A New Distributionally Robust Optimization Model Based on Maxiance Regularization

Guanyu Jin[a], Roger J. A. Laeven[a,c,d], and Henry Lam[b]

[a]Dept. of Quantitative Economics, University of Amsterdam,
1001 NJ Amsterdam, The Netherlands

[b]Department of Industrial Engineering & Operations Research, Columbia University,
500 W. 120th Street New York, NY 10027, USA

[c]EURANDOM, 5600 MB Eindhoven, The Netherlands

[d]CentER, Tilburg University, 5000 LE Tilburg, The Netherlands

Abstract

Distributionally robust optimization (DRO) has been frequently used to address overfitting in empirical risk minimization. In particular, DRO models based on ϕ-divergence have become a popular choice for such purposes due to their tractability and variance regularization effect. However, when outliers are present, ϕ-divergence DRO can in fact be counter-effective for addressing overfitting, since variance is sensitive to outliers. 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.

1 Introduction

A large part of the data-driven optimization literature has been concentrated on finding optimal solution for the following problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})], \tag{1}$$

where $l(\mathbf{x},\boldsymbol{\xi})$ is some real-valued loss function that evaluates a decision variable $\mathbf{x}\in\mathbb{R}^d$, and an uncertain parameter $\boldsymbol{\xi}\in\mathbb{R}^k$ following the true, but often unknown distribution $\mathbb{P}_0$. Problem (1) encompasses a broad class of decision problems in statistics and economics, with examples including regression/classification problems, portfolio optimization and inventory management problem. A very direct approach to solve (1) is to replace the true expectation with its sample average approximation (SAA), formulated as

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x},\boldsymbol{\xi})], \tag{2}$$

where $\hat{\mathbb{P}}_n$ is the empirical distribution constructed from an i.i.d. sample of $\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_n$. However, SAA is known to be sensitive to overfitting, see for example Smith and Winkler [2006] that discusses the shortcoming of SAA in decision analysis. To address this overfitting issue, 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:

$$\min_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbb{Q}:I_\phi(\mathbb{Q},\hat{\mathbb{P}}_n)\le r} \mathbb{E}_{\mathbb{Q}}[l(\mathbf{x},\boldsymbol{\xi})], \tag{3}$$

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 any convex function $\phi:[0,\infty)\to[0,\infty)$ with $\phi(1)=0$, serves as a variational measure between two probabilities $\mathbb{Q}\ll\mathbb{P}$.

A main reason that makes ϕ-divergence DRO (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, (3) provides a tractable methodology for minimizing both the bias and variance. When there are near-optimal solutions of (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 distinction.

However, the use of variance as a measure of variability is only justified when the distribution is symmetric. Otherwise, a large variance does not necessarily imply a large indifference among the majority of a population. Indeed, by its own 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 does not reflect any relative differences in the entire population. This makes variance sensitive to outliers, which is one of the main cause for overfitting. By contrast, the Gini's mean difference, introduced by Corrado Gini in 1912, which is a well-known measure among economists that study income inequality (see e.g., Gini 1912, 1921), is based on relative differences, and defined as:

$$\mathrm{GMD}(X) = \frac{1}{2}\mathbb{E}|X^{(1)} - X^{(2)}|, \tag{4}$$

where for any random variable X, $X^{(1)}, X^{(2)}$ denote its i.i.d. copies. The advantages of GMD as a substitute for variance have been demonstrated for many examples in economics and statistics 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 book *The Gini's Methodology* by Yitzhaki and Schechtman [2012]. In risk theory, the GMD, which is referred to as *maxiance* in Eeckhoudt and Laeven [2021], has also 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 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 presented 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 two loss distributions X and Y, where $X = \text{Unif}[5,7] + 25\epsilon$ is a sum of uniform distribution on $[5,7]$, with a Bernoulli variable ϵ that takes 1 with probability 0.05, simulating outliers. 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 true optimal choice for problem (1). Suppose we do not know the true distributions but only observe 1000 samples of both X and Y, as represented in Figure 1. It can be observed that except for the outliers, the samples of Y exhibit a larger variability than X. Since their true expected value 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 (1) (for example, in the particular realizations given in Figure 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 1000 data in Figure 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 realistic representation of variability, and is a more reasonable choice than the variance.

Motivated by this example, we construct a new DRO model, where not the variance but the maxiance is regularized. It turns out that such construction is completely dual from the standard DRO procedure, in the sense that it considers deviations in a different dimension: while the traditional DRO such as (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 the ϕ-divergence DRO in (3) as *primal* DRO. More precisely, given any positive loss random variable $X = [x_1; p_1, x_2; p_2, \ldots, x_n; p_n]$ with outcomes x_i's that take probabilities p_i's, we define the dual DRO risk measure $\rho^{\phi}_{\delta,\mathbf{p}}(X)$ by minimizing the expected loss, under the shifted outcome values that are penalized from the nominal outcomes $(x_1, \ldots, x_n)$, measured using a ϕ-divergence with a penalization constant $\delta > 0$:

$$\rho^{\phi}_{\delta,\mathbf{p}}(X) = \inf_{\Delta\mathbf{y} \in \mathbb{R}^n_+} \sum_{i=1}^{n} \overline{F}_i \Delta y_i + \frac{1}{\delta} I_\phi(\Delta\mathbf{y}, \Delta\mathbf{x}). \tag{5}$$

Here, we defined $\Delta\mathbf{x} = (\Delta x_1, \ldots, \Delta x_n)$ 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 probabilties, so that for any $\Delta\mathbf{y} \in \mathbb{R}^+_n = \{\mathbf{a} \in \mathbb{R}^n : a_i \geq 0\}$, we

[1] see Yaari [1987], Quiggin [1982], Schmeidler [1986, 1989] for more background on non-expected utility theory

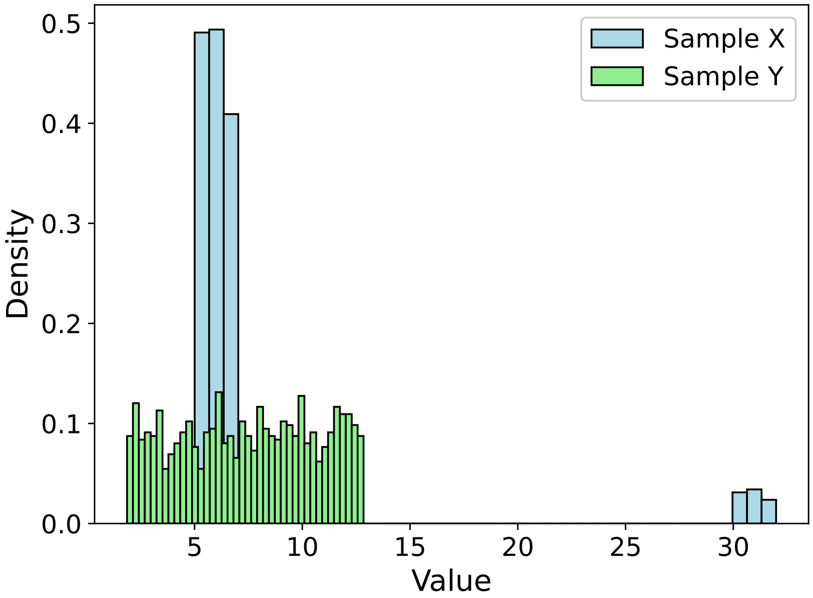

Figure 1: 1000 samples drawn for two distributions $X = \text{Unif}[5,7] + 25\epsilon$ and $Y = \text{Unif}[1.85, 12.85]$, where $\text{Unif}[a,b]$ is the uniform distribution on $[a,b]$ and $\epsilon \sim \text{Ber}(0.05)$ is a Bernoulli.

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 (5). We note that the dual DRO is designed to minimize both the expected loss and the maxiance, which is a measure of relative differences. If X is 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:

$$\min_{\mathbf{x} \in \mathcal{X}} \ \rho^{\phi}_{\delta(n), \hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})), \tag{6}$$

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 $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$[2], we are effectively searching for a decision $\mathbf{x}$ that minimizes the expected loss and relative differences of its distribution $l(\mathbf{x}, \boldsymbol{\xi}) \sim \hat{\mathbb{P}}_n$, which is the type of procedure that we aim to achieve.

Our main contributions may be summarized as follows:

- We prove the asymptotic equivalence between dual DRO (5) and mean-maxiance regularization, as $\delta \to 0$. This equivalence also holds uniformly across all decision variables, for the optimization problem (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.

- We show that the minimization problem (6) can be computed by solving a convex optimization problem that provides tight lower and upper bounds on (6). We then provide numerical examples where the solution of a dual DRO optimization problem can be more optimal than the ones obtained

[2] $\iota_{\boldsymbol{\xi}_i}$ denotes the Dirac delta distribution.

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 (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 effect 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 procedure for the optimal expected value problem (1) in the setting of heavy-tailed distributions. Bartl and Mendelson [2022] purposed a novel procedure for estimating the solution and objective value of (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 an $O(1/\sqrt{n})$ 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 purpose in this paper is computationally tractable, and can serve as an attractive new data-driven formulation for heavy-tailed distributions, as our numerical simulations suggests that maxiance is robust against outliers and adds a much less conservative regularization term to the empirical mean than the mean-variance model.

The remaining parts of the paper are organized as follows: we first introduce the maxiance and its relation to distortion risk measures in Section 2. We then prove our main results on asymptotic equivalence in Section 3. In Section 4, we study the minimization problem (6), and its statistical properties in Section 5. Finally, a concluding remark is provided in Section 6.

2 Preliminaries

2.1 ϕ-Divergence

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 $\text{dom}(\phi) \subset [0, \infty)$. Then, the ϕ-divergence between any two non-negative vectors $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n_+$, is defined as:

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

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.

For any convex function $f : \mathbb{R}^d \to \mathbb{R}$, we denote $f^*(\mathbf{y}) = \sup_{\mathbf{x}\in\mathrm{dom}(f)} \mathbf{y}^T\mathbf{x} - f(\mathbf{y})$ as the convex conjugate of f. For any concave function $g : \mathbb{R}^d \to \mathbb{R}$, we denote $g_*(\mathbf{y}) = \inf_{\mathbf{x}\in\mathrm{dom}(g)} \mathbf{y}^T\mathbf{x} - g(\mathbf{x})$ as 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:

$$(-\phi)_*(y) = -\phi^*(-y), \forall y \in \mathbb{R}. \tag{7}$$

In the following proposition, we summarize some properties of ϕ^* that are frequently used in this paper.

Proposition 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 $(\phi^*)'''(0) = -\phi'''(1)/(\phi''(1))^3$.

2.2 Maxiance and Distortion Risk Measures

For a random variable X, the maxiance is defined as

$$\overline{\mathtt{m}}_2(X) = \mathbb{E}[\max\{X^{(1)}, X^{(2)}\}] - \mathbb{E}[X], \tag{8}$$

where $X^{(1)}, X^{(2)}$ are i.i.d. draws of X. We note that the maxiance can also be equally expressed as $\overline{\mathtt{m}}_2(X) = \mathbb{E}[X] - \mathbb{E}[\min\{X^{(1)}, X^{(2)}\}]$[3]. Throughout this paper, we also refer the quantity $\mathbb{E}[\min\{X^{(1)}, X^{(2)}\}]$ as the second *dual* moment (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 variable. Indeed, one has $\overline{\mathtt{m}}_2(X) = \mathbb{E}|X^{(1)} - X^{(2)}|/2$.

An important tool that allows us to study the dual DRO model (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 $h(0) = 0$, $h(1) = 1$. Given a distortion function h, the corresponding distortion risk measure of a loss variable X is defined as the Choquet integral

$$\rho_h(X) = \int_0^\infty h\left(\mathbb{P}[X > t]\right) \mathrm{d}t + \int_{-\infty}^0 \left(h\left(\mathbb{P}[X > t]\right) - 1\right) \mathrm{d}t. \tag{9}$$

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

$$\rho_h(X) = \sum_{i=1}^n h\left(\overline{F}_i\right)(x_i - x_{i-1}), \tag{10}$$

[3] since $\mathbb{E}[\max\{X^{(1)}, X^{(2)}\}] + \mathbb{E}[\min\{X^{(1)}, X^{(2)}\}] = 2\mathbb{E}[X]$

where $\overline{F}_i = \sum_{j=i}^n p_j$ for $i = 1, \dots, n$, are the decumulative probabilities and we set $x_0 = 0$. The class of distortion risk measures contains several 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 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]. Moreover, for any integer n, we let $[n] := \{1, \dots, n\}$.

3 Asymptotic Equivalence of Dual DRO and Mean-Maxiance

We now study the dual DRO model in a discrete probability setting, since one often works with the empirical distribution in practice. 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 $0 < x_1 \le x_2 \le \ldots \le x_n$ and with probabilities $p_1, p_2, \ldots, p_n$. We assume that $\min_i p_i > 0$ (otherwise, discard those x_i's where $p_i = 0$). We denote $\Delta\mathbf{x} = (\Delta x_1, \dots, \Delta x_n)$ with $\Delta x_i = x_i - x_{i-1}$ for $i = 2, \dots, n$ and $\Delta x_1 = x_1$. The same definition is applied to $\Delta\mathbf{y}$. Then, we have that $\Delta\mathbf{x} \in \mathbb{R}_+^n$, where $\mathbb{R}_+^n$ is the set of n-dimensional vectors with non-negative entries.

Let ϕ be a divergence function. We recall the definition of the dual DRO model, where instead of the probabilities, we optimize with respect to the payoffs dimension $\Delta\mathbf{y}$, controlled by a ϕ-divergence penalization term:

$$\rho_{\delta,\mathbf{p}}^{\phi}(X) = \inf_{\Delta\mathbf{y}\in\mathbb{R}_+^n} \sum_{i=1}^n \overline{F}_i \Delta y_i + \frac{1}{\delta} I_\phi(\Delta\mathbf{y}, \Delta\mathbf{x}), \tag{11}$$

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 1. *Let ϕ be a ϕ-divergence function that is twice continuously differentiable at 1. Then, we have that*

$$\rho_{\delta,\mathbf{p}}^{\phi}(X) = \left(1 - \frac{\delta}{2\phi''(1)}\right)\mathbb{E}_{\mathbf{p}}[X] + \frac{\delta}{2\phi''(1)}\overline{\mathbb{m}}_{2,\mathbf{p}}(X) + o(\delta). \tag{12}$$

Although not stated in Theorem 1, we should note that the dual DRO in (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. In particular, as shown in the proof of Theorem 1, the dual DRO model in (11) is proportional to the following distortion risk measure:

$$\begin{aligned}\rho_{\delta,\mathbf{p}}^{\phi}(X) &= \frac{(-\phi)_*(\delta)}{\delta}\mathcal{R}_{\delta,\mathbf{p}}^{\phi}(X) \\ &\triangleq \frac{(-\phi)_*(\delta)}{\delta}\sum_{i=1}^n h_\delta(\overline{F}_i)\Delta x_i,\end{aligned} \tag{13}$$

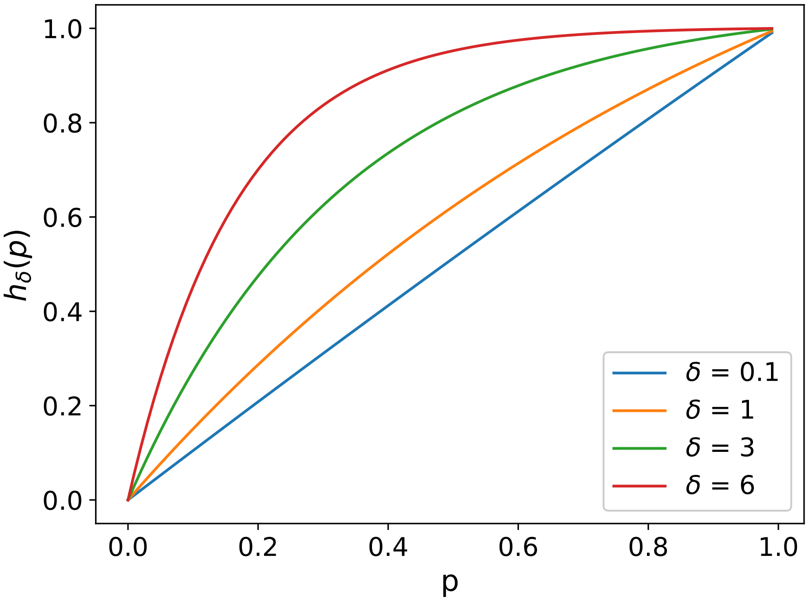

Figure 2: Shape of the distortion function $h_\delta(p) = \frac{(-\phi)_*(\delta p)}{(-\phi)_*(\delta)}$ with varying δ, where ϕ is the Kullback-Leibler divergence.

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 risk 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 (11), shapes the concavity of the distortion function h_δ. As shown in Figure 2, a larger δ gives more concavity to h_δ, which puts more probability weights on the worst outcomes of the loss X.

As we can see in (12), the asymptotic analysis of $\rho^\phi_{\delta,\mathbf{p}}(X)$ shows that there is still a factor $\delta/(2\phi''(1))$ at 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 2. *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 mentioned in (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 (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

$$\rho_{xp}(X) = \sup_{\mathbf{q}\in\mathbb{P}^n} \inf_{\Delta\mathbf{y}\in\mathbb{R}^n_+} \sum_{i=1}^n q_i y_i + \frac{1}{\delta_1} I_\phi(\Delta\mathbf{y}, \Delta\mathbf{x}) - \frac{1}{\delta_2} I_\psi(\mathbf{q}, \mathbf{p}). \tag{14}$$

The next result states the asymptotic equivalence of a primal-dual DRO model to mean-variance-maxiance regularization.

Theorem 2. *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*

$$\begin{aligned}\rho_{xp}(X) = &\left(1 - \frac{\delta_1}{2\phi''(1)}\right)\mathbb{E}_{\mathbf{p}}[X] + \frac{\delta_2}{2\psi''(1)}\mathrm{Var}_{\mathbf{p}}(X) \\ &+ \frac{\delta_1}{2\phi''(1)}\overline{\mathbb{m}}_{2,\mathbf{p}}(X) + o(\delta_1) + o(\delta_2) + O(\delta_1\delta_2 n).\end{aligned} \tag{15}$$

Therefore, we see that when a loss variable X only has finitely many n support points, then the above equivalence holds when $\delta_1, \delta_2 \to 0$. We note that since the dual DRO model is equivalent to the distortion risk measure (13), this also gives a regularization perspective for a distributionally robust distortion risk measure.

4 Optimization of Dual DRO Models

In this section, we discuss how to solve a dual DRO minimization problem in a data-driven setting. Suppose we have an i.i.d. sample of data $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$. This allows us to construct an empirical distribution on $\boldsymbol{\xi}$, namely $\hat{\mathbb{P}}_n = \sum_{i=1}^n \iota_{\boldsymbol{\xi}_i}/n$, where $\iota_\cdot$ denotes the Dirac delta measure. Then, we can consider the following dual DRO minimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \rho^\phi_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})), \tag{16}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a compact feasible set imposed on $\mathbf{x}$ consisting of convex constraints, and $l(\mathbf{x}, \boldsymbol{\xi})$ is a positive loss function that 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 (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 (3), due to its rank-dependent nature on the decision variable. To circumvent this, we use (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 (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 theorem that states an equivalence between distortion risk measures and robust optimization.

Theorem 3. *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 (10), and $M^h(\mathbf{p})$ is the set

$$\begin{aligned} M^h(\mathbf{p}) = \Bigg\{ \mathbf{q} \in \mathbb{R}^n \,\Bigg|\, & \mathbf{q} \geq \mathbf{0},\ \sum_{i=1}^{n} q_i = 1, \\ & \sum_{j \in J} q_j \leq h\left(\sum_{j \in J} p_j \right), \forall J \subset [n] \Bigg\}. \end{aligned} \tag{17}$$

Hence, by (13), it immediately follows from Theorem 3 that problem (16) is equivalent to a min-max problem, where the uncertainty set is given by (17), for $h = h_{\delta(n)}$. This leads to the following corollary.

Corollary 1. *We have for any $\delta(n) > 0$,*

$$\min_{\mathbf{x} \in \mathcal{X}} \rho^{\phi}_{\delta(n), \hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})) = \frac{(-\phi)_*(\delta(n))}{\delta(n)} \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{q} \in \mathcal{M}^{\phi}_{\delta(n)}} \sum_{i=1}^{n} q_i l(\mathbf{x}, \boldsymbol{\xi}_i). \tag{18}$$

where

$$\begin{aligned} \mathcal{M}^{\phi}_{\delta(n)} = \Bigg\{ \mathbf{q} \in \mathbb{R}^n \,\Bigg|\, & \mathbf{q} \geq \mathbf{0},\ \sum_{i=1}^{n} q_i = 1, \\ & \sum_{j \in J} q_j \leq h_{\delta(n)}\left(\sum_{j \in J} \frac{1}{n} \right), \forall J \subset [n] \Bigg\}. \end{aligned} \tag{19}$$

We note that although the right-hand side of (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 (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. [2025].

One way to efficiently compute (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 (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 it 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 (17)), where $\mathbf{1}/n = (1/n, \ldots, 1/n) \in \mathbb{R}^n$. Hence, (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. [2025], this lower bound can be computed as the following convex optimization problem with $O(n \cdot K)$ number of constraints.

Theorem 4. *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:*

$$
\begin{aligned}
\min_{\substack{\mathbf{x}\in\mathcal{X}\\ \beta,\lambda_{ij},\nu_j}} \quad & \beta + \sum_{j=1}^K \nu_j b_j + \frac{1}{n}\sum_{i=1}^m \sum_{j=1}^K \lambda_{ij} l_j \\
\text{s.t.} \quad & l(\mathbf{x}, \boldsymbol{\xi}_i) - \beta - \sum_{j=1}^K \lambda_{ij} \leq 0, \ \forall i \in [n] \\
& \lambda_{ij} \leq \nu_j, \ \forall i \in [n], \ \forall j \in [K] \\
& \lambda_{ij}, \nu_j \geq 0, \ \forall i \in [n], \ \forall j \in [K]
\end{aligned}
\tag{20}
$$

We further note that as outlined by Jin et al. [2025], for any constant $\epsilon > 0$, one can find a piecewise-linear approximation $h_L \leq h$, with the least 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 (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 (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 (20), by simply replacing the constants b_j with $b_j + \epsilon$. As further shown by Jin et al. [2025], 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 (18) as $\epsilon \to 0$. Since our loss function is assumed to be positive, this condition is automatically satisfied.

4.1 A Numerical Investigation on the Advantages of Dual DRO Optimization

We investigate numerically the difference between dual DRO (5), primal DRO (3), and SAA (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.

4.1.1 Example 1: Portfolio Optimization

We consider a portfolio optimization problem with two assets $\boldsymbol{\xi} = (\xi_1, \xi_2)$ that follow a similar distribution as the example presented in Figure 1. We let $\xi_1 = \text{Unif}[6, 8] + 26\epsilon_0$, where ϵ_0 is a Bernoulli variable that takes 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 portfolio 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.

To compare the performance between primal DRO, the dual DRO, SAA, we generate 100 samples of $\boldsymbol{\xi}$ of size $n \in \{50, 100, 200, 500, 1000, 2000\}$ and record the average (among the 100) optimal portfolio weight $\overline{\mathbf{x}}^*_{1,\text{model}}$ on ξ_1 obtained from each model $\in \{\text{saa}, \text{p.dro}, \text{d.dro}\}$.[4] We choose for both primal and dual DRO models 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 2, identity (13) and Duchi et al. [2021], we have that for each $\mathbf{x} \in \{\mathbf{x}_1, \mathbf{x}_2\}$,

$$\begin{aligned}&\sup_{\mathbb{Q}: I_\phi(\mathbb{Q}, \hat{\mathbb{P}}_n) \leq r(n)} \mathbb{E}_{\mathbb{Q}}[l(\mathbf{x}, \boldsymbol{\xi})] \\ &\quad = \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x}, \boldsymbol{\xi})] + \sqrt{\frac{\chi^2_{0.95,1}}{2n}} \sqrt{\text{Var}_{\hat{\mathbb{P}}_n}[l(\mathbf{x}, \boldsymbol{\xi})]} \\ &\qquad + o(1/n), \end{aligned} \tag{p.DRO}$$

and

$$\begin{aligned}&\frac{\delta(n)}{(-\phi)_*(\delta(n))} \rho^\phi_{\delta(n), \hat{\mathbb{P}}_n}(l(\mathbf{x}, \boldsymbol{\xi})) \\ &\quad = \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x}, \boldsymbol{\xi})] + \sqrt{\frac{\chi^2_{0.95,1}}{2n}} \cdot \overline{\mathbb{m}}_{2, \hat{\mathbb{P}}_n}[l(\mathbf{x}, \boldsymbol{\xi})] \\ &\qquad + o(1/n). \end{aligned} \tag{d.DRO}$$

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 1. As we can observe, the average optimal portfolio 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 more preferable asset ξ_1, in cases where SAA underperforms due to its sensitivity to 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 portfolio weight on ξ_1 for both DRO models conditioned on the realizations $\hat{\mu}_1 > \hat{\mu}_2$ and $\hat{\mu}_1 < \hat{\mu}_2$, where

[4] We solved the dual DRO using (20), and a piecewise-linear approximation of $h_{\delta(n)}$ with uniform approximation error $\epsilon = 0.0001$

$\hat{\mu}_1, \hat{\mu}_2$ are empirical estimations 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 respectively to the situations where SAA makes either a perfectly correct or incorrect decision. As we can observe from Table 1, when $\hat{\mu}_1 > \hat{\mu}_2$, which is the case when SAA outputs the correct optimal portfolio weight $(1, 0)$, 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 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.

Table 1: Average optimal portfolio 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 conditioned on sample realizations $\hat{\mu}_1 > \hat{\mu}_2$ and $\hat{\mu}_1 < \hat{\mu}_2$, where $\hat{\mu}_1, \hat{\mu}_2$ are empirical estimations 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

4.1.2 Example 2: 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 of a distribution, which is also the optimum of the following nominal problem:

$$\inf_{x \in \mathbb{R}} \mathbb{E}_{\mathbb{P}_0} |x - \xi|. \tag{21}$$

Using the example given by Duchi and Namkoong [2019], we let $\xi \in \{-1, 0, 1\}$ be a distribution that takes 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 (21) for a sample size of $n = 1000$, 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 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 2. 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 2: 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.

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 (1), by deriving the limiting distribution of the following quantity

$$\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right). \tag{22}$$

In addition, as suggested by Proposition 2, an upper bound on (1) can also be obtained by simply minimizing the distortion risk measures as defined in (13). Hence, we are also interested in examining the limit distribution of

$$\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right). \tag{23}$$

To study both (22) and (23), 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 $\mathcal{H}$[5]. 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 $\sup_{\mathbf{x}\in\mathcal{X}} |l(\mathbf{x},\tilde{\boldsymbol{\xi}})| \leq M_2(\tilde{\boldsymbol{\xi}}),\ \forall \tilde{\boldsymbol{\xi}} \in \mathbb{R}^k$. 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 $\overline{\mathtt{m}}_{2,\mathbb{P}_0}(\mathbf{x}) = \overline{\mathtt{m}}_{2,\mathbb{P}_0}(l(\mathbf{x},\boldsymbol{\xi}))$. 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 $\overline{\mathtt{m}}_{2,\hat{\mathbb{P}}_n}(\mathbf{x}) = \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x},\boldsymbol{\xi})] - \mathrm{dm}_{2,\hat{\mathbb{P}}_n}(\mathbf{x})$.

We first examine the measurability of (22) and (23) as functions of Ω to the outcome space $\mathbb{R}$.

Proposition 3. *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* $\tilde{\boldsymbol{\xi}}$*, and measurable in* $\tilde{\boldsymbol{\xi}}$*, for all* $\mathbf{x}$*.*

The derivation of the limit distributions for (22) and (23) requires the following theorem which states that the maxiance regularization effect of dual DRO models also holds uniformly in the decision space.

Theorem 5. *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 $\delta(n) :=$*

[5] 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 (22) and (23) can also be viewed as elements in $l^{\infty}(\mathcal{H})$.

$\delta(\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_n)$ be a measurable function such that $\delta(\boldsymbol{\xi}_1(\omega),\dots,\boldsymbol{\xi}_n(\omega)) > 0$ for all $\omega \in \Omega$ and $\delta(n) \xrightarrow{P} 0$. Then, we have that,

$$\begin{aligned}
&\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) \\
&\qquad = \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x},\boldsymbol{\xi})] - \frac{\delta(n)}{2\phi''(1)}\mathrm{dm}_{2,\hat{\mathbb{P}}_n}(\mathbf{x}) + \epsilon_n(\mathbf{x}) \\
&\mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) \\
&\qquad = \mathbb{E}_{\hat{\mathbb{P}}_n}[l(\mathbf{x},\boldsymbol{\xi})] + \frac{\delta(n)}{2\phi''(1)}\overline{\mathtt{m}}_{2,\hat{\mathbb{P}}_n}(\mathbf{x}) + \tilde{\epsilon}_n(\mathbf{x}),
\end{aligned}$$

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 (22) and (23) 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 6. *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 $\delta(n) = \sqrt{r/n}$,*

$$\begin{aligned}
&\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right) \\
&\qquad \rightsquigarrow \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} G(\mathbf{x}) - \frac{\sqrt{r}}{2\phi''(1)}\mathrm{dm}_{2,\mathbb{P}_0}(\mathbf{x}),
\end{aligned}$$

and,

$$\begin{aligned}
&\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right) \\
&\qquad \rightsquigarrow \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} G(\mathbf{x}) + \frac{\sqrt{r}}{2\phi''(1)}\overline{\mathtt{m}}_{2,\mathbb{P}_0}(\mathbf{x}),
\end{aligned}$$

where $G(\mathbf{x})$ is a mean-zero Gaussian process with covariance

$$\mathrm{Cov}(\mathbf{x}_1,\mathbf{x}_2) = \mathrm{Cov}(l(\mathbf{x}_1,\boldsymbol{\xi}), l(\mathbf{x}_2,\boldsymbol{\xi})).$$

As shown in Theorem 6, at a rate of $1/\sqrt{n}$, the gap between the dual DRO upper and lower bounds with the true nominal optimal value converges consistently 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 (1). This is slightly 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. 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 the primal DRO.

We note that as a consequence of Theorem 6, 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 $\min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))$ are respectively one-sided lower and upper bound 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 distribution function, we have that

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) \le \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right) = \Phi\left(\frac{\sqrt{r}}{2\phi''(1)} \frac{\mathrm{dm}_{2,\mathbb{P}_0}(\mathbf{x}^*)}{\sigma_{\mathbb{P}_0}(\mathbf{x}^*)}\right) \tag{24}$$

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] \le \min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))\right) = \Phi\left(\frac{\sqrt{r}}{2\phi''(1)} \frac{\overline{\mathrm{m}}_{2,\mathbb{P}_0}(\mathbf{x}^*)}{\sigma_{\mathbb{P}_0}(\mathbf{x}^*)}\right) \tag{25}$$

Since $\mathbf{x}^*$ is an unknown quantity, an explicit calculation of (24) and (25) will require a consistent estimator of $\mathbf{x}^*$, for example the solution of SAA based on a second independent sample. In Section EC.2, we show that it is also possible to let r depend on the data, such that the probability (24) and (25) are equal to some confidence parameter α, as $n \to \infty$ (although this does require first estimating $\mathbf{x}^*$ using half of the data). If $\mathcal{X}^*_{\mathbb{P}_0}$ contains more than one solution, then the non-centered Gaussian process in Theorem 6 becomes hard to evaluate. In this case, we can still obtain a lower bound on the probability of $\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) \le \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]$, namely that for any $\mathbf{x}^* \in \mathcal{X}^*_{\mathbb{P}_0}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) \le \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right) \ge \Phi\left(\frac{\sqrt{r}}{2\phi''(1)} \frac{\mathrm{dm}_{2,\mathbb{P}_0}(\mathbf{x}^*)}{\sigma_{\mathbb{P}_0}(\mathbf{x}^*)}\right).$$

Finally, we note that an upper bound on $\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]$ can always be obtained by selecting a feasible solution $\mathbf{x} \in \mathcal{X}$, such as the one calculated from $\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))$, and then estimate $\mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]$ with a second independent sample.

6 Conclusion

We developed the dual DRO model that is asymptotically equivalent to mean-maxiance regularization. As illustrated with numerical example, dual DRO provides much more robustness against outliers than ϕ-divergence DRO due to its maxiance regularization effect. This makes dual DRO a more 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 DRO, and is 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

EC.1 Proofs

Proof of Proposition 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 neighbourhood of $0 = \phi'(1)$. It then follows from the implicit function theorem that in a neighbourhood of 0, there exists a continuously differentiable function $z(y)$, such that $z(0) = 1$ and $y - \phi'(z(y)) = 0$. Therefore, in a neighbourhood 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 1. Following a similar argument as in Theorem 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 infimimum is obtained at $\Delta y_i = \Delta x_i = 0$, where $0\phi\left(\frac{0}{0}\right) = 0$. By Proposition 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

$$
\begin{aligned}
\rho^{\phi}_{\delta,\mathbf{p}}(X) &= \frac{1}{\delta} \sum_{i=1}^{n} (-\phi)_*(\delta \overline{F}_i) \Delta x_i \\
&= \frac{1}{\delta} \sum_{i=1}^{n} \left[(-\phi)_*(0) + (-\phi)'_*(0) \delta \overline{F}_i + \frac{1}{2} (-\phi)''_*(0) \delta^2 (\overline{F}_i)^2 + o(\delta^2)(\overline{F}_i)^2 \right] \Delta x_i \\
&= \sum_{i=1}^{n} \overline{F}_i \Delta x_i - \frac{\delta}{2\phi''(1)} \sum_{i=1}^{n} (\overline{F}_i)^2 \Delta x_i + o(\delta) \cdot \sum_{i=1}^{n} (\overline{F}_i)^2 \Delta x_i \\
&= \sum_{i=1}^{n} \overline{F}_i \Delta x_i - \frac{\delta}{2\phi''(1)} \sum_{i=1}^{n} (\overline{F}_i)^2 \Delta x_i + o(\delta) \\
&= \mathbb{E}_{\mathbf{p}}[X] - \frac{\delta}{2\phi''(1)} \mathbb{E}_{\mathbf{p}}[\min\{X^{(1)}, X^{(2)}\}] + o(\delta)
\end{aligned}
$$

$$= \left(1 - \frac{\delta}{2\phi''(1)}\right) \mathbb{E}_{\mathbf{p}}[X] + \frac{\delta}{2\phi''(1)} \overline{\mathtt{m}}_{2,\mathbf{p}}(X) + o(\delta).$$

□

Proof of Proposition 2. 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{\mathtt{m}}_{2,\mathbf{p}}(X) + o(\delta).
\end{aligned}$$

□

Proof of Theorem 2. Since $\delta_1, \delta_2 \to 0$, we may assume without loss of generality, that $\delta_1 \leq 1$. Following the proof of Theorem 1, 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

$$\begin{aligned}\sum_{i=1}^{n}\left(\sum_{j=i}^{n} q_j\right)^2 &= \sum_{i=1}^{n}\left(\sum_{j=i}^{n} p_j + \sum_{j=i}^{n} q_j - p_j\right)^2 \\ &= \sum_{i=1}^{n}\left(\sum_{j=i}^{n} p_j\right)^2 + 2\sum_{j=i}^{n} p_j \sum_{k=i}^{n} q_k - p_k + \left(\sum_{j=i}^{n} q_j - p_j\right)^2 \\ &= -\sum_{i=1}^{n}\left(\sum_{j=i}^{n} p_j\right)^2 + 2\sum_{j=i}^{n} p_j \sum_{k=i}^{n} q_k + \left(\sum_{j=i}^{n} q_j - p_j\right)^2 .\end{aligned} \tag{EC.26}$$

Hence, we have that

$$\begin{aligned}\rho_{xp}(X) = \sup_{\mathbf{q}\in\mathbb{P}^n} \sum_{i=1}^{n} q_i x_i - \frac{\delta_1}{\phi''(1)} \sum_{j=i}^{n} p_j \sum_{k=i}^{n} q_k \Delta x_i - \frac{1}{\delta_2} p_i \psi\left(\frac{q_i}{p_i}\right) - \frac{\delta_1}{2\phi''(1)}\left(\sum_{j=i}^{n} q_j - p_j\right)^2 \Delta x_i \\ + \frac{\delta_1}{2\phi''(1)}\left(\sum_{j=i}^{n} p_j\right)^2 \Delta x_i + o(\delta_1).\end{aligned}$$

We note that the last two terms above do not depend on $\mathbf{q}$. Hence, we examine only

$$\begin{aligned}&\sup_{\mathbf{q}\in\mathbb{P}^n} \sum_{i=1}^{n} q_i x_i - \frac{\delta_1}{\phi''(1)} \sum_{j=i}^{n} p_j \sum_{k=i}^{n} q_k \Delta x_i - \frac{1}{\delta_2} p_i \psi\left(\frac{q_i}{p_i}\right) - \frac{\delta_1}{2\phi''(1)}\left(\sum_{j=i}^{n} q_j - p_j\right)^2 \Delta x_i \\ &\triangleq \sup_{\mathbf{q}\in\mathbb{P}^n} E(\mathbf{q}).\end{aligned}$$

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 $E(\mathbf{p}) \geq \sum_{i=1}^{n} p_i x_i - \frac{1}{\phi''(1)} x_{\max}$. Therefore, we have that for all $\delta_2 > 0$, if $\mathbf{q}\in\mathbb{P}^n$ is any probability vector such that

$$\sum_{i=1}^{n} p_i \psi\left(\frac{q_i}{p_i}\right) > \delta_2 K,$$

where $K > 0$ is any constant (independent of δ_1, δ_2), such that

$$K \geq \left(1 + \frac{4}{\phi''(1)}\right) x_{\max} - \sum_{i=1}^{n} p_i x_i. \tag{EC.27}$$

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 (EC.27), 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 EC.1.1, 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

$$\sup_{\mathbf{q}\in\mathbb{P}^n:\sum_{i=1}^n p_i\psi\left(\frac{q_i}{p_i}\right)\leq\delta_2 K} E_l(\mathbf{q}) = \sup_{\mathbf{q}\in\mathbb{P}^n} E_l(\mathbf{q}). \tag{EC.28}$$

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 (EC.27) also ensures (EC.28).

We now examine:

$$\begin{aligned}
\sup_{\mathbf{q}\in\mathbb{P}^n} E_l(\mathbf{q}) &= \sup_{\mathbf{q}\in\mathbb{P}^n} \sum_{i=1}^n q_i x_i - \frac{\delta_1}{\phi''(1)} \sum_{j=i}^n p_j \sum_{k=i}^n q_k \Delta x_i - \frac{1}{\delta_2} p_i \psi\left(\frac{q_i}{p_i}\right) \\
&= \sup_{\mathbf{q}\in\mathbb{P}^n} \sum_{i=1}^n \left(x_i - \frac{\delta_1}{\phi''(1)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j \right) q_i - \frac{1}{\delta_2} p_i \psi\left(\frac{q_i}{p_i}\right) \\
&= \inf_{\lambda\in\mathbb{R}} \lambda + \sup_{\mathbf{q}\in\mathbb{R}^n_+} \sum_{i=1}^n \left(x_i - \frac{\delta_1}{\phi''(1)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j - \lambda \right) q_i - \frac{1}{\delta_2} p_i \psi\left(\frac{q_i}{p_i}\right) \\
&= \inf_{\lambda\in\mathbb{R}} \lambda + \frac{1}{\delta_2} \sum_{i=1}^n p_i \sup_{q_i\in\mathbb{R}_+} \left(\delta_2(x_i - \lambda) - \frac{\delta_1\delta_2}{\phi''(1)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j \right) \frac{q_i}{p_i} - \psi\left(\frac{q_i}{p_i}\right) \\
&= \inf_{\lambda\in\mathbb{R}} \lambda + \frac{1}{\delta_2} \sum_{i=1}^n p_i \sup_{t\in\mathbb{R}_+} \left(\delta_2(x_i - \lambda) - \frac{\delta_1\delta_2}{\phi''(1)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j \right) t - \psi(t) \\
&= \inf_{\lambda\in\mathbb{R}} \lambda + \frac{1}{\delta_2} \sum_{i=1}^n p_i \psi^* \left(\delta_2(x_i - \lambda) - \frac{\delta_1\delta_2}{\phi''(1)} \sum_{j=1}^i \sum_{k=j}^n p_k \Delta x_j \right).
\end{aligned} \tag{EC.29}$$

By proposition 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 on any compact set $[-b, b]$ with b sufficiently large (see Lemma EC.1.2). 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

$$\inf_{\lambda\in[-b,b]} \lambda + \frac{1}{\delta_2} \sum_{i=1}^n p_i \psi^* (\delta_2 B_i(\lambda)) = \inf_{\lambda\in[-b,b]} \lambda + \frac{1}{\delta_2} \sum_{i=1}^n p_i (\delta_2 B_i(\lambda)) + p_i \frac{1}{2\psi''(1)} \delta_2^2 (B_i(\lambda))^2 + o(\delta_2^2).$$

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 (EC.26) and (EC.29). 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

$$\begin{aligned}
\rho_{xp}(X) &= \mathbb{E}_{\mathbf{p}}[X] + \frac{\delta_2}{2\psi''(1)}\mathrm{Var}_{\mathbf{p}}(X) - \frac{\delta_1}{2\phi''(1)}\sum_{i=1}^{n}\left(\sum_{j=i}^{n} p_j\right)^2 \Delta x_i + o(\delta_1) + o(\delta_2) + O(\delta_1\delta_2\cdot n) \\
&= \mathbb{E}_{\mathbf{p}}[X] + \frac{\delta_2}{2\psi''(1)}\mathrm{Var}_{\mathbf{p}}(X) - \frac{\delta_1}{2\phi''(1)}\mathbb{E}_{\mathbf{p}}\left[\min\{X^{(1)}, X^{(2)}\}\right] + o(\delta_1) + o(\delta_2) + O(\delta_1\delta_2\cdot n) \\
&= \left(1 - \frac{\delta_1}{2\phi''(1)}\right)\mathbb{E}_{\mathbf{p}}[X] + \frac{\delta_2}{2\psi''(1)}\mathrm{Var}_{\mathbf{p}}(X) + \frac{\delta_1}{2\phi''(1)}\overline{\mathbb{m}}_{2,\mathbf{p}}(X) + o(\delta_1) + o(\delta_2) + O(\delta_1\delta_2\cdot n).
\end{aligned}$$

□

Lemma EC.1.1. *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 $p_{\min} := \min_i p_i > 0$.

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

$i = 1, \ldots, n$

$$p_i \psi\left(\frac{q_i}{p_i}\right) \leq \sum_{i=1}^{n} p_i \psi\left(\frac{q_i}{p_i}\right) \leq \delta_2 K.$$

Hence, for all i, we have

$$\psi\left(\frac{q_i}{p_i}\right) \leq \delta_2 K/p_i \leq \delta_2 K/p_{\min}.$$

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

$$\psi\left(\frac{q_i}{p_i}\right) \leq \delta_2 K/p_{\min} \Rightarrow \left|\frac{q_i}{p_i} - 1\right| \leq \epsilon, \forall i = 1, \ldots, n.$$

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

$$\begin{aligned}\sum_{i=1}^{n} p_i \psi\left(\frac{q_i}{p_i}\right) \leq \delta_2 K &\Rightarrow \left|\frac{q_i}{p_i} - 1\right| \leq \epsilon, \forall i \\ &\Rightarrow \left|R_2\left(\frac{q_i}{p_i}\right)\right| \leq \frac{\psi''(1)}{4}, \forall i.\end{aligned}$$

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 EC.1.2. *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 on 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

$$\sum_{i=1}^{n} p_i(\psi^*)'\left(\delta_2\left(x_i - \frac{\delta_1}{\psi''(0)}\sum_{j=1}^{i}\sum_{k=j}^{n} p_k \Delta x_j - \lambda^*\right)\right) = 1. \qquad \text{(EC.30)}$$

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 (EC.30) 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 on a compact set $[-b, b]$ with b sufficiently large. □

Proof of Proposition 3. 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,

$$\left\{\omega : \inf_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta_n(\omega)}} \sum_{i=1}^{n} q_i l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) \leq c\right\}$$

$$
\begin{aligned}
&= \bigcap_{n\geq 1} \bigcup_{\mathbf{x}\in\mathcal{X}} \left\{ \omega : \sup_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta_n(\omega)}} \sum_{i=1}^{n} q_i l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) \leq c + \frac{1}{n} \right\} \\
&= \bigcap_{n\geq 1} \bigcup_{\mathbf{x}\in\mathcal{X}} \left\{ \omega : \sum_{i=1}^{n} \frac{(-\phi)_* \left(\delta(\omega)\cdot \frac{n-i+1}{n}\right)}{(-\phi)_*(\delta(\omega))} \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)) \leq c + \frac{1}{n} \right\} \\
&= \bigcap_{n\geq 1} \bigcup_{\sigma\in\Pi(n)} \bigcup_{\mathbf{x}\in\mathcal{X}} \left\{ \omega : \sum_{i=1}^{n} \frac{(-\phi)_* \left(\delta(\omega)\cdot \frac{n-i+1}{n}\right)}{(-\phi)_*(\delta(\omega))} \Delta l(\mathbf{x}, \boldsymbol{\xi}_{\sigma(i)}(\omega)) \leq c + \frac{1}{n} \right\} \\
&\qquad \cap \left\{ \omega : l(\mathbf{x}, \boldsymbol{\xi}_{\sigma(1)}(\omega)) \leq \ldots \leq l(\mathbf{x}, \boldsymbol{\xi}_{\sigma(n)}(\omega)) \right\},
\end{aligned}
$$

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 5. 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 1, we have that for all possible realisations $\omega \in \Omega$, that

$$
\begin{aligned}
\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi}))(\omega) &= \frac{1}{\delta(n)(\omega)} \sum_{i=1}^{n} (-\phi)_* \left(\delta(n)(\omega) \frac{n-i+1}{n} \right) \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)) \\
&= \frac{1}{\delta(n)(\omega)} \sum_{i=1}^{n} \left(\delta(n)(\omega)\frac{n-i+1}{n} - \frac{1}{2\phi''(1)} \left(\delta(n)(\omega)\frac{n-i+1}{n} \right)^2 \right. \\
&\qquad \left. + \frac{(\phi^*)'''(\tilde{\xi}_n)}{6} \left(\delta(n)(\omega)\frac{n-i+1}{n} \right)^3 \right) \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)) \\
&= \frac{1}{n} \sum_{i=1}^{n} l(\mathbf{x}, \boldsymbol{\xi}_i(\omega)) - \frac{\delta(n)(\omega)}{2\phi''(1)} \sum_{i=1}^{n} \left(\frac{n-i+1}{n} \right)^2 \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)) \\
&\qquad + \sum_{i=1}^{n} \frac{(\phi^*)'''(\tilde{\xi}_n)(\omega)}{6} \delta^2(n)(\omega) \left(\frac{n-i+1}{n} \right)^3 \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)).
\end{aligned}
$$

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, it's third derivative is bounded in a neighbourhood 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

$$
\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{\delta(n)(\omega)} |\epsilon_n(\mathbf{x})(\omega)| &\leq C \cdot \delta(n)(\omega) \sup_{\mathbf{x}\in\mathcal{X}} \sum_{i=1}^{n} \frac{n-i+1}{n} \Delta l(\mathbf{x}, \boldsymbol{\xi}_{i(\mathbf{x})}(\omega)) \\
&= C \cdot \delta(n)(\omega) \sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} l(\mathbf{x}, \boldsymbol{\xi}_i(\omega))
\end{aligned}
$$

$$\leq C \cdot \delta(n)(\omega) \frac{1}{n} \sum_{i=1}^{n} M_2(\boldsymbol{\xi}_i(\omega)),$$

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 $\rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}$):

$$\begin{aligned}\mathbb{P}_0\left(\sup_{\mathbf{x}\in f\mathcal{X}} \frac{1}{\delta(n)}|\epsilon_n(\mathbf{x})| > \tilde{\epsilon}\right) &\leq \mathbb{P}_0\left(\left\{\sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{\delta(n)}|\epsilon_n(\mathbf{x})| > \tilde{\epsilon}\right\} \cap \{\delta(n) \leq \epsilon_0\}\right) \\ &+ \mathbb{P}_0\left(\left\{\sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{\delta(n)}|\epsilon_n(\mathbf{x})| > \tilde{\epsilon}\right\} \cap \{\delta(n) > \epsilon_0\}\right).\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}_0^*\left(\left\{\sup_{\mathbf{x}\in\mathcal{X}} \frac{1}{\delta(n)}|\epsilon_n(\mathbf{x})| > \tilde{\epsilon}\right\} \cap \{\delta(n) \leq \epsilon_0\}\right) &\leq \mathbb{P}_0\left(\left\{\delta(n)\frac{1}{n}\sum_{i=1}^{n} M_2(\boldsymbol{\xi}_i) > \frac{\tilde{\epsilon}}{C}\right\} \cap \{\delta(n) \leq \epsilon_0\}\right) \\ &\leq \mathbb{P}_0\left(\delta(n)\frac{1}{n}\sum_{i=1}^{n} M_2(\boldsymbol{\xi}_i) > \frac{\tilde{\epsilon}}{C}\right).\end{aligned}$$

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 6. Let $\delta(n) = \sqrt{r/n}$. Using Theorem 5, 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 EC.1.8, 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 5 that we also have the weak convergence

$$\sqrt{n}\left(\mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\cdot,\boldsymbol{\xi})) - \mathbb{E}_{\mathbb{P}_0}[l(\cdot,\boldsymbol{\xi})]\right) \rightsquigarrow G(\cdot) + \frac{\sqrt{r}}{2\phi''(1)}\overline{\mathtt{m}}_{2,\mathbb{P}_0}(\cdot), \text{ in } l^{\infty}(\mathcal{H}).$$

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 Proposition 1, we have that

$$\begin{aligned}&\min_{\mathbf{x}\in\mathcal{X}} \rho^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] \\ &\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta(n)}} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \sum_{i=1}^{n} q_i l(\mathbf{x},\boldsymbol{\xi}_i) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\end{aligned}$$

$$
\begin{aligned}
&= \max_{\mathbf{q}\in\mathcal{M}^{\phi}_{\delta(n)}} \min_{\mathbf{x}\in\mathcal{X}} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \sum_{i=1}^{n} q_i l(\mathbf{x}, \boldsymbol{\xi}_i) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x}, \boldsymbol{\xi})] \\
&= \max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} T(\mathbb{Q}) - T(\mathbb{P}_0),
\end{aligned}
$$

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 $\mathbf{q} \in \mathcal{M}^{\phi}_{\delta(n)}$ (as defined in (19)) with a measure $\mathbb{Q} = \sum_{i=1}^{n} q_i \iota_{\boldsymbol{\xi}_i}$ and we denote the set

$$
\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n) \triangleq \left\{ \sum_{i=1}^{n} q_i \iota_{\boldsymbol{\xi}_i} \;\middle|\; \mathbf{q} \in \mathcal{M}^{\phi}_{\delta(n)} \right\},
$$

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 with examining

$$
\begin{aligned}
&\frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \left| \max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} (T(\mathbb{Q}) - T(\mathbb{P}_0)) - \max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x}, \mathbb{P}_0)] \right| \\
&\le \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \left| \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \left(T(\mathbb{Q}) - T(\mathbb{P}_0) - \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x}, \mathbb{P}_0)] \right) \right| \\
&\triangleq \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})|.
\end{aligned}
$$

Our first goal is to show that for any $\epsilon > 0$,

$$
\limsup_{n\to\infty} \mathbb{P}^*_0 \left(\sqrt{n} \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})| \ge \epsilon \right) = 0. \tag{EC.31}
$$

To do this, we choose for any $\delta' > 0$, a sequence of measurable selection (see Lemma EC.1.6) $\mathbb{Q}_n \in \mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)$, such that $|\kappa(\mathbb{Q}_n)| \ge (1-\delta') \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})|$. This gives the bound

$$
\mathbb{P}^*_0 \left(\sqrt{n} \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})| \ge \epsilon \right) \le \mathbb{P}^*_0 \left(\sqrt{n} |\kappa_n(\mathbb{Q}_n)| \ge (1-\delta')\epsilon \right).
$$

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 EC.1.5, that all subsequences of $\sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)$ are asymptotically tight and measurable. It follows from the 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

$$
\limsup_{n\to\infty} \mathbb{P}^*_0 \left(\sqrt{n} |\kappa_n(\mathbb{Q}_n)| \ge (1-\delta')\epsilon \right) = \lim_{m\to\infty} \mathbb{P}^*_0 \left(\sqrt{n(m)} |\kappa_{n(m)}(\mathbb{Q}_{n(m)})| \ge (1-\delta')\epsilon \right) = 0,
$$

where the latter is obtained by applying the functional delta theorem (Theorem 1, Römisch 2006) to conclude that

$$\sqrt{n(m)}\left(T(\mathbb{Q}_{n(m)}) - T(\mathbb{P}_0) - \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \mathbb{E}_{\mathbb{Q}_{n(m)}}[IF(\mathbf{x},\mathbb{P}_0)]\right) \xrightarrow{P^*} 0.$$

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 (EC.31), 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

$$\begin{aligned}
&\inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \sqrt{n}\left(\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\sqrt{r/n})}{\sqrt{r/n}} \mathbb{E}_{\mathbb{Q}}[l(\mathbf{x},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right) + \frac{\sqrt{r}}{2\phi''(1)} T(\mathbb{P}_0) - \frac{\sqrt{r}}{2\phi''(1)} T(\mathbb{P}_0) + o_p(1)\\
&\rightsquigarrow \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} G(\mathbf{x}) - \frac{\sqrt{r}}{2\phi''(1)} \mathrm{dm}_{2,\mathbb{P}_0}(\mathbf{x}),
\end{aligned}$$

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 $|\inf_{h\in\mathcal{H}^*} f(h) - \inf_{h\in\mathcal{H}^*} g(h)| \leq \sup_{h\in\mathcal{H}} |(f-g)(h)| \leq \epsilon$, where $\mathcal{H}^* = \{l(x,.) : x \in \mathcal{X}^*_{\mathbb{P}_0}\} \subset \mathcal{H}$.

Similarly, we have

$$\begin{aligned}
&\sqrt{n}\left(\min_{\mathbf{x}\in\mathcal{X}} \mathcal{R}^{\phi}_{\delta(n),\hat{\mathbb{P}}_n}(l(\mathbf{x},\boldsymbol{\xi})) - \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})]\right)\\
&= \sqrt{n}\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} T(\mathbb{Q}) - T(\mathbb{P}_0)
\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{n} \max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x},\mathbb{P}_0)] + o_p(1) \\
&= \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \sqrt{n} \max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x},\mathbb{P}_0)] + o_p(1) \\
&= \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} \sqrt{n} \left(\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{Q}}[l(\mathbf{x},\boldsymbol{\xi})] - \mathbb{E}_{\mathbb{P}_0}[l(\mathbf{x},\boldsymbol{\xi})] \right) + o_p(1) \\
&\rightsquigarrow \inf_{\mathbf{x}\in\mathcal{X}^*_{\mathbb{P}_0}} G(\mathbf{x}) + \frac{\sqrt{r}}{2\phi''(1)} \overline{\mathfrak{m}}_{2,\mathbb{P}_0}(\mathbf{x}).
\end{aligned}
$$

□

The proof of above theorems are supplemented with a couple of technical Lemmas.

Lemma EC.1.3. *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 (19). Suppose $\delta(n) = \sqrt{r/n}$. Then, we have that for any $i =, 1 \ldots, n$,*

$$
n\sqrt{n}\left| q_i - \frac{1}{n} \right| \le \frac{\sqrt{r}}{2\phi''(1)} + o(1).
$$

Proof. Fix $i \in [n]$. By definition of $\mathcal{M}^{h_{\delta(n)}}(\hat{\mathbf{p}})$ (see definition in (17)), we have that

$$
\begin{aligned}
q_i \le h_{\delta(n)}(1/n) = \frac{(-\phi)_*(\delta(n)1/n)}{(-\phi)_*(\delta(n))} = \frac{(-\phi)_*\left(\sqrt{\frac{r}{n}}\frac{1}{n}\right)}{(-\phi)_*\left(\sqrt{\frac{r}{n}}\right)} &= \frac{\sqrt{\frac{r}{n}}\frac{1}{n} - \frac{1}{2\phi''(1)}\frac{r}{n}\frac{1}{n^2} + O(\frac{r\sqrt{r}}{n^4\sqrt{n}})}{\sqrt{\frac{r}{n}} - \frac{1}{2\phi''(1)}\frac{r}{n} + O(\frac{r}{n}\sqrt{\frac{r}{n}})} \\
&= \frac{\frac{1}{n} - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\frac{1}{n^2} + O(\frac{r}{n^4})}{1 - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}} + O(\frac{r}{n})}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
q_i - \frac{1}{n} &\le \frac{1}{1 - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}} + O(\frac{r}{n})} \left(\frac{1}{2\phi''(1)} \frac{1}{n} \sqrt{\frac{r}{n}} \left(1 - \frac{1}{n}\right) + O\left(\frac{r}{n^2}\right) \right) \\
&= \frac{\sqrt{r}}{2\phi''(1)} \frac{1}{n\sqrt{n}} + o\left(\frac{1}{n\sqrt{n}}\right).
\end{aligned}
$$

On the other hand, we have

$$
\begin{aligned}
q_i = 1 - \sum_{j\neq i} q_j &\ge 1 - h_{\delta(n)}\left(\sum_{j\neq i} 1/n\right) \\
&= 1 - \frac{(-\phi)_*\left(\sqrt{\frac{r}{n}}\left(1 - \frac{1}{n}\right)\right)}{(-\phi)_*\left(\sqrt{\frac{r}{n}}\right)} \\
&= 1 - \frac{\sqrt{\frac{r}{n}}\left(1-\frac{1}{n}\right) - \frac{1}{2\phi''(1)}\frac{r}{n}\left(1-\frac{1}{n}\right)^2 - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n}\sqrt{\frac{r}{n}}\left(1-\frac{1}{n}\right)^3 + o\left(\frac{r^2}{n^2}\right)}{\sqrt{\frac{r}{n}} - \frac{1}{2\phi''(1)}\frac{r}{n} - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n}\sqrt{\frac{r}{n}} + o\left(\frac{r^2}{n^2}\right)} \\
&= 1 - \frac{\left(1-\frac{1}{n}\right) - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\left(1-\frac{1}{n}\right)^2 - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n}\left(1-\frac{1}{n}\right)^3 + o\left(\frac{r^2}{n\sqrt{n}}\right)}{1 - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}} - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n} + o\left(\frac{r^2}{n\sqrt{n}}\right)}.
\end{aligned}
$$

Hence,

$$\begin{aligned}
q_i - \frac{1}{n} &\geq \left(1 - \frac{1}{n}\right) - \frac{\left(1-\frac{1}{n}\right) - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\left(1-\frac{1}{n}\right)^2 - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n}\left(1-\frac{1}{n}\right)^3 + o\left(\frac{r^2}{n\sqrt{n}}\right)}{1 - \frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}} - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n} + o\left(\frac{r^2}{n\sqrt{n}}\right)} \\
&= \frac{1}{1+o\left(1\right)}\left(-\frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\frac{1}{n}\left(1-\frac{1}{n}\right) - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n}\left(1-\frac{1}{n}\right)\left(1-\left(1-\frac{1}{n}\right)^2\right) + o\left(\frac{r^2}{n\sqrt{n}}\right)\right) \\
&= (1+o(1))\left(-\frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\frac{1}{n}\left(1-\frac{1}{n}\right) - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r}{n^2}\left(1-\frac{1}{n}\right)\left(2-\frac{1}{n}\right) + o\left(\frac{r^2}{n\sqrt{n}}\right)\right) \\
&= (1+o(1))\left(-\frac{1}{2\phi''(1)}\sqrt{\frac{r}{n}}\frac{1}{n}\left(1-\frac{1}{n}\right) + o\left(\frac{1}{n\sqrt{n}}\right)\right) \\
&= -\frac{\sqrt{r}}{2\phi''(1)}\frac{1}{n\sqrt{n}} + o\left(\frac{1}{n\sqrt{n}}\right).
\end{aligned}$$

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 EC.1.4. *Let ϕ be four times continuously differentiable around* 1*. Suppose $\delta(n) = \sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)/n}$ where $r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n) \overset{P}{\to} r_0$ 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 $\mathbf{q} \in \mathcal{M}^{\phi}_{\delta(n)}$ 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 EC.1.3, 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

$$\begin{aligned}
&n\sqrt{n}\left(q_i - \frac{1}{n}\right) \\
&\geq \frac{1}{1+o_p\left(1\right)}\left(-\frac{\sqrt{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}}{2\phi''(1)}\left(1-\frac{1}{n}\right) - \frac{\phi'''(1)}{(\phi''(1))^3}\frac{r(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)}{\sqrt{n}}\left(1-\frac{1}{n}\right)\left(2-\frac{1}{n}\right) + o_p(1)\right).
\end{aligned}$$

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 EC.1.5. *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 $\omega \in \Omega$. 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, for any sequence $\mathbb{Q}_n \in \mathcal{M}^{\phi}_{\delta(n)}(\hat{\mathbb{P}}_n)$, 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

$$\begin{aligned}
&\limsup_{\delta \to 0, n \to \infty} \mathbb{P}_0^* \left(\sup_{\|h - h'\| < \delta} \left| \sqrt{n}(\mathbb{Q}_n - \mathbb{P}_0)(h - h') \right| > \epsilon \right) \\
&\leq \limsup_{\delta \to 0, n \to \infty} \mathbb{P}_0^* \left(\sup_{\|h - h'\| < \delta} \left| \sqrt{n}(\mathbb{Q}_n - \hat{\mathbb{P}}_n)(h - h') \right| > \epsilon/2 \right) + \mathbb{P}_0^* \left(\sup_{\|h - h'\| < \delta} \left| \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_0)(h - h') \right| > \epsilon/2 \right).
\end{aligned}$$

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 EC.1.4 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 $e_n, \tilde{e}_n \overset{P}{\to} 0$. 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 EC.1.4 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 implies 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 an 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 EC.1.6. *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 EC.1.7. *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

$$\begin{aligned} \sum_{i=1}^{n} \left(\frac{n-i+1}{n} \right)^2 \Delta X_{(i)} &= \frac{1}{n^2} \sum_{1 \leq i \leq j \leq n} \min\{X_i, X_j\} \\ &= \frac{1}{n^2} \sum_{i=1}^{n} X_i + \frac{2}{n^2} \sum_{i<j} \min\{X_i, X_j\} \\ &= \frac{1}{n^2} \sum_{i=1}^{n} X_i + \frac{n-1}{n} \frac{2}{n(n-1)} \sum_{i<j} \min\{X_i, X_j\}. \end{aligned}$$

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 variable $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}$: $\mathrm{dm}_{2,\mathbb{P}}(\mathbf{x}) = \mathbb{E}_{\mathbb{P}\times\mathbb{P}}[\min\{l(\mathbf{x}, \boldsymbol{\xi}^{(1)}, l(\mathbf{x}, \boldsymbol{\xi}^{(2)}\}]$.

Lemma EC.1.8. *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*

$$\mathrm{dm}_{2,\hat{\mathbb{P}}_n}(.) \rightsquigarrow \mathrm{dm}_{2,\mathbb{P}_0}(.), \text{ in } l^\infty(\mathcal{H}). \tag{EC.32}$$

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 (EC.32) using Theorem 1.5.4 of van der Vaart and Wellner [2023]. The convergence of the marginals follows from Lemma EC.1.7. 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 equicontinuity in probability, and proves (EC.32). □

Lemma EC.1.9. *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 the Berge's maximum theorem (Berge 1963). □

EC.2 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 from 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 EC.1.7 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 7. *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 an unique solution $\mathbf{x}^*$, and we have a consistent estimator $\sqrt{r(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}$ 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 7. We revisit the proof of Theorem 6. Let $T(\mathbb{P}) \triangleq \min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}}[l(\mathbf{x},\boldsymbol{\xi})]$. Using Proposition 1 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

$$\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n) = \left\{\sum_{i=1}^n q_i \iota_{\boldsymbol{\xi}_i} \;\middle|\; q_i \geq 0, \sum_{i=1}^n q_i = 1, \sum_{j\in J} q_j \leq \frac{(-\phi)_*\left(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)\cdot\frac{|J|}{n}\right)}{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}, \forall J \subset [n]\right\}, \tag{EC.33}$$

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

$$\sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})| \triangleq \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} \left| \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} \left(T(\mathbb{Q}) - T(\mathbb{P}_0) - \mathbb{E}_{\mathbb{Q}}[IF(\mathbf{x}^*, \mathbb{P}_0)]\right) \right|.$$

Our first goal is to show that for any $\epsilon_0 > 0$, we have

$$\limsup_{n\to\infty} \mathbb{P}_0^* \left(\sqrt{n} \sup_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} |\kappa_n(\mathbb{Q})| > \epsilon_0 \right) = 0. \tag{EC.34}$$

We fix a $\delta_0 > 0$, and choose a measurable selection (see Lemma EC.1.6) $\mathbb{Q}_n(\omega) \in \mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1(\omega),\ldots,\boldsymbol{\xi}_n(\omega))}(\hat{\mathbb{P}}_n)$ 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 EC.1.5, 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 (EC.34). Using Theorem 5, we thus have that in a similar argument as in the proof of Theorem 6, that

$$\sqrt{n} \left| \left(\max_{\mathbb{Q}\in\mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} T(\mathbb{Q}) - T(\mathbb{P}_0) \right) + \left(1 - \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} \right) T(\mathbb{P}_0) \right.$$
$$\left. - \mathbb{E}_{\hat{\mathbb{P}}_n}[IF(\mathbf{x}^*, \mathbb{P}_0)] + \frac{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}{2\phi''(1)} \mathbb{E}_{\hat{\mathbb{P}}_n}[\min\{IF^{(1)}(\mathbf{x}^*, \mathbb{P}_0), IF^{(2)}(\mathbf{x}^*, \mathbb{P}_0)\}] \right| \xrightarrow{P^*} 0.$$

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

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}_0 \left(\sqrt{n} \max_{\mathbb{Q} \in \mathcal{M}^{\phi}_{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}(\hat{\mathbb{P}}_n)} \frac{(-\phi)_*(\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n))}{\delta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)} T(\mathbb{Q}) - T(\mathbb{P}_0) \le 0 \right) \\
&= \mathbb{P}\left(N(0, \mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))) - \Phi^{-1}(\alpha)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*, \boldsymbol{\xi}))} \le 0 \right) \\
&= \Phi(\Phi^{-1}(\alpha)) = \alpha.
\end{aligned}$$

Finally, if instead we have that $\sqrt{r(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_n)}$ converges in probability to $\frac{2\Phi^{-1}(\alpha)\phi''(1)\sqrt{\mathrm{Var}_{\mathbb{P}_0}(l(\mathbf{x}^*,\boldsymbol{\xi}))}}{\bar{\mathtt{m}}_{2,\mathbb{P}_0}(l(\mathbf{x}^*,\boldsymbol{\xi}))}$, 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}))}}{\bar{\mathtt{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}))}}{\bar{\mathtt{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}$$

□

References

Daniel Bartl and Shahar Mendelson. On monte-carlo methods in convex stochastic optimization. *Annals of Applied Probability*, 32(4):3146–3198, 2022.

Aharon Ben-Tal and Marc Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

Aharon Ben-Tal, Adi Ben-Israel, and Marc Teboulle. Certainty equivalents and information measures:

Duality and extremal Principles. *Journal of Mathematical Analysis And Applications*, 157(1):211–236, 1991.

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

Claude Berge. *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*. Courier Corporation, 1963.

Ruidi Chen and Ioannis Ch. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13):1–48, 2018. URL http://jmlr.org/papers/v19/17-295.html.

Dieter Denneberg. *Non-Additive Measure and Integral*. Springer, 1994.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019. URL http://jmlr.org/papers/v20/17-750.html.

John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

Louis R. Eeckhoudt and Roger J.A. Laeven. Dual moments and risk attitudes. *Operations Research*, 70 (3):1330–1341, 2021.

Louis R. Eeckhoudt, Roger J.A. Laeven, and Harris Schlesinger. Risk apportionment: The dual story. *Journal of Economic Theory*, 185:104971, 2020.

Rui Gao, Xi Chen, and Anton J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2022.

Corrado Gini. *Variabilità e Mutuabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. C. Cuppini, Bologna, 1912.

Corrado Gini. Measurement of inequality of incomes. *The Economic Journal*, 31(121):124–126, 1921.

Jun-Ya Gotoh, Michael Jong Kim, and Andrew E.B. Lim. Robust empirical optimization is almost the same as mean-variance optimization. *Operations Research Letters*, 46:448–452, 2018.

Jun-Ya Gotoh, Michael Jong Kim, and Andrew E.B. Lim. Worst-case sensitivity. *Paper*, 2020. Available on ArXiv.

Jun-Ya Gotoh, Michael Jong Kim, and Andrew E. B. Lim. Calibration of distributionally robust empirical optimization models. *Operations Research*, 69(5):1630–1650, 2021.

Guanyu Jin, Roger J.A. Laeven, and Dick den Hertog. Robust optimization of rank-dependent models with uncertain probabilities, 2025. Available on arXiv.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: theory and applications in machine learning. *Informs Tutorials in Operations Research*, 0(0):130–166, 2019.

Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

Ingram Olkin and Shlomo Yitzhaki. Gini regression analysis. *International Statistical Review*, 60(2): 185–196, 1992.

Krzysztof Postek, Dick den Hertog, and Bertrand Melenberg. Computationally tractable counterparts of distributionally robust constraints on risk measures. *SIAM Review*, 58(4):603–650, 2016.

John Quiggin. A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3(4): 323–343, 1982.

Ulrich Rieder. Measurable selection theorems for optimization problems. *Manuscripta Mathematica*, 24: 115–131, 1978.

Werner Römisch. Delta method, infinite dimensional. In *Encyclopedia of Statistical Sciences*, volume 16. Wiley, New York, 2006.

David Schmeidler. Integral representation without additivity. *Proceedings of the American Mathematical Society*, 97:255–261, 1986.

David Schmeidler. Subjective probability and expected utility without additivity. *Econometrica*, 57(3): 571–587, 1989.

Haim Shalit and Shlomo Yitzhaki. Mean-gini, portfolio theory, and the pricing of risky assets . *The Journal of Finance*, 39(5):1449–1468, 1984.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures On Stochastic Programming*. SIAM, 2009.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, 2023.

Bart P.G. van Parys and Bert Zwart. Robust mean estimation for optimization: The impact of heavy tails, 2025. Available on arXiv.

Menahem E. Yaari. The dual theory of choice under risk. *Econometrica*, 55(1):95–115, 1987.

Shlomo Yitzhaki. Stochastic dominance, mean variance, and Gini's mean difference. *The American Economic Review*, 72(1):178–185, 1982.

Shlomo Yitzhaki and Edna Schechtman. *The Gini Methodology: A Primer on a Statistical Methodology*. Springer New York, 2012.