5.10 of van der Vaart and Wellner [2023], it follows that under the L^1 semi-metric $\rho_{1,\mathbb{P}_0}(\mathbf{x},\mathbf{y}) = \mathbb{E}_{\mathbb{P}_0}[|l(\mathbf{x},\boldsymbol{\xi}) - l(\mathbf{y},\boldsymbol{\xi})|]$, $\mathcal{H}$ is totally bounded. We now show (4.33) using Theorem 1.5.4 of van der Vaart and Wellner [2023]. The convergence of the marginals follows from Lemma 4.A.7. To establish asymptotic tightness, we use Theorem 1.5.7 of van der Vaart and Wellner [2023] and show that $\mathrm{dm}_{2,\hat{\mathbb{P}}_n}(.)$ is asymptotically uniformly ρ_1-equicontinuous in

probability. Indeed, we note that for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$

$$\begin{aligned}
&\left|\mathrm{dm}_{2,\hat{\mathbb{P}}_n}(\mathbf{x}) - \mathrm{dm}_{2,\hat{\mathbb{P}}_n}(\mathbf{y})\right| \\
&= \left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \min\{l(\mathbf{x}, \boldsymbol{\xi}_i), l(\mathbf{x}, \boldsymbol{\xi}_j)\} - \min\{l(\mathbf{y}, \boldsymbol{\xi}_i), l(\mathbf{y}, \boldsymbol{\xi}_j)\}\right| \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \max\left\{|l(\mathbf{x}, \boldsymbol{\xi}_i) - l(\mathbf{y}, \boldsymbol{\xi}_i)|, |l(\mathbf{x}, \boldsymbol{\xi}_j) - l(\mathbf{y}, \boldsymbol{\xi}_j)|\right\} \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |l(\mathbf{x}, \boldsymbol{\xi}_i) - l(\mathbf{y}, \boldsymbol{\xi}_i)| + |l(\mathbf{x}, \boldsymbol{\xi}_j) - l(\mathbf{y}, \boldsymbol{\xi}_j)| \\
&= \frac{2}{n}\sum_{i=1}^{n} |l(\mathbf{x}, \boldsymbol{\xi}_i) - l(\mathbf{y}, \boldsymbol{\xi}_i)| = 2\rho_{1,\hat{\mathbb{P}}_n}(\mathbf{x}, \mathbf{y}).
\end{aligned}$$

Hence, for any n, ϵ, δ, we have

$$\mathbb{P}^*\left(\sup_{\rho_{1,\mathbb{P}_0}(\mathbf{x},\mathbf{y})<\delta}\left|\mathrm{dm}_{2,\hat{\mathbb{P}}_n}(\mathbf{x}) - \mathrm{dm}_{2,\hat{\mathbb{P}}_n}(\mathbf{y})\right| > \epsilon\right) \leq \mathbb{P}^*\left(\sup_{\rho_{1,\mathbb{P}_0}(\mathbf{x},\mathbf{y})<\delta} \rho_{1,\hat{\mathbb{P}}_n}(\mathbf{x}, \mathbf{y}) > \frac{\epsilon}{2}\right).$$

Now, we note that since $\mathcal{H}$ is Donsker, it is certainly Glivenko-Cantelli. Since Glivenko-Cantelli is preserved under continuous mapping of multiple function classes (Theorem 2.10.5, van der Vaart and Wellner 2023), we have that

$$\sup_{\mathbf{x},\mathbf{y}}\left|\rho_{1,\hat{\mathbb{P}}_n}(\mathbf{x}, \mathbf{y}) - \rho_{1,\mathbb{P}_0}(\mathbf{x}, \mathbf{y})\right| \to 0,$$

in outer probability. Hence, we have that for any $\delta < \epsilon/2$:

$$\begin{aligned}
\mathbb{P}^*\left(\sup_{\rho_{1,\mathbb{P}_0}(\mathbf{x},\mathbf{y})<\delta} \rho_{1,\hat{\mathbb{P}}_n}(\mathbf{x}, \mathbf{y}) > \frac{\epsilon}{2}\right) &= \mathbb{P}^*\left(\sup_{\rho_{1,\mathbb{P}_0}(\mathbf{x},\mathbf{y})<\delta} \rho_{1,\hat{\mathbb{P}}_n}(\mathbf{x}, \mathbf{y}) - \rho_{1,\mathbb{P}_0}(\mathbf{x}, \mathbf{y}) + \rho_{1,\mathbb{P}_0}(\mathbf{x}, \mathbf{y}) > \frac{\epsilon}{2}\right) \\
&\leq \mathbb{P}^*\left(\sup_{\rho_{1,\mathbb{P}_0}(\mathbf{x},\mathbf{y})<\delta} |\rho_{1,\hat{\mathbb{P}}_n}(\mathbf{x}, \mathbf{y}) - \rho_{1,\mathbb{P}_0}(\mathbf{x}, \mathbf{y})| > \frac{\epsilon}{2} - \delta\right) \\
&\leq \mathbb{P}^*\left(\sup_{\mathbf{x},\mathbf{y}} |\rho_{1,\hat{\mathbb{P}}_n}(\mathbf{x}, \mathbf{y}) - \rho_{1,\mathbb{P}_0}(\mathbf{x}, \mathbf{y})| > \frac{\epsilon}{2} - \delta\right) \\
&\overset{n\to\infty}{\to} 0.
\end{aligned}$$

Hence, this establishes the equi-continuity in probability and proves (4.33). □

Lemma 4.A.9. *Let $l(\mathbf{x}, \tilde{\boldsymbol{\xi}})$ be a loss function that is continuous in $\mathbf{x}$, for every $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^{n_{\boldsymbol{\xi}}}$. Assume we have a random vector $\boldsymbol{\xi}$ taking values in $\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_n \in \mathbb{R}^{n_{\boldsymbol{\xi}}}$, with probability $\mathbf{p} = (p_1, \ldots, p_n)$. Then, for any upper-semicontinuous concave distortion function h : [0.1] $\to$ [0, 1] (non-decreasing, $h(0) = 0$, $h(1) = 1$, we have that the rank-dependent*

evaluation

$$\mathcal{R}_{h,\mathbf{p}}(\mathbf{x}) \triangleq \sum_{i=1}^{n} h\left(\sum_{j=i}^{n} p_{j(\mathbf{x})}\right) \Delta l(\mathbf{x}, \tilde{\boldsymbol{\xi}}_{i(\mathbf{x})}),$$

is continuous in $\mathbf{x}$*.*

Proof. Using the dual representation of rank-dependent model (see Denneberg 1994), we have that

$$\mathcal{R}_{h,\mathbf{p}}(\mathbf{x}) = \max_{\mathbf{q}\in\mathcal{M}_h(\mathbf{p})} \sum_{i=1}^{n} q_i l(\mathbf{x}, \tilde{\boldsymbol{\xi}}_i),$$

where

$$\mathcal{M}_h(\mathbf{p}) = \left\{ \mathbf{q} \in \mathbb{R}^n \;\middle|\; \mathbf{q} \geq \mathbf{0}, \sum_{i=1}^{n} q_i = 1, \sum_{i\in J} q_i \leq h\left(\sum_{i\in J} p_i\right), \forall J \subset [n] \right\}.$$

Since h is upper semi-continuous, and $\mathcal{M}_h(\mathbf{p})$ is a set of probability vectors, it follows that $\mathcal{M}_h(\mathbf{p})$ is compact. Moreover, the function $(\mathbf{q}, \mathbf{x}) \mapsto \sum_{i=1}^{n} q_i l(\mathbf{x}, \tilde{\boldsymbol{\xi}}_i)$ is jointly continuous. Hence, continuity of $\mathcal{R}_{h,\mathbf{p}}(\mathbf{x})$ follows from an application of Berge's maximum theorem (Berge 1963). □

4.B Details on Dual DRO Confidence Bounds

To construct dual DRO confidence bounds, we can split the data $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$ in half: $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{n/2})$ and $(\boldsymbol{\xi}'_1 \ldots, \boldsymbol{\xi}'_{n/2})$, and use the first half to obtain a consistent empirical estimator $\hat{\mathbf{x}}_{n/2}$ of the optimal solution $\mathbf{x}^*$. Then, with the second half of the sample, which is independent of the first, we can compute the following empirical estimator

$$\sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)} = \frac{2\Phi^{-1}(\alpha)\phi''(1)\sqrt{\frac{1}{n/2}\sum_{i=1}^{n/2}\left(l(\hat{\mathbf{x}}_{n/2}, \boldsymbol{\xi}'_i) - \frac{1}{n/2}\sum_{i=1}^{n/2} l(\hat{\mathbf{x}}_{n/2}, \boldsymbol{\xi}'_i)\right)^2}}{\sum_{i=1}^{n/2}\left(\frac{n/2+i-1}{n/2}\right)^2 \Delta l(\hat{\mathbf{x}}_{n/2}, \boldsymbol{\xi}'_{(i)})},$$

where $0 < l(\hat{\mathbf{x}}_{n/2}, \boldsymbol{\xi}'_{(1)}) \leq \ldots \leq l(\hat{\mathbf{x}}_{n/2}, \boldsymbol{\xi}'_{(n/2)})$. By the law of large number, Lemma 4.A.7 and continuous mapping theorem, we have that $\sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}$ converges in probability to $\frac{2\Phi^{-1}(\alpha)\phi''(1)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))}}{\mathbb{E}_{\mathbb{P}_0}[\min\{l(\mathbf{x}^*, \boldsymbol{\xi}^{(1)}), l(\mathbf{x}^*, \boldsymbol{\xi}^{(2)})\}]}$. The guarantee of the dual DRO confidence bounds is then formally stated in the following theorem.

Theorem 4.B.10. *Let* $\mathcal{H}$ *be a* $\mathbb{P}_0$*-Donsker class with a square integrable envelope function* M_2*. Assume* ϕ *is four times continuously differentiable in a neighborhood of* 1*. If* $\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}, \boldsymbol{\xi})]$ *has a unique solution* $\mathbf{x}^*$*, and* $\sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}$ *is a consistent estimator that converges in probability to* $\frac{2\Phi^{-1}(\alpha)\phi''(1)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))}}{\mathbb{E}_{\mathbb{P}_0}[\min\{l(\mathbf{x}^*, \boldsymbol{\xi}^{(1)}), l(\mathbf{x}^*, \boldsymbol{\xi}^{(2)})\}]}$*, for any* $\alpha \in (0, 1)$*, and that*

for all $n \geq 1$, $\omega \in \Omega$, *we have* $r(\boldsymbol{\xi}_1(\omega), \ldots, \boldsymbol{\xi}_n(\omega)) > 0$. *Then, for* $\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n) = \sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)/n}$, *we have that the following coverage guarantee:*

$$\lim_{n\to\infty} \mathbb{P}_0 \left(\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] \leq 0 \right) = \alpha.$$

On the other hand, if $\sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}$ *converges in probability to* $\frac{2\Phi^{-1}(\alpha)\phi''(1)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*,\boldsymbol{\xi}))}}{\bar{\mathfrak{m}}_{2,\mathbb{P}_0}(l(\mathbf{x}^*,\boldsymbol{\xi}))}$, *then we have that*

$$\lim_{n\to\infty} \mathbb{P}_0 \left(\min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] \geq 0 \right) = \alpha.$$

Proof of Theorem 4.B.10. We revisit the proof of Theorem 4.10. Let $T(\mathbb{P}) \triangleq \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[l(\mathbf{x},\boldsymbol{\xi})]$. Using Lemma 4.6 and the minimax theorem, we have that

$$\begin{aligned}
&= \min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] \\
&\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}} \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} \sum_{i=1}^n q_i l(\mathbf{x},\boldsymbol{\xi}_i) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] \\
&= \max_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}} \min_{\mathbf{x}\in\mathcal{X}} \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} \sum_{i=1}^n q_i l(\mathbf{x},\boldsymbol{\xi}_i) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] \\
&= \max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} T(\mathbb{Q}) - T(\mathbb{P}_0),
\end{aligned}$$

where

$$\begin{aligned}
&\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n) \\
&= \left\{ \sum_{i=1}^n q_i \iota_{\boldsymbol{\xi}_i} \;\middle|\; q_i \geq 0, \sum_{i=1}^n q_i = 1, \sum_{j\in J} q_j \leq \frac{(-\phi)_*\left(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)\cdot\frac{|J|}{n}\right)}{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}, \forall J \subset [n] \right\},
\end{aligned} \tag{4.34}$$

where $\iota_{\boldsymbol{\xi}_i}$ denotes the Dirac delta measure. Again, we denote the influence function $IF(\mathbf{x}^*, \mathbb{P}_0) \triangleq l(\mathbf{x}^*, X) - \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}^*, X)]$, and we examine the quantity

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})| \\
&\triangleq \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} \left| \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} \left(T(\mathbb{Q}) - T(\mathbb{P}_0) - \mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x}^*,\mathbb{P}_0)]\right) \right|.
\end{aligned}$$

Our first goal is to show that for any $\epsilon_0 > 0$, we have

$$\limsup_{n\to\infty} \mathbb{P}_0^* \left(\sqrt{n} \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})| > \epsilon_0 \right) = 0. \tag{4.35}$$

Fix $\delta_0 > 0$, and choose a measurable selection (see Lemma 4.A.6)

$$\mathbb{Q}_n(\omega) \in \mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1(\omega),\ldots,\boldsymbol{\xi}_n(\omega))}(\hat{\mathbb{P}}_n)$$

with $\mathbb{Q}_n(\omega) = \sum_{i=1}^n q_i(\omega)\iota_{\boldsymbol{\xi}_i(\omega)}$, such that

$$|\kappa_n(\mathbb{Q}_n(\omega))| \geq (1-\delta_0) \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1(\omega),\ldots,\boldsymbol{\xi}_n(\omega))}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})|.$$

Therefore, monotonicity of outer measure gives that

$$\limsup_{n\to\infty} \mathbb{P}_0^* \left(\sqrt{n} \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})| > \epsilon_0 \right) \leq \limsup_{n\to\infty} \mathbb{P}_0^* \left(\sqrt{n}|\kappa_n(\mathbb{Q}_n)| > (1-\delta_0)\epsilon_0 \right).$$

As a mapping $\mathbb{Q}_n : \Omega \to l^\infty(\mathcal{H})$, we have that by Lemma 4.A.5, the sequence $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)$ is asymptotically tight. Hence, Prohorov's theorem (van der Vaart and Wellner 2023) implies that there exists a subsequence such that the weak convergence $\sqrt{n(m)}(\mathbb{Q}_{n(m)} - \mathbb{P}_0) \rightsquigarrow Z$ holds for a tight limit Z and that

$$\limsup_{n\to\infty} \mathbb{P}_0^* \left(\sqrt{n}|\kappa_n(\mathbb{Q}_n)| > (1-\delta_0)\epsilon_0 \right) = \lim_{m\to\infty} \mathbb{P}_0^* \left(\sqrt{n(m)}|\kappa_{n(m)}(\mathbb{Q}_{n(m)})| \geq (1-\delta')\epsilon \right) = 0,$$

where the latter is obtained by applying the functional delta theorem (Theorem 3.10.4, van der Vaart and Wellner 2023), and that $\frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} \xrightarrow{P} 1$, to conclude that

$$\frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} \sqrt{n(m)}(T(\mathbb{Q}_{n(m)}) - T(\mathbb{P}_0) - \mathbb{E}_{\mathbb{Q}_{n(m)}}[IF(\mathbf{x}^*, \mathbb{P}_0)]) \xrightarrow{P^*} 0.$$

Hence, this proves (4.35). Using Theorem 4.9, we thus have that in a similar argument as in the proof of Theorem 4.10, that

$$\sqrt{n} \left| \left(\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} T(\mathbb{Q}) - T(\mathbb{P}_0) \right) + \left(1 - \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} \right) \right.$$
$$\left. \cdot T(\mathbb{P}_0) - \mathbb{E}_{\hat{\mathbb{P}}_n}[IF(\mathbf{x}^*, \mathbb{P}_0)] + \frac{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}{2\phi''(1)} \mathbb{E}_{\hat{\mathbb{P}}_n}[\min\{IF^{(1)}(\mathbf{x}^*, \mathbb{P}_0), IF^{(2)}(\mathbf{x}^*, \mathbb{P}_0)\}] \right| \xrightarrow{P^*} 0.$$

Therefore, using the Slutsky's theorem (example 1.4.7 of van der Vaart and Wellner 2023) and the fact that convergence in outer probability implies weak convergence (Lemma

1.10.2 of van der Vaart and Wellner 2023), we have that

$$\begin{aligned}
&\sqrt{n} \max_{\mathbb{Q} \in \mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)} T(\mathbb{Q}) - T(\mathbb{P}_0) \\
&\rightsquigarrow N(0, \mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))) - \frac{1}{2\phi''(1)} \frac{2\Phi^{-1}(\alpha)\phi''(1)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))}}{\mathbb{E}_{\mathbb{P}_0}[\min\{l(\mathbf{x}^*, \boldsymbol{\xi}^{(1)}), l(\mathbf{x}^*, \boldsymbol{\xi}^{(2)})\}]} \\
&\qquad \cdot \left(\mathbb{E}_{\mathbb{P}_0}[\min\{IF^{(1)}(\mathbf{x}^*, \mathbb{P}_0), IF^{(2)}(\mathbf{x}^*, \mathbb{P}_0)\}] + T(\mathbb{P}_0)\right) \\
&= N(0, \mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))) - \Phi^{-1}(\alpha)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))}.
\end{aligned}$$

Hence, we have that

$$\begin{aligned}
&\lim_{n \to \infty} \mathbb{P}_0 \left(\sqrt{n} \max_{\mathbb{Q} \in \mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)} T(\mathbb{Q}) - T(\mathbb{P}_0) \leq 0 \right) \\
&= \mathbb{P}\left(N(0, \mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))) - \Phi^{-1}(\alpha)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))} \leq 0 \right) \\
&= \Phi(\Phi^{-1}(\alpha)) = \alpha.
\end{aligned}$$

Finally, if instead we have that $\sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}$ converges in probability to $\frac{2\Phi^{-1}(\alpha)\phi''(1)}{\overline{\mathrm{m}}_{2,\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))} \cdot \sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))}$, then we have that

$$\begin{aligned}
&\sqrt{n}\left(\min_{\mathbf{x} \in \mathcal{X}} \mathcal{R}^{\phi}_{\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n), \hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})) - \min_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}, \boldsymbol{\xi})] \right) \\
&= \sqrt{n}\left(\min_{\mathbf{x} \in \mathcal{X}} \frac{\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}{(-\phi)_*(\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n))} \rho^{\phi}_{\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n), \hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})) - \min_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}, \boldsymbol{\xi})] \right) \\
&= \sqrt{n}\left(\min_{\mathbf{x} \in \mathcal{X}} \rho^{\phi}_{\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n), \hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})) - \min_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}, \boldsymbol{\xi})] \right) \\
&\qquad + \sqrt{n}\left(\left(\frac{\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}{(-\phi)_*(\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n))} - 1 \right) \min_{\mathbf{x} \in \mathcal{X}} \rho^{\phi}_{\delta(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n), \hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})) \right) \\
&\rightsquigarrow N(0, \mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi})) - \frac{\Phi^{-1}(\alpha)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))}}{\overline{\mathrm{m}}_{2,\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))} \mathbb{E}_{\mathbb{P}_0}[\min\{l(\mathbf{x}^*, \boldsymbol{\xi}^{(1)}), l(\mathbf{x}^*, \boldsymbol{\xi}^{(2)})\}] \\
&\qquad + \frac{\Phi^{-1}(\alpha)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))}}{\overline{\mathrm{m}}_{2,\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}^*, \boldsymbol{\xi})] \\
&= N(0, \mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi})) + \Phi^{-1}(\alpha)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))}.
\end{aligned}$$

□

Bibliography

M. Allais. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica*, 21:503–546, 1953.

R. B. Ash. *Information Theory*. Dover, New York, NY, USA, 1990.

D. Bartl and S. Mendelson. On monte-carlo methods in convex stochastic optimization. *Annals of Applied Probability*, 32(4):3146–3198, 2022.

D. Bartl and L. Tangpi. Non-asymptotic rates for the estimation of risk measures. *Mathematics of Operations Research*, 48(4):1811–2382, 2022.

D. Bartl, S. Drapeau, and L. Tangpi. Computational aspects of robust optimized certainty equivalents and option pricing. *Mathematical Finance*, 30(1):287–309, 2020.

A. Beck and A. Ben-Tal. Duality in robust optimization: Primal worst equals dual best. *Operations Research Letters*, 37:1–6, 2009.

F. Bellini, B. Klar, A. Müller, and E. R. Gianin. Generalized quantiles as risk measures. *Insurance: Mathematics and Economics*, 54:41–48, 2014.

D. Belomestny and V. Krätschmer. Central limit theorems for law-invariant coherent risk measures. *Journal of Applied Probability*, 49:1–21, 2012.

A. Ben-Tal and A. Ben-Israel. A recourse certainty equivalent for decisions under uncertainty. *Annals of Operations Research*, 30:1–44, 1991.

A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization*. SIAM, 2019.

A. Ben-Tal and M. Teboulle. Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Management Science*, 32(11):1445–1466, 1986.

A. Ben-Tal and M. Teboulle. Penalty functions and duality in stochastic programming via ϕ-divergence functionals. *Mathematics of Operations Research*, 12(2):224–240, 1987.

A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

A. Ben-Tal, A. Ben-Israel, and M. Teboulle. Certainty equivalents and information measures: Duality and extremal principles. *Journal of Mathematical Analysis And Applications*, 157(1):211–236, 1991.

A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, Princeton, NJ, 2009.

A. Ben-Tal, D. den Hertog, A. de Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

A. Ben-Tal, R. Brekelmans, D. den Hertog, and J. Vial. Globalized robust optimization for nonlinear uncertain inequalities. *INFORMS Journal on Computing*, 29(2):197–376, 2017.

C. Berge. *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*. Courier Corporation, 1963.

C. Bernard, S. M. Pesenti, and S. Vanduffel. Robust distortion risk measures. *Mathematical Finance*, 34:774–818, 2024.

D. Bertsimas and D. B. Brown. Constructing uncertainty sets for robust linear optimization. *Operations Research*, 57(6):1483–1495, 2009.

D. Bertsimas and D. den Hertog. *Robust and Adaptive Optimization*. Dynamic Ideas LLC Belmont, Massachusetts, 2022.

D. Bertsimas, I. Dunning, and M. Lubin. Reformulation versus cutting-planes for robust optimization. *Computational Management Science*, 13(2):195–217, 2016.

R. G. Bland, D. Goldfarb, and M. J. Todd. The ellipsoid method: A survey. *Operations Research*, 29(6):1039–1091, 1981.

T. Breuer and I. Csiszár. Measuring distibution model risk. *Mathematical Finance*, 26 (2):395–411, 2016.

C. J. Bélisle, H. E. Romeijn, and R. L. Smith. Hit-and-Run algorithms for generating multivariate distributions. *Mathematics of Operations Research*, 18(2):255–266, 1993.

J. Cai, J. Y.-M. Li, and T. Mao. Distributionally robust optimization under distorted expectations. *Operations Research*, 73(2), 2023.

G. Carlier and R.-A. Dana. Law invariant concave utility functions and optimization problems with monotonicity and comonotonicity constraints. *Statistics & Risk Modeling*, 24 (1):127–152, 2006.

P. R. Chares. *Cones and Interior-Point Algorithms for Structured Convex Optimization Involving Powers and Exponentials*. Ecole polytechnique de Louvain, Université catholique de Louvain, 2009.

H. Chen, D. Dentcheva, Y. Lin, and G. J. Stock. Central limit theorems for vector-valued composite functionals with smoothing and applications, 2024. Available on ArXiv.

R. Chen and I. C. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13): 1–48, 2018.

A. S. Cherny and D. B. Madan. New measures for performance evaluation. *The Review of Financial Studies*, 22(7):2571–2606, 2009.

S. H. Chew, E. Karni, and Z. Safra. Risk aversion in the theory of expected utility with rank dependent probabilities. *Journal of Economic Theory*, 42:370–381, 1987.

G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1954.

J. Chu and L. Tangpi. Non-asymptotic estimation of risk measures using stochastic gradient Langevin dynamics. *SIAM Journal on Financial Mathematics*, 15(2):503–536, 2024.

M. Cox. An algorithm for approximating convex functions by means by first degree splines. *The Computer Journal*, 14(3):272–275, 1971.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975.

E. Delage, S. Guo, and H. Xu. Shortfall risk models when information of loss function is incomplete. *Operations Research*, 70(6):3511–3518, 2022.

F. Delbaen. Risk measures for non-integrable random variables. *Mathematical Finance*, 19:329–333, 2009.

D. Denneberg. Distorted probabilities and insurance premiums. *Methods of Operations Research*, 63:3–5, 1990a.

D. Denneberg. Premium calculation: Why standard deviation should be replaced by absolute deviation. *ASTIN Bulletin: The Journal of the IAA*, 20(2):181–190, 1990b.

D. Denneberg. *Non-Additive Measure and Integral*. Springer, 1994.

J. C. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 49(3):1378–1406, 2021.

J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: a generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

R. Durrett. *Probability: Theory and Examples*, volume 5. Cambridge Series in Statistical and Probabilistic Mathematics, 2019.

L. Eeckhoudt, C. Gollier, and H. Schlesinger. The risk-averse (and prudent) newsboy. *Management Science*, 41(5):786–794, 1995.

L. R. Eeckhoudt and R. J. Laeven. Dual moments and risk attitudes. *Operations Research*, 70(3):1330–1341, 2021.

L. R. Eeckhoudt, R. J. Laeven, and H. Schlesinger. Risk apportionment: The dual story. *Journal of Economic Theory*, 185:104971, 2020.

A. Eichhorn and W. Römisch. Stochastic integer programming: Limit theorems and confidence intervals. *Mathematics of Operations Research*, 32(1):118–135, 2007.

D. Ellsberg. Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

G. C. Ferrante. Bounds on binomial tails with applications. *IEEE Transactions on Information Theory*, 67(12):8273–8279, 2021.

H. Föllmer and T. Knispel. *Handbook of the Fundamentals of Financial Decision Making: In 2 Parts*. Scientific Publishing Co. Pte. Ltd, 2013.

H. Föllmer and A. Schied. *Stochastic Finance, 4th ed.* Berlin: De Gruyter, 2016.

R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2022.

I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

E. Gine and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, Cambridge, 2016.

C. Gini. *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. C. Cuppini, Bologna, 1912.

C. Gini. Measurement of inequality of incomes. *The Economic Journal*, 31(121):124–126, 1921.

P. Glasserman and X. Xu. Robust risk measurement and model risk. *Quantitative Finance*, 14(1):29–58, 2014.

M. J. Goovaerts and R. J. Laeven. Actuarial risk measures for financial derivative pricing. *Insurance: Mathematics and Economics*, 42(2):540–547, 2008.

B. L. Gorissen, H. Blanc, A. Ben-Tal, and D. den Hertog. Technical Note–Deriving robust and globalized robust solutions of uncertain linear programs with general convex uncertainty sets. *Operations Research*, 62(3):672–679, 2014.

J. Gotoh and Y. Takano. Newsvendor solutions via conditional value-at-risk minimization. *European Journal of Operational Research*, 179(1):80–96, 2007.

J. Gotoh, M. J. Kim, and A. E. Lim. Robust empirical optimization is almost the same as mean-variance optimization. *Operations Research Letters*, 46:448–452, 2018.

J. Gotoh, M. J. Kim, and A. E. Lim. Worst-Case Sensitivity. *Paper*, 2020. Available on ArXiv.

J. Gotoh, M. J. Kim, and A. E. B. Lim. Calibration of distributionally robust empirical optimization models. *Operations Research*, 69(5):1630–1650, 2021.

V. Guigues, A. Nemirovski, and A. Juditsky. Non-asymptotic confidence bounds for optimal value of a stochastic program. *Optimization Methods and Software*, 32:1033–1058, 2017.

V. Guigues, V. Krätschmer, and A. Shapiro. A central limit theorem and hypotheses testing for risk-averse stochastic programs. *SIAM Journal on Optimization*, 28:1337–1366, 2018.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL `https://www.gurobi.com`.

L. P. Hansen and T. J. Sargent. Robust control and model uncertainty. *American Economic Review*, 91:60–66, 2001.

L. P. Hansen and T. J. Sargent. *Robustness*. Princeton University Press, 2008.

X. D. He and X. Y. Zhou. Portfolio choice via quantiles. *Mathematical Finance*, 21(2):203–231, 2011.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

A. Imamoto and B. Tang. Optimal piecewise linear approximation of convex functions . *Proceedings of the World Congress on Engineering and Computer Science*, 2008.

G. Jin, R. J. Laeven, D. den Hertog, and A. Ben-Tal. Composite Uncertainty Sets, 2024. Available on arXiv.

G. Jin, V. Krätschmer, and R. J. Laeven. Sample Average Approximation of Risk Functionals: Non-Asymptotic Error Bounds, 2025a. Working Paper.

G. Jin, R. J. Laeven, and D. den Hertog. Robust optimization of rank-dependent models with uncertain probabilities, 2025b. Available on arXiv.

G. Jin, R. J. Laeven, and H. Lam. A New Distributionally Robust Optimization Model Based on Maxiance Regularization, 2025c. Working Paper.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

M. Kaina and L. Rüschendorf. On convex risk measures on L^p-spaces. *Mathematical Methods of Operations Research*, 69:475–495, 2009.

S. Kim, R. Pasupathy, and S. G. Henderson. A Guide to Sample-Average Approximation. In *Handbook of Simulation Optimization*, chapter 8, pages 207–243. Springer, 2015.

V. Krätschmer. Nonasymptotic upper estimates for errors of the sample average approximation method to solve risk-averse stochastic programs. *SIAM Journal on Optimization*, 34:1264–1294, 2024a.

V. Krätschmer. First order asymptotics of the sample average approximation method to solve risk averse stochastic programs. *Mathematical Programming*, 208:209–242, 2024b.

T. Kruse, J. C. Schneider, and N. Schweizer. Technical note—the joint impact of F-divergences and reference models on the contents of uncertainty sets. *Operations Research*, 67(2):428–435, 2019.

T. Kruse, J. C. Schneider, and N. Schweizer. A toolkit for robust risk assessment using F-divergences. *Management Science*, 67(10):6529–6552, 2021.

D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: theory and applications in machine learning. *Informs Tutorials in Operations Research*, 0(0):130–166, 2019.

S. Kusuoka. On law invariant coherent risk measures. *Advances in Mathematical Economics*, 3:83–95, 2001.

R. J. Laeven and M. Stadje. Entropy coherent and entropy convex measures of risk. *Mathematics of Operations Research*, 38(2):265–293, 2013.

R. J. Laeven and M. Stadje. A rank-dependent theory for decision under risk and ambiguity, 2024. Available on ArXiv.

H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

H. Lam and C. Mottet. Tail analysis without parametric models: A worst-case perspective. *Operations Research*, 65(6):1696–1711, 2017.

V. L. Levin. A characterization theorem for normal integrands with applications to descriptive function theory, functional analysis and nonconvex optimization. *Set-Valued Analysis*, 2:395–414, 1994.

D. G. Luenberger. *Optimization by Vector Space Methods*. Wiley Professional Paperback Series, 1969.

F. Maccheroni, M. Marinacci, and A. Rustichini. Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica*, 74(6):1447–1498, 2006.

A. Mas-Colell, M. D. Whinston, and J. R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics, Invited Papers at the Twelfth British Combinatorial Conference*, pages 148–188, 1989.

P. Muliere and M. Scarsini. A note on stochastic dominance and inequality measures. *Journal of Economic Theory*, 49(2):314–323, 1989.

A. Mutapcic and S. Boyd. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software*, 24(3):381–406, 2009.

W. Newey and J. Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55:819–847, 1987.

M. Norton, V. Khokhlov, and S. Uryasev. Calculating CVaR and bPOE for common probability distributions with application to portfolio optimization and density estimation. *Annals of Operations Research*, 299:1281–1315, 2021.

I. Olkin and S. Yitzhaki. Gini regression analysis. *International Statistical Review*, 60(2): 185–196, 1992.

L. Pardo. *Statistical Inference Based on Divergence Measures*. Chapman & Hall/ CRC Boca Raton, 2006.

S. Pesenti, Q. Wang, and R. Wang. Optimizing distortion riskmetrics with distributional uncertainty. *Paper*, 2020. Available on ArXiv.

G. C. Pflug and D. Wozabal. Asymptotic distribution of law-invariant risk functionals. *Finance and Stochastics*, 14:397–418, 2010.

K. Postek, D. den Hertog, and B. Melenberg. Computationally tractable counterparts of distributionally robust constraints on risk measures. *SIAM Review*, 58(4):603–650, 2016.

D. Prelec. The probability weighting function. *Econometrica*, 66(3):497–527, 1998.

J. Quiggin. A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3(4):323–343, 1982.

U. Rieder. Measurable selection theorems for optimization problems. *Manuscripta Mathematica*, 24:115–131, 1978.

R. T. Rockafellar. Integrals which are convex functionals. *Pacific Journal of Mathematics*, 24(3):525–539, 1968.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer Verlag, Berlin-Heidelberg-New York, 2004.

W. Römisch. Delta method, infinite dimensional. In *Encyclopedia of Statistical Sciences*, volume 16. Wiley, New York, 2006.

A. Ruszczynski and A. Shapiro. Optimization of convex risk functions. *Mathematics of Operations Research*, 31:433–451, 2006.

H. S. Ryoo and N. V. Sahinidis. A branch-and-reduce approach to global optimization. *Journal of Global Optimization*, 8(2):107–138, 1996.

D. Schmeidler. Integral representation without additivity. *Proceedings of the American Mathematical Society*, 97:255–261, 1986.

D. Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571–587, 1989.

J. C. Schneider and N. Schweizer. Robust measurement of (heavy-tailed) risks: Theory and implementation. *Journal of Economics Dynamics and Control*, 61(1-2):183–203, 2015.

H. Shalit and S. Yitzhaki. Mean-gini, portfolio theory, and the pricing of risky assets . *The Journal of Finance*, 39(5):1449–1468, 1984.

A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Mathematics of Operations Research*, 18:829–845, 1993.

A. Shapiro. Consistency of sample estimates of risk averse stochastic program. *Journal of Applied Probability*, 50:533–541, 2013.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures On Stochastic Programming*. SIAM, 2009.

J. E. Smith and R. L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

M. Talagrand. Sharper bounds for gaussian and empirical processes. *Annals of Statistics*, 22:28–76, 1994.

M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563, 1996.

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

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, 2023.

B. P. van Parys and B. Zwart. Robust mean estimation for optimization: The impact of heavy tails, 2025. Available on arXiv.

B. P. van Parys, P. M. Esfahani, and D. Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402, 2021.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 3 edition, 1944.

P. P. Wakker. *Prospect Theory for Risk and Rationality*. Cambridge University Press, 2010.

A. Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

S. Wang. Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics*, 17(1):43–54, 1995.

S. Wang. A class of distortion operators for pricing financial and insurance risks. *The Journal of Risk and Insurance*, 67(1):15–36, 2000.

W. Wang and H. Xu. Preference robust distortion risk measure and its application. *Mathematical Finance*, 33(2):389–434, 2023.

W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

M. E. Yaari. The dual theory of choice under risk. *Econometrica*, 55(1):95–115, 1987.

S. Yitzhaki. Stochastic dominance, mean variance, and Gini's mean difference. *The American Economic Review*, 72(1):178–185, 1982.

S. Yitzhaki and E. Schechtman. *The Gini Methodology: A Primer on a Statistical Methodology*. Springer New York, 2012.

Summary in English

This dissertation presents four research papers on the subject of decision under uncertainty, with a focus on decision-making in a risk-averse setting. The aim is to develop robust optimization and statistical techniques to obtain and analyze solutions for stochastic optimization problems, where the uncertain objective function is evaluated by a risk measure.

In Chapter 1, we study distributionally robust optimization problems for a rich class of risk measures with ambiguity sets defined by ϕ-divergences. The risk measures are allowed to be non-linear in probabilities, are represented by Choquet integrals possibly induced by a probability weighting function, and encompass many well-known examples. Optimization for this class of risk measures is challenging due to their rank-dependent nature. We show that for various shapes of probability weighting functions, including concave, convex and inverse S-shaped, the robust optimization problem can be reformulated into a rank-independent problem. In the case of a concave probability weighting function, the problem can be reformulated further into a convex optimization problem that admits explicit conic representability for a collection of canonical examples. While the number of constraints in general scales exponentially with the dimension of the state space, we circumvent this dimensionality curse and develop two types of algorithms. They yield tight upper and lower bounds on the exact optimal value and are formally shown to converge asymptotically. This is illustrated numerically in a robust newsvendor problem and a robust portfolio choice problem.

In Chapter 2, we study a robust optimization approach to combat tail uncertainty by proposing a unifying framework to construct uncertainty sets for a broad class of risk measures, given a specified nominal model. Our framework is based on a parametrization of robust risk measures using two (or multiple) ϕ-divergence functions, which enables us to provide uncertainty sets that are tailored to both the sensitivity of each risk measure to tail losses and the tail behavior of the nominal distribution. In addition, our formulation allows for a tractable computation of robust risk measures, and elicitation of ϕ-divergences that describe a decision maker's risk and ambiguity preferences.

In Chapter 3, we study Sample Average Approximation (SAA) for risk-averse stochastic programs, where the risk of an uncertain goal function is measured by a law-invariant convex risk measure. We provide non-asymptotic upper estimation on the deviation prob-

ability of absolute error between SAA and the true optimal value under more general assumptions for the goal function and risk measure, as compared to previous works by Bartl and Tangpi [2022] and Krätschmer [2024a]. In particular, the estimation is based on conditions from empirical process theory, which allow the goal function to be non-linear, unbounded, and do not require any pathwise analytical properties such as continuity or convexity. Conditions on the risk measures are also readily verifiable in the coherent case, with prominent examples including distortion risk measures and expectiles. Furthermore, our non-asymptotic bound contains explicit constants that can be used to construct confidence regions for the true optimal value and can also be adapted to provide estimates for stochastic programs with inverse S-shaped distortion risk measures.

Finally, in Chapter 4, we propose a new DRO model, which we call *dual* DRO, that is based on *maxiance* regularization, a variability measure also known as the Gini's mean difference. As we illustrate with numerical examples, this difference in regularization scheme ensures that the dual DRO model is more robust against outliers than ϕ-divergence DRO, while also providing a tradeoff between bias and variability. Furthermore, we show that optimization of dual DRO models is just as tractable as ϕ-divergence DRO, and can be used as a tractable lower bound on the true optimal expectation value problem with the same $O(1/\sqrt{n})$ asymptotic convergence rate as ϕ-divergence DRO.

Nederlandse Samenvatting (Summary in Dutch)

Dit proefschrift bestaat uit vier onderzoeksartikelen over het onderwerp van beslissingen onder onzekerheden, met een focus op beslissingsproblemen in risico-averse instellingen. Het doel is om robuuste optimalisatie- en statistische technieken te ontwikkelen voor het oplossen van stochastische beslissingsproblemen, waarbij de onzekere doelfunctie door een risicomaat geëvalueerd wordt.

In Hoofdstuk 1 bestuderen we robuuste optimalisatieproblemen voor een brede klasse van risicomaten, waarbij de onzekerheidsverzameling via een ϕ-divergentie gedefinieerd is. De risicomaten mogen niet-lineair zijn in de kansen, zijn gerepresenteerd via een Choquet integraal met een wegingsfunctie en omvat vele bekende voorbeelden. Optimalisatieproblemen met deze klasse van risicomaten zijn extra uitdagend vanwege de rangafhankelijkheid. We laten voor verschillende vormen van de kanswegingsfunctie, zoals concaaf, convex en omgekeerd S-vormig, zien dat het robuuste optimalisatieprobleem geherformuleerd kan worden tot een rangonafhankelijk probleem. In het geval van een concave kanswegingsfunctie kan het probleem verder geherformuleerd worden tot een convex optimalisatieprobleem dat een expliciete kegelrepresentatie toelaat onder veel voorbeelden. Ondanks dat het aantal voorwaarden van het herformuleerde probleem exponentieel stijgt met de dimensie van de toestandsruimte, omzeilen we deze moeilijkheid door het ontwikkelen van twee typen algoritmes. Beide geven strakke boven- en ondergrenzen die convergeren naar de exacte optimale waarde. Dit is geïllustreerd met voorbeelden uit robuust voorraadbeheer en portefeuilleoptimalisatie.

In Hoofdstuk 2 gebruiken we robuuste optimalisatie om staartonzerkeheid aan te pakken. We introduceren een universeel raamwerk om onzekerheidsverzameling te construeren voor een brede klasse van risicomaten, gegeven een nominaal model. Ons raamwerk is gebaseerd op een parametrisatie van robuuste risicomaten met twee of meerdere ϕ-divergenties. Deze methode biedt ons de mogelijkheid om onzekerheidsverzamelingen te bouwen die afgestemd zijn op de gevoeligheid voor staartonzekerheden van risicomaten en het staartgedrag van nominale modellen. Daarnaast maakt onze formulering een goed hanteerbare berekening van robuuste risicomaten mogelijk, evenals het afleiden van ϕ-divergenties die de risico- en ambiguïteitsvoorkeuren van een beslisser beschrijven.

In Hoofdstuk 3 bestuderen we de "Sample Average Approximation"(SAA) van risico-averse stochastische optimalisatieproblemen, waarbij het risico van een onzekere doelfunctie gemeten wordt door een distributie-invariante risicomaat. We geven een niet-asymptotische bovengrens voor de afwijkingskans tussen de SAA-schatting en de werkelijke optimale waarde onder voorwaarden op de doelfunctie en risicomaten die algemener zijn dan Bartl and Tangpi [2022] en Krätschmer [2024a]. In het bijzonder mag de doelfunctie niet-lineair en onbegrensd zijn, zonder enige aanname over analytische eigenschappen zoals convexiteit en continuïteit. De voorwaarden op de risicomaten zijn ook direct verifieerbaar in het coherente geval, met belangrijke voorbeelden zoals distortie-risicomaten en expectielen. Daarnaast kunnen onze bovengrenzen gebruikt worden om betrouwbaarheidsinterval voor optimale waarde te construeren, ook voor stochastische optimalisatieproblemen met omgekeerd S-vormige distortierisicomaten.

Ten slotte introduceren we in Hoofdstuk 4 een nieuw DRO-model, dat we de zogenaamde duale DRO noemen en dat gebaseerd is op maxiance-regularisatie, een maat voor variatie die ook bekend staat als de "Gini's mean difference". Zoals we illustreren met numerieke voorbeelden, zorgt maxiance-regularisatie ervoor dat het duale DRO-model veel robuuster is tegen uitschieters dan DRO's die gebaseerd zijn op ϕ-divergentie. Daarnaast is het duale DRO-model ook bijna equivalent met een bias-variatie-model en biedt dezelfde computationele voordelen als ϕ-divergentie-DRO-modellen. Bovendien kan het duale DRO-model dienen als een ondergrensschatting voor de exacte optimale waarde, met een statistische convergentiesnelheid van $O(1/\sqrt{n})$.