Distributionally Robust Shortfall Risk Optimization Model and Its Approximation [*]

Shaoyan Guo [†] and Huifu Xu[‡]

February 20, 2018

Abstract

Utility-based shortfall risk (SR) measures have received increasing attention over the past few years for its potential to quantify the risk of large tail losses more effectively than conditional value at risk. In this paper, we consider a distributionally robust version of the shortfall risk (DRSR) measure where the true probability distribution is unknown and the worst distribution from an ambiguity set of distributions is used to calculate the SR. We start by showing that the DRSR is a convex risk measure and under some special circumstance a coherent risk measure. We then move on to study an optimization problem with the objective of minimizing the DRSR of a random function and investigate numerical tractability of the optimization problem with the ambiguity set being constructed in various ways including moment conditions, ϕ-divergence, Kantorovich metric and mixture distribution. In the case when the underlying random variables are continuously distributed, we propose some discrete approximation schemes for the ambiguity sets and derive error bounds for the approximation under the Kantorovich metric. Quantitative convergence of the optimal values of the approximation problems is consequently established under moderate conditions. Specifically, we show that the error of the optimal value is linearly bounded by the error of each of the approximate ambiguity sets and subsequently derive a confidence interval of the optimal value under each of the approximation schemes. Some preliminary numerical test results are reported for the proposed modeling and computational schemes.

keywords DRSR, Kantorovich metric, moment conditions, ϕ-divergence ball, Kantorovich ball, mixture distribution, quantitative convergence analysis

1 Introduction

Quantitative measure of risk is a key element for financial institutions and regulatory authorities. It provides a way to compare different financial positions. A financial position can be mathematically characterized by a random variable $Z : (\Omega, \mathscr{F}, P) \to {\rm I\!R}$, where Ω is a sample space with sigma algebra $\mathscr{F}$ and P is a probability measure. A risk measure ρ assigns to Z a number that signifies the risk of the position. A good risk measure should have some virtues, such as being sensitive to excessive losses, penalizing concentration and encouraging diversification, and supporting dynamically consistent risk managements over multiple horizons [15].

[*] The research is supported by EPSRC grant EP/M003191/1.

[†] School of Mathematical Sciences, Dalian University of Technology, Dalian, 116024,China.(syguo@dlut.edu.cn).

[‡] School of Mathematics, University of Southampton, Southampton, SO17 1BJ, UK. (h.xu@soton.ac.uk).

Artzner et al. [1] considered the axiomatic characterizations of risk measures and first introduced the concept of coherent risk measure, which satisfies: (a) positive homogeneity ($\rho(\alpha Z) = \alpha\rho(Z)$ for $\alpha \geq 0$); (b) subadditivity ($\rho(Z+Y) \leq \rho(Z)+\rho(Y)$); (c) monotonicity (if $Z \geq Y$, then $\rho(Z) \leq \rho(Y)$); (d) translation invariance (if $m \in \mathbb{R}$, then $\rho(Z+m) = \rho(Z) - m$). Frittelli and Rosazza Gianin [12], Heath [17] and Föllmer and Schied [9] extended the notion of coherent risk measure to convex risk measure by replacing the positive homogeneity and the subadditivity with convexity, that is, $\rho(\alpha Z + (1-\alpha)Y) \leq \alpha\rho(Z) + (1-\alpha)\rho(Y)$, for all $\alpha \in [0,1]$. Obviously positive homogeneity and subadditivity imply convexity but not vice versa. In other words, a coherent risk measure is a convex risk measure but conversely it may not be true.

A well-known coherent risk measure is conditional value at risk (CVaR) defined by $\mathrm{CVaR}_\alpha(Z) := \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\lambda(Z)d\lambda$, where $\mathrm{VaR}_\lambda(Z)$ denotes the value at risk (VaR) which in this context is the smallest amount of cash that needs to be added to Z such that the probability of the financial position falling into a loss does not exceed a specified level λ, that is, $\mathrm{VaR}_\lambda(Z) := \inf\{t \in \mathbb{R} : P(Z+t<0) \leq \lambda\}$. In a financial context, CVaR has a number of advantages over the commonly used VaR, and CVaR has been proposed as the primary tool for banking capital regulation in the draft Basel III standard [2]. However, CVaR has a couple of obvious deficiencies.

One is that CVaR is not invariant under randomization, a property which is closely related to the weak dynamic consistency of risk measurements, that is, if $\mathrm{CVaR}_\alpha(Z_i) \leq 0$, for $i = 1,2$ and $Z := \begin{cases} Z_1, & \text{with probability}\, p, \\ Z_2, & \text{with probability}\, 1-p, \end{cases}$ for $p \in (0,1)$, then we do not necessarily have $\mathrm{CVaR}_\alpha(Z) \leq 0$; see [32, Example 3.4]. The other is that CVaR is not particularly sensitive to heavy tailed losses [15, Section 5]. Here, we illustrate this by a simple example. Let

$$X_1 := \begin{cases} 100, & p_1 = 98\% \\ -100, & p_2 = 1\% \\ -200, & p_3 = 1\% \end{cases}, X_2 := \begin{cases} 100, & p_1 = 98\% \\ -1, & p_2 = 1\% \\ -299, & p_3 = 1\% \end{cases}, X_3 := \begin{cases} 100, & p_1 = 98\% \\ 99, & p_2 = 1\% \\ -399, & p_3 = 1\% \end{cases}. \tag{1.1}$$

It is easy to calculate that $\mathrm{CVaR}_{0.02}(X_1) = \mathrm{CVaR}_{0.02}(X_2) = \mathrm{CVaR}_{0.02}(X_3) = 150$.

To overcome the deficiencies, a special category of convex risk measure, called *utility-based shortfall risk measure* (abbreviated as SR hereafter) was introduced by Föllmer and Schied [9] and attracted more and more attention in recent years; see [7, 15, 18]. Let $l : \mathbb{R} \to \mathbb{R}$ be a convex, increasing and non-constant function. Let λ be a pre-specified constant in the interior of the range of l to reveal the risk level. The SR of a financial position Z is defined as

$$(\mathrm{SR}) \quad \mathrm{SR}^P_{l,\lambda}(Z) := \inf\{t \in \mathbb{R} : t + Z \in \mathcal{A}_P\}, \tag{1.2}$$

where $\mathcal{A}_P := \{Z \in L^\infty : \mathbb{E}_P[l(-Z(\omega))] \leq \lambda\}$ is called the acceptance set and L^∞ denotes the set of bounded random variables. From the definition, we can see that the SR is the smallest amount of cash that must be added to the position Z to make it acceptable, i.e., $t + Z \in \mathcal{A}_P$. Observe that when $l(\cdot)$ takes a particular characteristic function of the form $\mathbb{1}_{(0,+\infty]}(\cdot)$, that is $\mathbb{1}_{(0,+\infty]}(z) = 1$ if $z \in (0,+\infty]$, otherwise 0, in this case $\mathrm{SR}^P_{l,\lambda}(Z)$ coincides with $\mathrm{VaR}_\lambda(Z)$. Of course, here l is nonconvex.

Compared to VaR and CVaR, the shortfall risk measure not only satisfies convexity and invariance under randomization, but also can be used more appropriately for dynamic measurement of risks over time. To see invariance under randomization, we note that SR defined as in (1.2) is a function on the space of random variables, it can also be represented as a function on the space of probability measures; see [32, Remark 2.1]. In the latter case, the acceptance set can be characterized by $\mathscr{N} := \{\mu \in \mathscr{P}(C) : \int l(-x)\mu(dx) \le \lambda\}$, where $\mathscr{P}(C)$ denotes the space of probability measures with support set being contained in a compact set $C \subset {\rm I\!R}$. If $\mu, \nu \in \mathscr{N}$, i.e., $\int l(-x)\mu(dx) \le \lambda, \int l(-x)\nu(dx) \le \lambda$, then for any $\alpha \in (0,1)$, $\int l(-x)(\alpha\mu + (1-\alpha)\nu)(dx) \le \lambda$, which means $\alpha\mu + (1-\alpha)\nu \in \mathscr{N}$. Moreover, the SR is found to be more sensitive to financial losses from extreme events with heavy tailed distributions; see [15, Section 5]. Indeed, if we set $l(z) = e^z$ and $\lambda = e$, then we can easily calculate the shortfall risk values of X_1, X_2 and X_3 in the previous example (1.1) with $\mathrm{SR}^P_{l,\lambda}(X_1) \approx 194, \mathrm{SR}^P_{l,\lambda}(X_2) \approx 293$, and $\mathrm{SR}^P_{l,\lambda}(X_3) \approx 393$. Furthermore, if we choose $l(z) = e^{\beta z}$ with $\beta > 0$, the resulting SR coincides, up to an additive constant, with the entropic risk measure, that is,

$$\mathrm{SR}^P_{l,\lambda}(Z) = \inf\{t \in {\rm I\!R} : \mathbb{E}_P[e^{-\beta(Z+t)}] \le \lambda\} = \frac{1}{\beta}\left(\log \mathbb{E}_P[e^{-\beta Z}] - \log\lambda\right).$$

In the case when $l(z) = z^\alpha \mathbb{1}_{[0,\infty)}(z)$ with $\alpha \ge 1$ the associated risk measure focuses on downside risk only and thus neglects the tradeoff between gains and losses.

Dunkel and Weber [7] are perhaps the first to discuss the computational aspects of SR. They characterized SR as a stochastic root finding problem and proposed the stochastic approximation (SA) method combined with importance sampling techniques to calculate it. Hu and Zhang [18] proposed an alternative approach by reformulating SR as the optimal value of a stochastic optimization problem and applying the well-known sample average approximation (SAA) method to solve the latter when either the true probability distribution is unknown or it is prohibitively expensive to compute the expected value of the underlying random functions. A detailed asymptotic analysis of the optimal values obtained from solving the sample average approximated problem was also provided.

In some practical applications, however, the true probability distribution may be unknown and it is expensive to collect a large set of samples or the samples are not trustworthy. However, it might be possible to use some partial information such as empirical data, computer simulation, prior moments or subjective judgements to construct a set of distributions which contains or approximates the true probability distribution in good faith. Under these circumstances, it might be reasonable to consider a distributionally robust version of (1.2) in order to hedge the risk arising from ambiguity of the true probability distribution,

$$\text{(DRSR)} \quad \mathrm{SR}^{\mathcal{P}}_{l,\lambda}(Z) := \inf\{t \in {\rm I\!R} : t + Z \in \mathcal{A}_{\mathcal{P}}\}, \tag{1.3}$$

where $\mathcal{A}_{\mathcal{P}} := \{Z \in L^\infty : \sup_{P\in\mathcal{P}} \mathbb{E}_P[l(-Z)] \le \lambda\}$, and $\mathcal{P}$ is a set of probability distributions. Föllmer and Schied seem to be the first to consider the notion of distributionally robust SR. In [10, Corollary 4.119], they established a robust representation theorem for DRSR. More recently, Wiesemann et al. [33] demonstrated how an DRSR optimization problem may be reformulated as a tractable convex programming problem when l is piecewise affine and the ambiguity set is constructed through some moment conditions, see [33, Example 6] for details.

In this paper, we take on the research by giving a more comprehensive treatment of DRSR. We start by looking into the properties of DRSR and then move on to discuss some optimization problems associated with DRSR. Specifically, for a loss $c(x,\xi)$ associated with decision vector $x \in X \subset {\rm I\!R}^n$ and random vector $\xi \in {\rm I\!R}^k$, we consider an optimization problem which aims to minimize the distributionally robust shortfall risk measure of the random loss:

$$\text{(DRSRP)} \quad \min_{x\in X} \; \mathrm{SR}^{\mathcal{P}}_{l,\lambda}(-c(x,\xi)), \tag{1.4}$$

where $\mathrm{SR}^{\mathcal{P}}_{l,\lambda}(\cdot)$ is defined as in (1.3). We present a detailed discussion on (DRSRP) including tractable reformulation for the problem when the ambiguity set has a specific structure. For the cases when the tractable reformulation is not possible, we propose discrete approximation schemes and quantify the approximation error of the ambiguity set and its propagation to the optimal value of (DRSRP). As far as we are concerned, the main contribution of the paper can be summarized as follows.

- We demonstrate that DRSR is the worst-case SR (Proposition 2.1) and using the property to show that it is a convex risk measure. In the case when $\mathrm{SR}^{P}_{l,\lambda}(\cdot)$ is a coherent risk measure for each $P \in \mathcal{P}$, we show that $\mathrm{SR}^{\mathcal{P}}_{l,\lambda}(\cdot)$ is also a coherent risk measure, see Remark 2.1.

- We investigate tractability of (DRSRP) by considering particular cases where the ambiguity set $\mathcal{P}$ is constructed through moment conditions, ϕ-divergence ball, Kantorovich ball and mixture distribution. For instance, when $\mathcal{P}$ is defined through some moment conditions, and $l(\cdot)$ is piecewise linear, we find that (DRSRP) can be reformulated as a linear program (3.13) with linear semi-definite constraint using standard approach in the literature of distributionally robust optimization. Similar observations are made when $\mathcal{P}$ is constructed through ϕ-divergence ball and Kantorovich ball.

- Since the structure of $\mathcal{P}$ often involves sample data, we analyse convergence of the ambiguity set as the sample size increases. For instance, when the nominal distribution in the ϕ-divergence ball is constructed through iid samples, we derive a quantitative convergence result about how the ambiguity set approximates the true probability distribution as the sample size increases, similar convergence result is established for the Kantorovich ball, see Propositions 3.1 and 3.3. These results not only give rise to appropriate error bounds for the ambiguity sets $\mathcal{P}_N$, but also present a confidence interval for the true probability distribution with a given set of samples. To quantify how the errors arising from the ambiguity set propagate to the optimal value of (DRSRP), we show under some moderate conditions that the error of the optimal value is linearly bounded by the error of the ambiguity set and subsequently derive a confidence interval for the optimal value of (DRSRP) for each discrete approximation scheme of the ambiguity sets (Theorem 4.2 and Corollary 4.1). Similar convergence results are established for mathematical programs with DRSR constraints (Theorem 4.3).

- Finally, as an application, we apply the (DRSRP) model to a portfolio management problem and carry out various tests on the numerical schemes for the (DRSRP) model with simulated data and real data. One of the important findings is that the (DRSRP) model outperforms the SAA model (with the ambiguity set $\mathcal{P}$ in (DRSRP) being replaced by the empirical probability distribution) in all of the tests (Section 5).

Throughout the paper, we use $\mathbb{R}^n$ to represent n dimensional Euclidean space, $\|x\|$ the Euclidean norm of a vector $x \in \mathbb{R}^n$ and $d(x, A) := \inf_{x' \in A} \|x - x'\|$ the distance from a point x to a set A. For two compact sets $A, B \subset \mathbb{R}^n$, we write $\mathbb{D}(A, B) := \sup_{x \in A} d(x, B)$ for the deviation of A from B and $\mathbb{H}(A, B) := \max\{\mathbb{D}(A, B), \mathbb{D}(B, A)\}$ for the Hausdorff distance between A and B. For two matrices $Y_1, Y_2 \in \mathbb{R}^{n \times n}$, $\langle Y_1, Y_2 \rangle := \mathrm{Trace}(Y_1^T Y_2)$, $Y_1 \preceq 0$ signifies the negative semi-definiteness of symmetric matrix Y_1. We use $\mathbb{B}$ to denote the unit ball in a matrix or vector space. Finally, for a sequence of subset $\{S_N\}$ in a metric space, denote by $\limsup_{N \to \infty} S_N$ its outer limit, that is $\limsup_{N \to \infty} S_N := \{x : \exists\, x_{N_k} \in S_{N_k} \text{such that}\, x_{N_k} \to x \,\text{as}\, k \to \infty\}$.

2 Properties of DRSR

In this section, we investigate the properties of DRSR. It is easy to observe that $\mathrm{SR}_{l,\lambda}^{\mathcal{P}}(Z)$ is the optimal value of the following minimization problem:

$$\begin{array}{ll} \min\limits_{t\in\mathbb{R}} & t \\ \text{s.t.} & \sup\limits_{P\in\mathcal{P}} \mathbb{E}_P[l(-Z-t)] \leq \lambda. \end{array} \tag{2.5}$$

The following proposition states that the DRSR is the worst-case SR and it preserves convexity of SR.

Proposition 2.1 *Let $\mathrm{SR}_{l,\lambda}^{\mathcal{P}}(Z)$ be defined as in (1.3), $Z \in L^\infty$ and $l : \mathbb{R} \to \mathbb{R}$ be a convex, increasing and non-constant function, let λ be a pre-specified constant in the range of l. Then $\mathrm{SR}_{l,\lambda}^{\mathcal{P}}(Z)$ is finite,*

$$\mathrm{SR}_{l,\lambda}^{\mathcal{P}}(Z) = \sup_{P\in\mathcal{P}} \mathrm{SR}_{l,\lambda}^{P}(Z), \tag{2.6}$$

and $\mathrm{SR}_{l,\lambda}^{\mathcal{P}}(Z)$ is a convex risk measure.

Proof. Since Z is bounded, then there exist constants α, β such that $Z(\omega) \in [\alpha, \beta]$ for all $\omega \in \Omega$. Thus

$$l(-\beta - t) \leq \sup_{P\in\mathcal{P}} \mathbb{E}_P[l(-Z-t)] \leq l(-\alpha - t), \forall t \in \mathbb{R}.$$

Since $l(-\beta - t) \to \infty$ as $t \to -\infty$ and $l(-\alpha - t) \leq \lambda$ for t sufficiently large, we conclude that the feasible set of the left hand side of (2.5) is bounded. To show equality (2.6), we note that

$$\begin{aligned} \hat{t} \quad &:= \quad \sup_{P\in\mathcal{P}} \mathrm{SR}_{l,\lambda}^{P}(Z) \leq \inf_t \sup_{P\in\mathcal{P}} \{t \in \mathbb{R} : \mathbb{E}_P[l(-Z-t)] \leq \lambda\} \\ &\leq \quad \inf_t \{t \in \mathbb{R} : \sup_{P\in\mathcal{P}} \mathbb{E}_P[l(-Z-t)] \leq \lambda\} = \mathrm{SR}_{l,\lambda}^{\mathcal{P}}(Z) =: t^*. \end{aligned}$$

To show the converse inequality, note that $\hat{t} \geq \mathrm{SR}_{l,\lambda}^{P}(Z), \forall P \in \mathcal{P}$. Thus

$$\mathbb{E}_P[l(-Z-\hat{t})] \leq \lambda, \forall P \in \mathcal{P},$$

which implies $\hat{t}$ is a feasible solution of (2.5) and hence $t^* \leq \hat{t}$. □

Remark 2.1 It may be helpful to make some comments on Proposition 2.1.

(i) The relationship established in (2.6) means that DRSR is the worst-case SR. This observation allows one to calculate DRSR via SR for each $P \in \mathcal{P}$ if it is easy to do so. Moreover, Giesecke et al. [15] showed that SR is a coherent risk measure if and only if the loss function l takes a specific form:

$$l(z) := \lambda - \alpha[z]_- + \beta[z]_+,\ \beta \geq \alpha \geq 0,$$

where $[z]_-$ denotes the negative part of z and $[z]_+$ denotes the positive part. Using this result, we can easily show through equation (2.6) that DRSR is a coherent risk measure when l takes the specific form in that the operation $\sup_{P\in\mathcal{P}}$ preserves the positive homogeneity and subadditivity.

(ii) The restriction of Z to L^∞ implies that the support [1] of the probability distribution of Z is bounded. This condition may be relaxed to the case when there exist $t_l, t_u \in \mathbb{R}$ such that $\sup_{P\in\mathcal{P}} \mathbb{E}_P[l(-Z-t_l)] > \lambda$ and $\sup_{P\in\mathcal{P}} \mathbb{E}_P[l(-Z-t_u)] < \lambda$; see [18].

We now move on to discuss the property of DRSR when it is applied to a random function. This is to pave a way for us to develop full investigation on (DRSRP) in Sections 3-4. To this end, we need to make some assumptions on the random function c and the loss function l.

Throughout this section, we use Ξ to denote the image space of random variable $\xi(\omega)$ and $\mathscr{P}(\Xi)$ to denote the set of all probability measures defined on the measurable space $(\Xi, \mathscr{B})$ with Borel sigma algebra $\mathscr{B}$. To ease notation, we will use ξ to denote either the random vector $\xi(\omega)$ or an element of $\mathbb{R}^k$ depending on the context.

Assumption 2.1 *Let X, $l(\cdot)$ and $c(\cdot,\cdot)$ be defined as in (DRSRP) (1.4). We assume the following.*

(a) X is a convex and compact set and Ξ is a compact set.

(b) l is convex, increasing, non-constant and Lipschitz continuous with modulus L.

(c) $c(\cdot,\xi)$ is finite valued, convex w.r.t. $x \in X$ for each $\xi \in \Xi$ and there exists a positive constant κ such that

$$|c(x,\xi) - c(x,\xi')| \le \kappa\|\xi - \xi'\|, \forall x \in X, \xi, \xi' \in \Xi.$$

The proposition below summarises some important properties of $l(c(x,\xi)-t)$ and $\sup\limits_{P\in\mathcal{P}} \mathbb{E}_P[l(c(x,\xi)-t)] - \lambda$ as a function of (x,t).

Proposition 2.2 *Let $g(x,t,\xi) := l(c(x,\xi)-t)$ and $v(x,t) := \sup\limits_{P\in\mathcal{P}} \mathbb{E}_P[g(x,t,\xi)] - \lambda$. The following assertions hold.*

(i) Under Assumption 2.1 (b) and (c), $g(\cdot,\cdot,\xi)$ is convex w.r.t. (x,t) for each fixed $\xi \in \Xi$, $g(x,t,\cdot)$ is uniformly Lipschitz continuous w.r.t. ξ with modulus $L\kappa$, and $v(x,t)$ is a convex function w.r.t. (x,t).

(ii) If, in addition, Assumption 2.1 (a) holds and λ is a pre-specified constant in the interior of the range of l, then there exist a point $(x_0,t_0) \in X \times \mathbb{R}$ and a constant $\eta > 0$ such that

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P[l(c(x_0,\xi)-t_0)] - \lambda < -\eta \tag{2.7}$$

and (DRSRP) *has a finite optimal value.*

Proof. Part (i). It is well known that the composition of a convex function by a monotonic increasing convex function preserves convexity. The remaining claims can also be easily verified.

[1] The support of the probability distribution is the smallest closed set such that a probability measure of its complement is zero.

Part (ii). Since $c(x, \xi)$ is finite valued and convex in x, it is continuous in x for each fixed ξ. Together with its uniform continuity in ξ, we are able to show that $c(x, \xi)$ is continuous over $X \times \Xi$. By the boundedness of X and Ξ, there is a positive constant α such that $c(x, \xi) \leq \alpha$ for all $(x, \xi) \in X \times \Xi$. With the boundedness of c and the monotonic increasing, convex and non-constant property of l, we can easily show Part (ii) analogous to the proof of the first part of Proposition 2.1. We omit the details. $\square$

3 Structure of (DRSRP') and approximation of the ambiguity set

In this section, we investigate the structure and numerical solvability of (DRSRP). Using the formulation (2.5) for DRSR, we can reformulate (DRSRP) as

$$\text{(DRSRP')} \quad \begin{array}{ll} \min\limits_{x \in X, t \in T} & t \\ \text{s.t.} & \sup\limits_{P \in \mathcal{P}} \mathbb{E}_P[l(c(x, \xi) - t)] \leq \lambda, \end{array} \tag{3.8}$$

where T is a compact set in $\mathbb{R}$ which contains t_0 defined as in (2.7) and its existence is ensured by Proposition 2.2 under some moderate conditions. Obviously, the structure of (DRSRP') is determined by the distributionally robust constraint. The latter relies heavily on the concrete structure of the ambiguity set $\mathcal{P}$ and the loss function l.

In the literature of distributionally robust optimization, various statistical methods have been proposed to build ambiguity sets based on available information of the underlying uncertainty, see for instance [33, 34] and the references therein. Here we consider four approaches which use moment information, ϕ-divergence ball, Kantorovich ball and mixture distribution because (DRSRP') is a special class of DRO. We discuss tractable formulation of (DRSRP') and its approximation in each case.

3.1 Ambiguity set based on moment conditions

Consider the case where the ambiguity set is defined by a general system of moment conditions:

$$\mathcal{P} := \{P \in \mathscr{P}(\Xi) : \mathbb{E}_P[\Psi(\xi, u)] = 0, \mathbb{E}_P[\Phi(\xi, u)] \preceq 0\}, \tag{3.9}$$

where $\Psi : \Xi \times \mathbb{R}^m \to \mathbb{R}^{n_1}$ is a vector-valued function and $\Phi : \Xi \times \mathbb{R}^m \to \mathbb{R}^{n_2 \times n_2}$ is a matrix-valued mapping, $u \in \mathbb{R}^m$ is a parameter, the mathematical expectations of Ψ and Φ are taken componentwise. In the literature of distributionally robust optimization, ambiguity set built on moment conditions can be traced back to earlier work by Scarf [28], who considered a distributionally robust model for a newsvendor problem. More recent works using moment conditions can be found in Delage and Ye [6], Wiesemann et al. [33] and Xu et al. [34].

Under some moderate conditions on the moment systems, we can easily derive Lagrangian dual formulation of $\sup_{P \in \mathcal{P}} \mathbb{E}_P[l(c(x, \xi) - t)]$ (see [34]) and consequently recast (DRSRP') as

$$\begin{array}{ll} \min\limits_{x,t,y,Y} & t \\ \text{s.t.} & l(c(x, \xi) - t) - y^T \Psi(\xi, u) - \langle Y, \Phi(\xi, u) \rangle \leq \lambda, \ \forall \xi \in \Xi, \\ & Y \succeq 0, \\ & x \in X, y \in \mathbb{R}^{n_1}, t \in T, \end{array} \tag{3.10}$$

which is a minimization problem with semi-infinite constraints. Under Assumption 2.1, the semi-infinite constraints are convex. Moreover, when l, c and $\mathcal{P}$ take specific forms, we can further reformulate (3.10) as a convex semi-definite program (SDP for short). For example, consider the case that $c(x,\xi) := -x^T\xi$, $\Xi = \mathbb{R}^k$ and

$$\mathcal{P}(\mu,\Sigma,\gamma_1,\gamma_2) := \left\{ P \in \mathscr{P}(\mathbb{R}^k) : \begin{array}{l} \mathbb{E}_P[\xi-\mu]^T\Sigma^{-1}\mathbb{E}_P[\xi-\mu] \le \gamma_1 \\ \mathbb{E}_P[(\xi-\mu)(\xi-\mu)^T] \preceq \gamma_2\Sigma \end{array} \right\}, \tag{3.11}$$

where μ and $\Sigma \succ 0$ are estimates of the mean and covariance of ξ and γ_1, γ_2 are parameters. This type of ambiguity set was first considered by Delage and Ye [6]. When the loss function l is piecewise affine and convex, that is,

$$l(z) := \max_{j=1,\ldots,K} a_j z + b_j, \tag{3.12}$$

Problem (3.10) can be reformulated as a convex SDP through the S-lemma [25]:

$$\begin{array}{cl} \min\limits_{x,t,\lambda_0,\zeta,Y_1,Y_2,y,\eta} & t \\ \text{s.t.} & \langle \Sigma, \gamma_2 Y_1 + Y_2\rangle + \lambda_0 - \mu^T Y_1 \mu - 2\mu^T y + \gamma_1\eta \le \lambda, \\ & \zeta + 2Y_1\mu + 2y = 0, \\ & \left[\begin{array}{cc} Y_1 & (\zeta + a_j x)/2 \\ (\zeta + a_j x)^T/2 & \lambda_0 + a_j t - b_j \end{array} \right] \succeq 0, j = 1,\ldots,K, \\ & Y_1 \succeq 0, \\ & \left[\begin{array}{cc} Y_2 & y \\ y^T & \eta \end{array} \right] \succeq 0, \\ & x \in X, t \in T. \end{array} \tag{3.13}$$

In the case when $l(\cdot)$ is a general increasing and convex function on $\mathbb{R}$, we may develop inner and outer piecewise affine convex approximations for the feasible set of problem (3.10). To see how it works, let $z_1 < z_2 < \ldots < z_K$ be K points in a compact set $O \subset \mathbb{R}$ where l is differentiable with $l'(z) = \frac{dl}{dz}$, let

$$\underline{a}_j := l'(z_j),\ \underline{b}_j := l(z_j) - l'(z_j)z_j \text{ for } j = 1,\ldots,K,$$

and $\underline{l}(z) := \max\limits_{j=1,\ldots,K} \underline{a}_j z + \underline{b}_j$. Let

$$\overline{a}_j := \frac{l(z_{j+1}) - l(z_j)}{z_{j+1} - z_j},\ \overline{b}_j := -\overline{a}_j z_j + l(z_j) \text{ for } j = 1,\ldots,K-1,$$

and $\overline{l}(z) := \max\limits_{j=1,\ldots,K-1} \overline{a}_j z + \overline{b}_j$. By the convexity of $l(\cdot)$, $\underline{l}(z) \le l(z) \le \overline{l}(z), \forall z \in O$. If we replace $l(\cdot)$ with $\underline{l}(\cdot)$ and $\overline{l}(\cdot)$ respectively in (3.10), then the resulting feasible set will give an outer and inner approximation of the feasible set of (3.10) and the approximate programs will give lower and upper bounds for the optimal value of (3.10). In either case, the approximate program can be solved through (3.13).

Remark 3.1 When μ and Σ in $\mathcal{P}(\mu,\Sigma,\gamma_1,\gamma_2)$ are estimated via iid samples, i.e.,

$$\mu_N := \frac{1}{N}\sum_{i=1}^N \xi^i \text{ and } \Sigma_N := \frac{1}{N}\sum_{i=1}^N (\xi^i - \mu_N)^T(\xi^i - \mu_N), \tag{3.14}$$

and γ_1 and γ_2 are some positive constants depending on the samples, written as γ_1^N and γ_2^N, we may ask ourselves as to how $\mathcal{P}(\mu_N, \Sigma_N, \gamma_1^N, \gamma_2^N)$ evolves as the size of data increases. Delage and Ye [6] proved that under the boundedness of support of ξ, the true probability distribution of ξ lies in set $\mathcal{P}(\mu_N, \Sigma_N, \gamma_1^N, \gamma_2^N)$ with probability at least $1-\delta$; see [6, Corollaries 3 and 4]. So [30] relaxed the bounded condition by replacing it with some moment growth condition, that is, there exists an absolute constant $s > 0$ such that for any $q \geq 1$, the following holds:

$$\mathbb{E}_P[\|\Sigma^{-1/2}(\xi - \mu)\|^q] \leq (sq)^{q/2}. \tag{3.15}$$

By specifying γ_1^N and γ_N^2 as follows:

$$\gamma_1^N := \frac{t_m^N}{1 - t_c^N - t_m^N}, \text{ and } \gamma_2^N := \frac{1 + t_m^N}{1 - t_c^N - t_m^N}, \tag{3.16}$$

where

$$t_m^N := \frac{4se^2 \ln^2(2/\delta)}{N}, \; t_c^N := \frac{4s'(2e/3)^{3/2} \ln^{3/2}(4k/\delta)}{\sqrt{N}}$$

with $\delta \in (0, 2e^{-3})$, $s' = \max\{s, 1\}$, So [30, Theorem 9] showed that the true probability distribution of ξ lies in $\mathcal{P}(\mu_N, \Sigma_N, \gamma_1^N, \gamma_2^N)$ with probability $1-\delta$ for N being sufficiently large.

Sun and Xu [31, Section 4] presented some quantitative convergence analysis of the ambiguity set $\mathcal{P}(\mu_N, \Sigma_N, \gamma_1^N, \gamma_2^N)$ and its impact on the underlying DRO problem as the sample size N increases. Similar analysis is given by Zhang et al. for DRO problems with general conic constrained moment conditions; see [37, Sections 2 and 3]. Note also that when tractable reformulation such as (3.13) is impossible, we may resort to discrete approximation scheme considered by Xu et al. [34, Section 3] for solving (DRSRP'). Quantitative convergence analysis of the ambiguity set and the approximate (DRSRP') may be established as in [19], we leave interested readers to explore as it is not the main focus here.

3.2 Ambiguity set constructed through ϕ-divergence

Let us now consider the case that the only available information about the random vector ξ is its empirical data and the size of such data is limited (not very large). In stochastic programming, a well-known approach in such situation is to use empirical distribution constructed through the data to approximate the true probability distribution. However, if the sample size is not big enough or there is a reason from computational point of view to use a small size of empirical data (e.g., in multistage decision-making problems), then the quality of such approximation may be compromised. ϕ-divergence is subsequently proposed to address this dilemma.

Let $p = (p_1, \ldots, p_M)^T \in \mathbb{R}_+^M$ and $q = (q_1, \ldots, q_M)^T \in \mathbb{R}_+^M$ be two probability vectors, that is, $\sum_{i=1}^M p_i = 1$ and $\sum_{i=1}^M q_i = 1$. The so-called ϕ-divergence between p and q is defined as

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

where $\phi(t)$ is a convex function for $t \geq 0$, $\phi(1) = 0$, $0\phi(a/0) := a \lim_{t\to\infty} \phi(t)/t$ for $a > 0$ and $0\phi(0/0) := 0$. In this subsection, we consider some common ϕ-divergences which are defined as follows.

(a) Kullback-Leibler: $I_{\phi_{KL}}(p,q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$ with $\phi_{KL}(t) = t\log t - t + 1$;

(b) Burg entropy: $I_{\phi_B}(p,q) = \sum_i q_i \log\left(\frac{q_i}{p_i}\right)$ with $\phi_B(t) = -\log t + t - 1$;

(c) J-divergence: $I_{\phi_J}(p,q) = \sum_i (p_i - q_i)\log\left(\frac{p_i}{q_i}\right)$ with $\phi_J(t) = (t-1)\log t$;

(d) χ^2-distance: $I_{\phi_{\chi^2}}(p,q) = \sum_i \frac{(p_i-q_i)^2}{p_i}$ with $\phi_{\chi^2}(t) = \frac{1}{t}(t-1)^2$;

(e) Modified χ^2-distance: $I_{\phi_{m\chi^2}}(p,q) = \sum_i \frac{(p_i-q_i)^2}{q_i}$ with $\phi_{m\chi^2}(t) = (t-1)^2$;

(f) Hellinger distance: $I_{\phi_H}(p,q) = \sum_i(\sqrt{p_i} - \sqrt{q_i})^2$ with $\phi_H(t) = (\sqrt{t}-1)^2$;

(g) Variation distance: $I_{\phi_V}(p,q) = \sum_i |p_i - q_i|$ with $\phi_V(t) = |t-1|$.

Lemma 3.1 *(Relationships between ϕ-divergences) For two probability vectors $p, q \in \mathbb{R}^M_+$, the following inequalities hold.*

(i) $I_{\phi_V}(p,q) \le \sqrt{2I_{\phi_{KL}}(p,q)}$;

(ii) $I_{\phi_V}(p,q) \le \sqrt{2I_{\phi_B}(p,q)}$;

(iii) $I_{\phi_V}(p,q) \le \sqrt{I_{\phi_J}(p,q)}$;

(iv) $I_{\phi_V}(p,q) \le \sqrt{I_{\phi_{\chi^2}}(p,q)}$;

(v) $I_{\phi_V}(p,q) \le \sqrt{I_{\phi_{m\chi^2}}(p,q)}$;

(vi) $I_{\phi_H}(p,q) \le I_{\phi_V}(p,q) \le 2\sqrt{I_{\phi_H}(p,q)}$.

Proof. Parts (i), (v) and (vi) follow straightforwardly from [14]. We prove the rest of the inequalities. Observe that $I_{\phi_B}(p,q) = I_{\phi_{KL}}(q,p)$ and $I_{\phi_V}(p,q) = I_{\phi_V}(q,p)$. By Part (i),

$$I_{\phi_V}(p,q) = I_{\phi_V}(q,p) \le \sqrt{2I_{\phi_{KL}}(q,p)} = \sqrt{2I_{\phi_B}(p,q)},$$

which gives rise to the inequality in Part (ii). Part (iii) follows from a combination of Parts (i) and (ii) in that $I_{\phi_J}(p,q) = I_{\phi_{KL}}(p,q) + I_{\phi_B}(p,q)$. Part (iv) holds because

$$I_{\phi_V}(p,q) = \sum_i \frac{|p_i - q_i|}{\sqrt{p_i}}\sqrt{p_i} \le \sqrt{I_{\phi_{\chi^2}}(p,q)},$$

where the inequality follows from the Cauchy-Schwarz inequality and the fact that p is a probability vector. □

Let $\{\zeta^1, \ldots, \zeta^M\} \subset \Xi$ denote the M-distinct points in the support of ξ and Ξ_i denote the Voronoi partition of Ξ centered at ζ^i for $i = 1, \ldots, M$. Let $\xi^1, \ldots, \xi^N$ be an iid sample of ξ where $N >> M$ and N_i denote the number of samples falling into area Ξ_i. Define empirical distribution

$$P_N(\cdot) := \sum_{i=1}^M \frac{N_i}{N}\mathbb{1}_{\zeta^i}(\cdot), \tag{3.17}$$

and ambiguity set

$$\mathcal{P}_N^M := \left\{ \sum_{i=1}^M p_i \mathbb{1}_{\zeta^i}(\cdot) : I_\phi(p, p_N) \leq r, \sum_{i=1}^M p_i = 1, p_i \geq 0, \forall i = 1, \ldots, M \right\}, \tag{3.18}$$

where $p_N = \left(\frac{N_1}{N}, \ldots, \frac{N_M}{N}\right)^T$. Using $\mathcal{P}_N^M$ for the ambiguity set in (DRSRP'), we can derive a dual formulation of (DRSRP') as follows:

$$\begin{array}{cl} \min\limits_{x\in X, t\in T, \tau, u} & t \\ \text{s.t.} & \tau + ru + u \sum\limits_{i=1}^M [p_N]_i \phi^*(s_i) \leq \lambda, \\ & s_i \leq \lim\limits_{t\to\infty} \dfrac{\phi(t)}{t}, \ i = 1, \ldots, M, \\ & s_i = [l(c(x, \zeta^i) - t) - \tau]/u, \ i = 1, \ldots, M, \\ & u \geq 0, \end{array} \tag{3.19}$$

where p_N is defined as in (3.18) and we write $[p_N]_i$ for the i-th component of p_N, ϕ^* denotes the Fenchel conjugate of ϕ, i.e., $\phi^*(s) = \sup_{t\geq 0}\{st - \phi(t)\}$, see similar formulation in [3]. Note that $u \sum_{i=1}^M \phi^*(l(c(x, \zeta^i) - t) - \tau)/u]$ is a convex function of x, u, τ and t, see [20]. Thus, problem (3.19) is a convex program.

It is important to note that the tractable reformulation (3.19) relies heavily on the discrete structure of the nominal distribution. Note that it is possible to use a continuous distribution for the nominal distribution, in which case the summation in the first constraint of problem (3.19) will become $\mathbb{E}[\phi^*((l(c(x, \zeta) - t) - \tau))/u)]$ (before introducing new variables s_i). In such a case, we will need to use sample average approximation approach to deal with the expected value.

Let $\mathscr{L}$ denote the space of all Lipschitz continuous functions $h : \Xi \to \mathbb{R}$ with modulus being bounded by 1 and $P, Q \in \mathscr{P}(\Xi)$ be two probability measures. Recall that the Kantorovich metric (or distance) between P and Q, denoted by $\mathsf{dl}_K(P, Q)$, is defined by

$$\mathsf{dl}_K(P, Q) := \sup_{h\in\mathscr{L}} \left\{ \int_\Xi h(\xi)P(d\xi) - \int_\Xi h(\xi)Q(d\xi) \right\}.$$

Using the Kantorovich metric, we can define deviation of a set of probability measures $\mathcal{P}$ from another set of probability measures $\mathcal{Q}$

$$\mathbb{D}_K(\mathcal{P}, \mathcal{Q}) := \sup_{P\in\mathcal{P}} \inf_{Q\in\mathcal{Q}} \mathsf{dl}_K(P, Q),$$

and the Hausdorff distance between the two sets

$$\mathbb{H}_K(\mathcal{P}, \mathcal{Q}) := \max\left\{\mathbb{D}_K(\mathcal{P}, \mathcal{Q}), \mathbb{D}_K(\mathcal{Q}, \mathcal{P})\right\}. \tag{3.20}$$

An important property of the Kantorovich metric is that it metrizes weak convergence of probability measures [4] when the support is bounded, that is, a sequence of probability measures $\{P_N\}$ converges to P weakly if and only if $\mathsf{dl}_K(P_N, P) \to 0$ as N tends to infinity.

Recall that for a given set of points $\{\zeta^1, \ldots, \zeta^M\}$, the Voronoi partition of Ξ is defined as M subsets of Ξ, denoted by $\Xi_1, \ldots, \Xi_M$, with $\bigcup_{i=1,\ldots,M} \Xi_i = \Xi$ and

$$\Xi_i \subseteq \left\{ y : \|y - \zeta^i\| = \min_{j=1,\ldots,M} \|y - \zeta^j\| \right\}.$$

By [24, Lemma 4.9],

$$\begin{aligned} \mathsf{dl}_K\left(\sum_{i=1}^M P^*(\Xi_i)\mathbb{1}_{\zeta^i}(\cdot), P^*\right) &= \int \min_{1\le i\le M} d(\xi, \zeta^i) dP^* \qquad (3.21)\\ &= \sum_{i=1}^M \int_{\Xi_i} d(\xi,\zeta^i) dP^* \le \beta_M, \end{aligned}$$

where

$$\beta_M := \max_{\xi\in\Xi} \min_{1\le i\le M} d(\xi,\zeta^i). \qquad (3.22)$$

Using this, we can estimate the Kantorovich distance between $\mathcal{P}_N^M$ and the true probability distribution P^*.

Proposition 3.1 *Let $\mathcal{P}_N^M$ be defined as in (3.18) and P^* be the true probability distribution of ξ. Let β_M be defined as in (3.22) and δ be a positive number such that $M\delta < 1$. If ϕ is chosen from one of the functions listed in (a)-(g) preceding Lemma 3.1, then with probability at least $1 - M\delta$,*

$$\mathbb{H}_K(\mathcal{P}_N^M, P^*) \le \beta_M + \frac{D}{2}\max\{2\sqrt{r}, r\} + \frac{D}{2}\Delta(M, N, \delta), \qquad (3.23)$$

where $\Delta(M,N,\delta) := \min\left(\frac{M}{\sqrt{N}}\left(2+\sqrt{2\ln\frac{1}{\delta}}\right), 4 + \frac{1}{\sqrt{N}}\left(2+\sqrt{2\ln\frac{1}{\delta}}\right)\right)$, D is the diameter of Ξ, that is, $\sup\{\|\xi' - \xi''\| : \xi', \xi'' \in \Xi\}$, and r is defined in (3.18). In the case when ξ follows a discrete distribution with support set $\{\zeta^1, \ldots, \zeta^M\}$, we have

$$\mathbb{H}_K(\mathcal{P}_N^M, P^*) \le \frac{D}{2}\max\{2\sqrt{r}, r\} + \frac{D}{2}\Delta(M,N,\delta) \qquad (3.24)$$

with probability at least $1 - M\delta$, where $D = \sup\{\|\zeta^i - \zeta^j\|, 1 \le i \ne j \le M\}$.

Proof. By the triangle inequality of the Hausdorff distance with the Kantorovich metric,

$$\mathbb{H}_K(\mathcal{P}_N^M, P^*) \le \sup_{P\in\mathcal{P}_N^M} \mathsf{dl}_K\left(P, \sum_{i=1}^M P^*(\Xi_i)\mathbb{1}_{\zeta^i}(\cdot)\right) + \mathsf{dl}_K\left(\sum_{i=1}^M P^*(\Xi_i)\mathbb{1}_{\zeta^i}(\cdot), P^*\right).$$

By (3.21), $\mathsf{dl}_K\left(\sum_{i=1}^M P^*(\Xi_i)\mathbb{1}_{\zeta^i}(\cdot), P^*\right) \le \beta_M$. Moreover, it follows by [14, Theorem 4], the Kantorovich distance is bounded by $D/2$ times the total variation distance, that is,

$$\mathsf{dl}_K\left(P, \sum_{i=1}^M P^*(\Xi_i)\mathbb{1}_{\zeta^i}(\cdot)\right) \le \frac{D}{2}\sum_{i=1}^M |p_i - P^*(\Xi_i)|.$$

Observe that

$$\begin{aligned}\sum_{i=1}^M |p_i - P^*(\Xi_i)| &\leq \sum_{i=1}^M (|p_i - [p_N]_i| + |[p_N]_i - P^*(\Xi_i)|)\\ &= I_{\phi_V}(p, p_N) + \sum_{i=1}^M |[p_N]_i - P^*(\Xi_i)|.\end{aligned}$$

By Lemma 3.1,

$$I_{\phi_V}(P_N^M, p_N) \leq \max\left\{2\sqrt{I_\phi(P_N^M, p_N)}, I_{\phi_V}(P_N^M, p_N)\right\} \leq \max\{2\sqrt{r}, r\}.$$

Thus, in order to show (3.23), it suffices to show

$$\sum_{i=1}^M |[p_N]_i - P^*(\Xi_i)| \leq \Delta(M, N, \delta). \tag{3.25}$$

Let $a \in \mathbb{R}^M$ be a vector with $\|a\|_\infty = \max_{1\leq i\leq M} |a_i| = 1$, and $\phi_a(\xi) := \sum_{i=1}^M a_i \mathbb{1}_{\Xi_i}(\xi)$. Then $\sup_{\xi\in\Xi} |\phi_a(\xi)| \leq 1$ and it follows by [29, Theorem 3] that

$$\left|\frac{1}{N}\sum_{k=1}^N \phi_a(\xi^k) - \mathbb{E}_{P^*}[\phi_a(\xi)]\right| \leq \frac{1}{\sqrt{N}}\left(2 + \sqrt{2\ln\frac{1}{\delta}}\right) \tag{3.26}$$

with probability at least $1-\delta$ for the fixed a. In particular, if we set $a = e_i$, for $i = 1, \cdots, M$, where $e_i \in \mathbb{R}^M$ is a vector with i-th component being 1 and the rest being 0, then we obtain

$$|[p_N]_i - P^*(\Xi_i)| = \left|\frac{1}{N}\sum_{k=1}^N \phi_{e_i}(\xi^k) - \mathbb{E}_{P^*}[\phi_{e_i}(\xi)]\right| \leq \frac{1}{\sqrt{N}}\left(2 + \sqrt{2\ln\frac{1}{\delta}}\right) \tag{3.27}$$

with probability at least $1-\delta$ for each $i = 1, \cdots, M$. This gives

$$\sum_{i=1}^M |[p_N]_i - P^*(\Xi_i)| \leq \frac{M}{\sqrt{N}}\left(2 + \sqrt{2\ln\frac{1}{\delta}}\right) \tag{3.28}$$

with probability at least $1 - M\delta$ and hence we have shown (3.25) for the first part of its bound in $\Delta(M, N, \delta)$.

To show the second part of the bound, we need a bit more complex argument to estimate the left hand side of (3.25). Let $A := \{a \in \mathbb{R}^M : \|a\|_\infty = 1\}$. Let ν be a small positive number (less or equal to 2) and $A_k := \{a^1, \cdots, a^k\}$ be a ν-net of A, that is, for any $a \in A$, there is a point $a_i(a) \in A_k$ depending on a such that $\|a - a^i(a)\|_\infty \leq \nu$. Observe that

$$\sum_{i=1}^M |[p_N]_i - P^*(\Xi_i)| = \sup_{\|a\|_\infty=1} |p_N^T a - {p^*}^T a|, \tag{3.29}$$

where we write p^* for the M-dimensional vector with i component $P^*(\Xi_i)$. Then

$$\begin{aligned}|p_N^T a - {p^*}^T a| &\leq |p_N^T(a - a^i(a))| + |p_N^T a^i(a) - {p^*}^T a^i(a)| + |{p^*}^T a^i(a) - {p^*}^T a|\\ &\leq 2\nu + |p_N^T a^i(a) - {p^*}^T a^i(a)|.\end{aligned}$$

By (3.26), for each a^i, $i = 1, \cdots, k$

$$|p_N^T a^i - p^{*T} a^i| \le \frac{1}{\sqrt{N}} \left(2 + \sqrt{2 \ln \frac{1}{\delta}}\right) \tag{3.30}$$

with probability at least $1 - \delta$, thus inequality (3.30) holds uniformly for all $i = 1, \cdots, k$ with probability at least $1 - k\delta$. This enables us to conclude that

$$\begin{aligned} \sum_{i=1}^{M} |[p_N]_i - P^*(\Xi_i)| &= \sup_{\|a\|_\infty = 1} |p_N^T a - p^{*T} a| \\ &\le 2\nu + \frac{1}{\sqrt{N}} \left(2 + \sqrt{2 \ln \frac{1}{\delta}}\right) \end{aligned} \tag{3.31}$$

with probability at least $1 - k\delta$. Since when $\nu = 2$, A_k will be a trivial ν-net, then we can set $k = M$ and obtain from (3.31)

$$\sum_{i=1}^{M} |[p_N]_i - P^*(\Xi_i)| \le 4 + \frac{1}{\sqrt{N}} \left(2 + \sqrt{2 \ln \frac{1}{\delta}}\right) \tag{3.32}$$

with probability at least $1 - M\delta$. This completes the proof of (3.25) and hence inequality (3.23).

In the case when ξ follows a discrete distribution with support set $\{\zeta^1, \ldots, \zeta^M\}$,

$$\mathbb{H}_K(\mathcal{P}_N^M, P^*) \le \sup_{P \in \mathcal{P}_N^M} \frac{D}{2} \left(I_{\phi_V}(P, p_N) + \sum_{i=1}^{M} |[p_N]_i - P^*(\Xi_i)|\right).$$

The rest follows from similar analysis for the proof of (3.23). □

It might be helpful to make a few comments on the above technical results. First, if we set $\delta = \frac{1}{10M}$, then $1 - \delta M = 90\%$ and the third term at the right hand side of (3.23) is

$$\frac{D}{2} \min\left(\frac{M}{\sqrt{N}} \left(2 + \sqrt{2 \ln(10M)}\right), 4 + \frac{1}{\sqrt{N}} \left(2 + \sqrt{2 \ln(10M)}\right)\right). \tag{3.33}$$

In order for the first part of (3.33) to be small, N must be significantly larger than M. The approach works for the case when there is a large data set which is not scattered evenly over Ξ, but rather they form clumps, locally dense areas, modes, or clusters. In the case that N is less than $(M-1)^2$, the second part of (3.33) is smaller than the first part, which means the second part provides a lower bound. Second, the true distribution in the local areas may be further described by moment conditions, see [21, 33]. Third, Pflug and Pichler proposed a practical way for identifying the optimal location of discrete points $\zeta^1, \ldots, \zeta^M$ and computing the probability of each Voronoi partition, see [24, Algorithms 4.1-4.5]. Forth, the inequality (3.23) gives a bound for the Hausdorff distance of the true probability distribution P^* and the ambiguity set $\mathcal{P}_N^M$, it does not indicate the true probability distribution P^* being located in $\mathcal{P}_N^M$.

Since the ambiguity set $\mathcal{P}_N^M$ does not constitute any continuous distribution irrespective of $r > 0$, then when the true probability distribution P^* is continuous, P^* lies outside $\mathcal{P}_N^M$ with probability 1. If the true probability distribution P^* is discrete, Pardo [22] showed that the estimated ϕ-divergence $\frac{2N}{\phi''(1)} I_\phi(p^*, p_N)$ asymptotically follows a χ^2_{M-1}-distribution with $M - 1$

degree of freedom, where p^* denotes the probability vector corresponding to probability measure P^* and M is the cardinality of Ξ (the support of P^*), which means if we set

$$r := \frac{\phi''(1)}{2N}\chi^2_{M-1,1-\delta}, \tag{3.34}$$

then with probability $1-\delta$, $I_\phi(p^*, p_N) \leq r$. The latter indicates that the ambiguity set (3.18) lies in the $1-\delta$ confidence region.

For general ϕ-divergences, we are unable to establish the quantitative convergence as in Proposition 3.1. However, if P^* follows a discrete distribution with support $\{\zeta^1, \ldots, \zeta^M\}$, the following qualitative convergence result holds.

Proposition 3.2 *[20, Proposition 2] Suppose that $\phi(t) \geq 0$ has a unique root at $t = 1$ and the samples are independent and identically distributed from the true distribution P^*. Then*

$$\mathbb{H}_K(\mathcal{P}_N^M, P^*) \to 0, \text{w.p.1},$$

as $N \to \infty$, where r is defined as in (3.34).

Proof. The proof is analogous to [20, Proposition 2], we provide it for completeness. First, we claim that

$$\sup_{p\in\mathcal{P}_N^M} \|p - p^*\|_\infty \to 0.$$

Observe that by the strong law of large numbers, $[p_N]_i = \frac{N_i}{N}$ converges to p^* uniformly, i.e.,

$$\|p_N - p^*\|_\infty = \sup_{i=1,\ldots,M} |[p_N]_i - p_i^*| \to 0 \,\text{as}\, N \to \infty,$$

and for any $p \in \mathcal{P}_N^M$,

$$\|p - p^*\|_\infty \leq \|p - p_N\|_\infty + \|p_N - p^*\|_\infty,$$

we only need to prove that $\sup_{p\in\mathcal{P}_N^M} \|p - p_N\|_\infty \to 0$. We will prove it by contradiction.

Let $\epsilon > 0$ be such that $\min_i p_i^* > \frac{\epsilon}{2}$ and N_1 such that $\|p_N - p^*\|_\infty \leq \frac{\epsilon}{2}$ for $N \geq N_1$, this means $[p_N]_i > 0$ for $i = 1, \ldots, M$. Suppose that there exists $p \in \mathcal{P}_N^M$ such that $\|p - p_N\|_\infty > \frac{\epsilon}{2}$ for $N > N_1$, that is, there exists i_0 such that $|p_{i_0} - [p_N]_{i_0}| > \frac{\epsilon}{2}$, i.e. $p_{i_0} > [p_N]_{i_0} + \frac{\epsilon}{2}$ or $p_{i_0} < [p_N]_{i_0} - \frac{\epsilon}{2}$. In either case, since $\phi(t) \geq 0$ is a convex function with only one root at $t = 1$, we have

$$\begin{aligned}
\phi\left(\frac{p_{i_0}}{[p_N]_{i_0}}\right) &\geq \min\left\{\phi\left(\frac{[p_N]_{i_0} + \frac{\epsilon}{2}}{[p_N]_{i_0}}\right), \phi\left(\frac{[p_N]_{i_0} - \frac{\epsilon}{2}}{[p_N]_{i_0}}\right)\right\} \\
&\geq \min\left\{\phi\left(1 + \frac{\epsilon}{2}\right), \phi\left(1 - \frac{\epsilon}{2}\right)\right\},
\end{aligned}$$

where the last inequality follows from the fact that

$$\frac{a + \frac{\epsilon}{2}}{a} \geq 1 + \frac{\epsilon}{2} \quad \text{and} \quad \frac{a - \frac{\epsilon}{2}}{a} \leq 1 - \frac{\epsilon}{2}$$

for $0 \le a \le 1$ and the property of ϕ. Thus

$$\begin{aligned} I_\phi(p, p_N) &= \sum_{i=1}^M [p_N]_i \phi\left(\frac{p_i}{[p_N]_i}\right) \ge (\min_i [p_N]_i) \phi\left(\frac{p_{i_0}}{[p_N]_{i_0}}\right) \\ &\ge \min_i \left\{p_i^* - \frac{\epsilon}{2}\right\} \min\left\{\phi\left(1+\frac{\epsilon}{2}\right), \phi\left(1-\frac{\epsilon}{2}\right)\right\}. \end{aligned} \tag{3.35}$$

The right hand side of (3.35) is positive because ϕ has a unique root at $t = 1$. By choosing $N_2 > N_1$ such that

$$\min_i \left\{p_i^* - \frac{\epsilon}{2}\right\} \min\left\{\phi\left(1+\frac{\epsilon}{2}\right), \phi\left(1-\frac{\epsilon}{2}\right)\right\} \ge \frac{r_0}{N_2},$$

with $r_0 = \frac{\phi''(1)}{2}\chi^2_{M-1,1-\delta}$, we now obtain that $\|p - p_N\|_\infty > \frac{\epsilon}{2}$ implies $I_\phi(p, p_N) > \frac{r_0}{N}$ for all $N > N_2$, which is a contradiction with $p \in \mathcal{P}_N^M$. Hence we have for any $\epsilon > 0$, there exists $N_2 > 0$ such that $\|p - p_N\|_\infty \le \frac{\epsilon}{2}$ holds for any $p \in \mathcal{P}_N^M$ and $N > N_2$.

Note that

$$\mathbb{H}_K(\mathcal{P}_N^M, P^*) = \sup_{p \in \mathcal{P}_N^M} \mathsf{dl}_K(p, p^*) \le \sup_{p \in \mathcal{P}_N^M} D' M \|p - p_N\|_\infty,$$

where D' is the diameter of discrete support set of P^*, the conclusion follows. □

3.3 Kantorovich ball

An alternative approach to the ϕ-divergence is to consider Kantorovich ball centered at a nominal distribution, that is,

$$\mathcal{P}_N = \{P \in \mathscr{P}(\Xi) : \mathsf{dl}_K(P, P_N) \le r\}, \tag{3.36}$$

where $P_N(\cdot) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\xi^i}(\cdot)$, $\xi^1, \ldots, \xi^N$ are iid samples of ξ. Differing from the ambiguity set based on ϕ-divergence, the Kantorovich ball contains both discrete and continuous distributions. In particular, if there exists a positive number $a > 0$ such that

$$A := \int_\Xi \exp(\|\xi\|^a) P^*(d\xi) < \infty, \tag{3.37}$$

then for any $r > 0$, there exist positive constants C_1 and C_2 such that

$$\text{Prob}(\mathsf{dl}_K(P^*, P_N) \ge r) \le \begin{cases} C_1 \exp(-C_2 N r^{\max\{k,2\}}) & \text{if } r \le 1, \\ \exp(-C_2 N r^a) & \text{if } r > 1, \end{cases} \tag{3.38}$$

for all $N \ge 1$, $k \ne 2$, where C_1 and C_2 are positive constants only depending on a, A and k, "Prob" is a probability distribution over space $\Xi \times \cdots \times \Xi$ (N times) with Borel-sigma algebra $\mathscr{B} \otimes \cdots \otimes \mathscr{B}$, and k is the dimension of ξ; see [11] for details. By setting the right hand side of the above inequality to δ and solving for r, we may set

$$r_N(\delta) := \begin{cases} \left(\frac{\log(C_1\delta^{-1})}{C_2 N}\right)^{1/\max\{k,2\}} & \text{if } N \ge \frac{\log(C_1\delta^{-1})}{C_2}, \\ \left(\frac{\log(C_1\delta^{-1})}{C_2 N}\right)^{1/a} & \text{if } N < \frac{\log(C_1\delta^{-1})}{C_2}, \end{cases} \tag{3.39}$$

and consequently the ambiguity set (3.36) contains the true probability distribution P^* with probability $1-\delta$ when $r = r_N(\delta)$.

In [8, 13, 35], the dual formulation of distributionally robust optimization problem with the ambiguity set (3.36) has been established. Based on these results, the dual of DRSRP can be written as

$$\begin{array}{ll} \min\limits_{x\in X, t\in T, \eta, s} & t \\ \text{s.t.} & \eta r + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \le \lambda, \\ & \sup\limits_{\xi\in\Xi} \left[l(c(x,\xi)-t) - \eta\|\xi-\xi^i\| \right] \le s_i, i = 1,\ldots,N. \end{array} \tag{3.40}$$

In the case when $c(x,\xi) = -x^T\xi$, $\Xi = \{\xi \in \mathbb{R}^k : G\xi \le d\}$ and

$$l(c(x,\xi)-t) = \max_{j=1,\ldots,K} a_j(-x^T\xi - t) + b_j = \max_{j=1,\ldots,K} \langle -a_j x, \xi\rangle - a_j t + b_j,$$

problem (3.40) can be recast as

$$\begin{array}{ll} \min\limits_{x\in X, t\in T, \eta, s, \gamma_{ij}} & t \\ \text{s.t.} & \eta r_N + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \le \lambda, \\ & b_j - a_j t - \langle a_j x, \xi_i\rangle + \langle \gamma_{ij}, d - A\xi^i\rangle \le s_i,\ i = 1,\ldots,N, j = 1,\ldots,K, \\ & \|G^T\gamma_{ij} + a_j x\| \le \eta,\ i = 1,\ldots,N, j = 1,\ldots,K, \\ & \gamma_{ij} \ge 0,\ i = 1,\ldots,N, j = 1,\ldots,K. \end{array}$$

The proposition below gives a bound for the Hausdorff distance of $\mathcal{P}_N$ and P^* under the Kantorovich metric.

Proposition 3.3 *Let $\mathcal{P}_N$ be defined as in (3.36) and P^* denote the true probability distribution. Let $r_N(\delta)$ be defined as in (3.39). If the radius of the Kantorovich ball in (3.36) is equal to $r_N(\delta)$, then with probability at least $1-\delta$,*

$$\mathbb{H}_K(\mathcal{P}_N, P^*) \le 2r_N(\delta). \tag{3.41}$$

Proof. We first prove that

$$\mathbb{H}_K(\mathcal{P}_N, P^*) \le \mathsf{dl}_K(P_N, P^*) + r. \tag{3.42}$$

To see this, for any $P' \in \mathcal{P}_N$, we have

$$\mathsf{dl}_K(P', P^*) \le \mathsf{dl}_K(P', P_N) + \mathsf{dl}_K(P_N, P^*) \le r + \mathsf{dl}_K(P_N, P^*),$$

which implies

$$\mathbb{D}_K(\mathcal{P}_N, P^*) \le r + \mathsf{dl}_K(P_N, P^*).$$

On the other hand,

$$\mathbb{D}_K(P^*, \mathcal{P}_N) = \inf_{Q\in\mathcal{P}_N} \mathsf{dl}_K(P^*, Q) \le \mathsf{dl}_K(P^*, P') \le r + \mathsf{dl}_K(P_N, P^*).$$

A combination of the last two equations yields (3.42).

Let us now estimate the first term in (3.42), i.e., $\mathsf{dl}_K(P_N, P^*)$. By the definition of $r_N(\delta)$, we have with probability $1-\delta$, $\mathsf{dl}_K(P_N, P^*) \le r_N(\delta)$. The conclusion follows. □

In the case when the centre of the Kantorovich ball P_N in (3.36) is replaced by that defined as in (3.17), we have

$$\mathbb{H}_K(\mathcal{P}_N, P^*) \le \beta_M + r + \Delta(M, N, \delta) \tag{3.43}$$

with probability at least $1 - M\delta$, where $\Delta(M, N, \delta)$ is defined as in Proposition 3.1. To see this, we can use the triangle inequality of the Hausdorff distance with the Kantorovich metric to derive

$$\begin{aligned} & \mathbb{H}_K(\mathcal{P}_N, P^*) \\ \le\ & \sup_{P\in\mathcal{P}_N} \mathsf{dl}_K\left(P, \sum_{j=1}^M P^*(\Xi_j)\mathbb{1}_{\zeta^j}(\cdot)\right) + \mathsf{dl}_K\left(\sum_{j=1}^M P^*(\Xi_j)\mathbb{1}_{\zeta^j}(\cdot), P^*\right). \end{aligned} \tag{3.44}$$

Since

$$\begin{aligned} \mathsf{dl}_K\left(P, \sum_{j=1}^M P^*(\Xi_j)\mathbb{1}_{\zeta^j}(\cdot)\right) &\le \mathsf{dl}_K\left(P, P_N\right) + \mathsf{dl}_K\left(P_N, \sum_{j=1}^M P^*(\Xi_j)\mathbb{1}_{\zeta^j}(\cdot)\right) \\ &\le r + \frac{D}{2}\sum_{j=1}^M \left|[p_N]_j - P^*(\Xi_j)\right|, \end{aligned} \tag{3.45}$$

we establish (3.43) by combining (3.44), (3.45), (3.28) and (3.31).

3.4 Mixture distribution

We now turn to discuss the case when the true distribution P^* of ξ lies in a convex combination of some known distributions P_j, $j = 1, \ldots, J$. In other words, P^* can be represented as a mixture distribution of P_j although we don't know the representation. Consequently, we may construct the ambiguity set as follows:

$$\mathcal{P} := \left\{ \sum_{j=1}^J \alpha_j P_j : \sum_{j=1}^J \alpha_j = 1, \alpha_j \ge 0, j = 1, \ldots, J \right\}. \tag{3.46}$$

Distributionally robust optimization under mixture distribution can be traced back to Hall et al. [16] and Peel and McLachlan [23]. More recently, Zhu and Fukushima [36] studied robust optimization of the CVaR of a random function under mixture distribution.

The (DRSRP') with $\mathcal{P}$ being defined as in (3.46) can be written as

$$\begin{aligned} \min_{x\in X, t\in T} \quad & t \\ \text{s.t.} \quad & \mathbb{E}_{P_j}[l(c(x,\xi) - t)] \le \lambda, j = 1, \ldots, J. \end{aligned} \tag{3.47}$$

Note that in order to solve problem (3.47), it might be desirable to use sample average approximation to avoid computation of the expected values w.r.t. probability distributions P_j. Let

$P_j^N(\cdot) = \frac{1}{N}\sum_{k=1}^N \mathbb{1}_{\xi_j^k}(\cdot)$ be an empirical distribution of P_j for $j = 1, \ldots, J$. We consider

$$\begin{array}{ll} \min\limits_{x\in X, t\in T} & t \\ \text{s.t.} & \mathbb{E}_{P_j^N}[l(c(x,\xi) - t)] \leq \lambda, j = 1, \ldots, J \end{array} \tag{3.48}$$

to approximate problem (3.47). By (3.38), we know P_j^N converges to P_j under the Kantorovich metric at exponential rate w.r.t. increase of sample size N under some condition. If we write $\hat{\mathcal{P}}$ for $\{P_j, j = 1, \ldots, J\}$ and $\hat{\mathcal{P}}_N$ for $\{P_j^N, j = 1, \ldots, J\}$, then we can easily establish the convergence of $\hat{\mathcal{P}}_N$ to $\hat{\mathcal{P}}$ under the Kantorovich metric.

Proposition 3.4 *Assume that* $\mathsf{dl}_K(P_j^N, P_j) \to 0$ *for* $j = 1, \ldots, J$ *as* $N \to \infty$. *For* N *sufficiently large*

$$\mathbb{H}_K(\hat{\mathcal{P}}_N, \hat{\mathcal{P}}) = \max\left\{\mathsf{dl}_K(P_j^N, P_j) : j = 1, \ldots, J\right\} \tag{3.49}$$

and

$$\mathbb{H}_K(\hat{\mathcal{P}}_N, \hat{\mathcal{P}}) \leq \max_{j=1,\ldots,J} r_N^j(\delta) \tag{3.50}$$

probability $1-\delta$, *where* $r_N^j(\delta)$ *is defined as in (3.39) for* P_j, $j = 1, \ldots, J$.

Proof. The proof is similar to [31, Proposition 6] where the discrepancy of two ambiguity sets defined via mixture distribution is quantified by the total variation metric. Denote by α the minimal Kantorovich distance between each pair of probability distributions in $\hat{\mathcal{P}}$, i.e., $\alpha := \min\{\mathsf{dl}_K(P_i, P_j) : i, j = 1, \ldots, J, i \neq j\}$. Since $\mathsf{dl}_K(P_j^N, P_j) \to 0$ for $j = 1, \ldots, J$ as $N \to \infty$, there exists N_0 such that for $N \geq N_0$,

$$\mathsf{dl}_K(P_j^N, P_j) \leq \frac{\alpha}{8}, \forall j = 1, \ldots, J.$$

Observe that for any $i \neq j$,

$$\mathsf{dl}_K(P_j^N, P_i) \geq \mathsf{dl}_K(P_i, P_j) - \mathsf{dl}_K(P_j^N, P_j) \geq \frac{7}{8}\alpha.$$

Hence,

$$\mathsf{dl}_K(P_j^N, \hat{\mathcal{P}}) = \min_{i\in\{1,\ldots,J\}}\{\mathsf{dl}_K(P_j^N, P_i)\} = \mathsf{dl}_K(P_j^N, P_j).$$

By the definition of $\mathbb{D}_K$, the inequality gives rise to

$$\mathbb{D}_K(\hat{\mathcal{P}}_N, \hat{\mathcal{P}}) = \max\{\mathsf{dl}_K(P_j^N, P_j) : j = 1, \ldots, J\}.$$

Likewise, we can show

$$\mathbb{D}_K(\hat{\mathcal{P}}, \hat{\mathcal{P}}_N) = \max\{\mathsf{dl}_K(P_j^N, P_j) : j = 1, \ldots, J\}.$$

A combination of the two equalities yields (3.49). Inequality (3.50) follows from (3.49) and (3.39). □

Following our discussion in section 3.3, condition for convergence of P_j^N to P_j under the Kantorovich metric is fulfilled if (3.37) holds. In particular the latter is satisfied when the support set of Ξ is bounded.

4 Approximation of (DRSRP')

In Section 3, we discussed four approaches to construct the ambiguity set $\mathcal{P}$ and argued in each approach that there is a need to approximate $\mathcal{P}$ by $\mathcal{P}_N$ because the ambiguity set contains samples. Under these circumstances, we have to solve (DRSRP') via solving the following minimization problem:

$$\text{(DRSRP'-N)} \quad \begin{cases} \min\limits_{X \in X, t \in T} & t \\ \text{s.t.} & \sup\limits_{P \in \mathcal{P}_N} \mathbb{E}_P[l(c(x,\xi) - t)] \leq \lambda. \end{cases} \tag{4.51}$$

Consequently, it is necessary to investigate finite sample guarantees on the quality of the optimal solutions obtained from solving (DRSRP'-N), a concept proposed by Esfahani and Kuhn [8], as well as convergence of the optimal values.

Let ϑ_N denote the optimal value of (DRSRP'-N) and S_N the corresponding optimal solution set. Let $x_N \in S_N$ and P^* denote the true probability distribution. We investigate the data-driven solution x_N with performance guarantee of the following type

$$\text{Prob}(\text{SR}^{P^*}_{l,\lambda}(-c(x_N,\xi)) \leq \vartheta_N) \geq 1 - \delta, \tag{4.52}$$

where $\delta \in (0,1)$ is called as a significance parameter, ϑ_N is a certificate for the out-of-sample performance of x_N and the probability on the left-hand side of (4.52) indicates ϑ_N's reliability. The following theorem states that the finite sample guarantee condition is fulfilled for the ambiguity sets discussed in Section 3, that is, when the size of the ambiguity sets are chosen carefully, the certificate ϑ_N can provide a $1-\delta$ confidence bound of the type (4.52) on the out-of-sample performance of x_N.

Theorem 4.1 (Finite sample guarantee) *The following assertions hold:*

(i) Let $\mathcal{P}(\mu_N, \Sigma_N, \gamma_1^N, \gamma_2^N)$ be defined as in (3.11) with $\mu_N, \Sigma_N, \gamma_1^N, \gamma_2^N$ being defined as in (3.14) and (3.16). Suppose the moment growth condition (3.15) holds, then with $\mathcal{P}_N = \mathcal{P}(\mu_N, \Sigma_N, \gamma_1^N, \gamma_2^N)$, the finite sample guarantee (4.52) holds.

(ii) Suppose the true probability distribution P^* is discrete, i.e., $\Xi = \{\zeta^1, \ldots, \zeta^M\}$. Let $\mathcal{P}_N^M$ be defined as (3.18) with r being given as (3.34), then with $\mathcal{P}_N = \mathcal{P}_N^M$, the finite sample guarantee (4.52) holds.

(iii) Let $\mathcal{P}_N$ be defined as in (3.37) with $r = r_N(\delta)$ being given in (3.39). Under condition (3.38), the finite sample guarantee (4.52) holds.

The results follow straightforwardly from (3.34), (3.38), (3.39) and the definition of finite sample guarantee. We now move on to investigate convergence of ϑ_N and S_N.

Theorem 4.2 (Convergence of the optimal values and optimal solutions) *Let $\mathcal{P}^* \subset \mathscr{P}(\Xi)$ be such that*

$$\lim_{N \to \infty} \mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*) = 0.$$

Let ϑ^* denote the optimal value of (DRSRP') *with $\mathcal{P}$ being replaced by $\mathcal{P}^*$. Let S^* be the corresponding optimal solutions. Under Assumption 2.1,*

$$|\vartheta_N - \vartheta^*| \leq \frac{2D_X L\kappa}{\eta}\mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*) \tag{4.53}$$

for N sufficiently large and

$$\limsup_{N\to\infty} S_N = S^*, \tag{4.54}$$

where D_X denotes the diameter of X, η is defined as in Proposition 2.2, and L, κ are defined as in Assumption 2.1.

Proof. Let

$$v^*(x,t) := \sup_{P\in\mathcal{P}^*} \mathbb{E}_P[l(c(x,\xi)-t)] - \lambda$$

and

$$v_N(x,t) := \sup_{P\in\mathcal{P}_N} \mathbb{E}_P[l(c(x,\xi)-t)] - \lambda.$$

By the definition

$$\begin{array}{rcl} v_N(x,t) - v^*(x,t) &=& \sup\limits_{P\in\mathcal{P}_N} \mathbb{E}_P[g(x,t,\xi)] - \sup\limits_{Q\in\mathcal{P}^*} \mathbb{E}_Q[g(x,t,\xi)] \\ &=& \sup\limits_{P\in\mathcal{P}_N} \inf\limits_{Q\in\mathcal{P}^*} (\mathbb{E}_P[g(x,t,\xi)] - \mathbb{E}_Q[g(x,t,\xi)]) \\ &\leq& \sup\limits_{P\in\mathcal{P}_N} \inf\limits_{Q\in\mathcal{P}^*} L\kappa \mathsf{dl}_K(P,Q) \\ &=& L\kappa\mathbb{D}_K(\mathcal{P}_N, \mathcal{P}^*), \end{array}$$

where the first inequality is due to equi-Lipschitz continuity of g in ξ and the definition of the Kantorovich metric. Likewise, we can establish

$$v^*(x,t) - v_N(x,t) \leq L\kappa\mathbb{D}_K(\mathcal{P}^*, \mathcal{P}_N).$$

Combining the above two inequalities, we obtain

$$\sup_{x\in X, t\in T} |v_N(x,t) - v^*(x,t)| \leq L\kappa\mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*). \tag{4.55}$$

Let

$$\mathcal{F}^* := \{(x,t) \in X\times T : v^*(x,t) \leq \lambda\},$$

and

$$\mathcal{F}_N := \{(x,t) \in X\times T : v_N(x,t) \leq \lambda\}.$$

By Proposition 2.2, v^* and v_N are convex on $X \times T$. Moreover, the Slater condition (2.7) allows us to apply Robinson's error bound for the convex inequality system (see [26]), i.e., there exists a positive constant C_1 such that for any $(x,t)\in X\times T$,

$$d((x,t), \mathcal{F}^*) \leq C_1[v^*(x,t) - \lambda]_+.$$

Let $(x,t) \in \mathcal{F}_N$. The inequality above enables us to estimate

$$\begin{aligned} d((x,t), \mathcal{F}^*) &\leq C_1[v^*(x,t) - \lambda]_+ \\ &\leq C_1(|v^*(x,t) - v_N(x,t)| + [v_N(x,t) - \lambda]_+) \\ &= C_1|v^*(x,t) - v_N(x,t)| \\ &\leq C_1 L\kappa \mathbb{H}_K(\mathcal{P}^*, \mathcal{P}_N). \end{aligned} \tag{4.56}$$

The last inequality follows from (4.55) and Robinson's error bound [26] ensures that the constant C_1 is bounded by D_X/η, where D_X is the diameter of X. This shows

$$\mathbb{D}(\mathcal{F}_N, \mathcal{F}^*) \leq \frac{D_X L\kappa}{\eta} \mathbb{H}_K(\mathcal{P}^*, \mathcal{P}_N).$$

On the other hand, the uniform convergence of v_N to v ensures

$$v_N(x_0, t_0) - \lambda < -\eta/2$$

for N sufficiently large, which means the convex inequality $v_N(x,t) - \lambda \leq 0$ satisfies the Slater condition. By applying Robinson's error bound for the inequality, we obtain

$$d((x,t), \mathcal{F}_N) \leq C_2|v^*(x,t) - v_N(x,t)| \leq C_2 L\kappa \mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*) \tag{4.57}$$

for $(x,t) \in \mathcal{F}^*$ and N is sufficiently large, where C_2 is bounded by $2D_X/\eta$. Combining (4.56) and (4.57), we obtain

$$\mathbb{H}(\mathcal{F}_N, \mathcal{F}^*) \leq \frac{2D_X L\kappa}{\eta} \mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*). \tag{4.58}$$

Let (x^*, t^*) be an optimal solution to (DRSRP') with $\mathcal{P}$ being replaced by $\mathcal{P}^*$ and (x_N, t_N) the optimal solution of (DRSRP'-N). Note that $\mathcal{F}_N, \mathcal{F}^* \subset X \times T$. Let

$$\Pi_T \mathcal{F} := \{t \in T : \text{there exists } x \in X \text{ such that } (x,t) \in \mathcal{F}\}.$$

Since $t_N = \min\{t : t \in \Pi_T \mathcal{F}_N\}$ and $t^* = \min\{t : t \in \Pi_T \mathcal{F}^*\}$, then

$$|t_N - t^*| \leq \mathbb{H}(\Pi_T \mathcal{F}_N, \Pi_T \mathcal{F}^*).$$

Thus

$$|\vartheta_N - \vartheta^*| = |t_N - t^*| \leq \mathbb{H}(\Pi_T \mathcal{F}_N, \Pi_T \mathcal{F}^*) \leq \mathbb{H}(\mathcal{F}_N, \mathcal{F}^*),$$

which yields (4.53) via (4.58).

Now, we move on to show (4.54). Let $(x_N, t_N) \in S_N$. Since X and T are compact, there exist a subsequence $\{(x_{N_k}, t_{N_k})\}$ and a point $(\hat{x}, \hat{t}) \in X \times T$ such that $(x_{N_k}, t_{N_k}) \to (\hat{x}, \hat{t})$. It follows by (4.58) and (4.53) that $(\hat{x}, \hat{t}) \in \mathcal{F}^*$ and $\hat{t} = \vartheta^*$. This shows $(\hat{x}, \hat{t}) \in S^*$. □

Theorem 4.2 is instrumental in that it provides a unified quantitative convergence result for the optimal value of (DRSRP'-N) in terms of $\mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*)$ when $\mathcal{P}_N$ is constructed in various ways discussed in Section 3. Based on the theorem and some quantitative convergence results about $\mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*)$, we can establish confidence intervals for the true optimal value ϑ^* in the following corollary.

Corollary 4.1 *Under the assumptions in Theorem 4.2, the following assertions hold.*

(i) If $\mathcal{P}^*$ comprises the true probability distribution only and $\mathcal{P}_N$ is defined by (3.18), then under conditions of Proposition 3.1,

$$\vartheta^* \in [\vartheta_N - \Theta, \vartheta_N + \Theta]$$

with probability $1 - M\delta$, where $\Theta := \frac{2D_X L\kappa}{\eta}\left[\beta_M + \frac{D}{2}\max\{2\sqrt{r}, r\} + \frac{D}{2}\Delta(M,N,\delta)\right]$ with $\Delta(M,N,\delta) = \min\left(\frac{M}{\sqrt{N}}\left(2+\sqrt{2\ln\frac{1}{\delta}}\right), 4 + \frac{1}{\sqrt{N}}\left(2+\sqrt{2\ln\frac{1}{\delta}}\right)\right)$, β being defined as in (3.22) and D being the diameter of Ξ.

(ii) If $\mathcal{P}^*$ comprises the true probability distribution only and $\mathcal{P}_N$ is defined by (3.36), then under conditions of Proposition 3.3,

$$\vartheta^* \in \left[\vartheta_N - \frac{4D_X L\kappa r_N(\delta)}{\eta}, \vartheta_N + \frac{4D_X L\kappa r_N(\delta)}{\eta}\right]$$

with probability $1-\delta$.

(iii) If $\mathcal{P}$ and $\mathcal{P}_N$ are defined as $\hat{\mathcal{P}}$ and $\hat{\mathcal{P}}_N$ in Proposition 3.4, then with probability $1-\delta$,

$$\vartheta^* \in \left[\vartheta_N - \left(2D_X L\kappa \max_{j=1,\ldots,J} r_N^j(\delta)\right)/\eta, \vartheta_N + \left(2D_X L\kappa \max_{j=1,\ldots,J} r_N^j(\delta)\right)/\eta\right].$$

4.1 Extension

Now we turn to extend the convergence result to optimization problems with DRSR constraints:

$$\text{(DRSRCP)} \quad \begin{array}{cl} \min\limits_{x\in X} & f(x) \\ \text{s.t.} & \mathrm{SR}_{l,\lambda}^{\mathcal{P}}(-c(x,\xi)) \le \gamma, \end{array} \tag{4.59}$$

where decision maker wants to optimize an objective $f(x)$ while requiring the DRSR risk level to be contained under threshold γ. By replacing $\mathcal{P}$ with $\mathcal{P}_N$, we may associate (DRSRCP) with

$$\text{(DRSRCP-N)} \quad \begin{array}{cl} \min\limits_{x\in X} & f(x) \\ \text{s.t.} & \mathrm{SR}_{l,\lambda}^{\mathcal{P}_N}(-c(x,\xi)) \le \gamma. \end{array} \tag{4.60}$$

Tractable reformulation of problem (DRSRCP) or (DRSRCP-N) may be derived as we did in Section 3. In what follows, we establish a theoretical quantitative convergence result for (DRSRCP-N).

Let $\hat{\mathcal{F}}$, $\hat{S}$ and $\hat{\vartheta}$ denote respectively the feasible set, the set of the optimal solutions and the optimal value of (DRSRCP). Likewise, we define $\hat{\mathcal{F}}_N$, $\hat{S}_N$ and $\hat{\vartheta}_N$ for its approximate problem (DRSRCP-N).

Theorem 4.3 *Let Assumptions 2.1 hold. Suppose that there exists $x_0 \in X$ such that*

$$\mathrm{SR}_{l,\lambda}^{\mathcal{P}}(-c(x_0,\xi)) < \gamma$$

and $\mathbb{H}_K(\mathcal{P}_N, \mathcal{P}) \to 0$ as $N \to \infty$. Then the following assertions hold.

(i) There is a constant $C > 0$ such that

$$\mathbb{H}(\hat{\mathcal{F}}_N, \hat{\mathcal{F}}) \leq C\mathbb{H}_K(\mathcal{P}_N, \mathcal{P})$$

for N sufficiently large.

(ii) $\lim\limits_{N\to\infty} \hat{\vartheta}_N = \hat{\vartheta}$ and $\limsup\limits_{N\to\infty} \hat{S}_N = \hat{S}$.

(iii) If, in addition, f is Lipschitz continuous with modulus β, then

$$|\hat{\vartheta}_N - \hat{\vartheta}| \leq \beta\mathbb{H}(\hat{\mathcal{F}}_N, \hat{\mathcal{F}}). \tag{4.61}$$

Moreover, if (DRSRCP) *satisfies the second order growth condition at the optimal solution set $\hat{S}$, i.e., there exist positive constants α and ε such that*

$$f(x) - \hat{\vartheta} \geq \alpha d(x, \hat{S})^2, \ \forall x \in \hat{\mathcal{F}} \cap (\hat{S} + \varepsilon\mathbb{B}),$$

then

$$\mathbb{D}(\hat{S}_N, \hat{S}) \leq \max\left\{2C, \sqrt{8C\beta/\alpha}\right\} \sqrt{\mathbb{H}_K(\mathcal{P}_N, \mathcal{P})} \tag{4.62}$$

when N is sufficiently large.

Proof. Part (i) can be established through an analogous proof of Theorem 4.2. We omit the details.

Part (ii). First we rewrite (DRSRCP) and (DRSRCP-N) as

$$\inf_{x\in\mathbb{R}^n} \tilde{f}(x) := f(x) + \delta_{\hat{\mathcal{F}}}(x) \tag{4.63}$$

and

$$\inf_{x\in\mathbb{R}^n} \tilde{f}_N(x) := f(x) + \delta_{\hat{\mathcal{F}}_N}(x), \tag{4.64}$$

where $\delta_{\hat{\mathcal{F}}}(x)$ is the indicator function of $\hat{\mathcal{F}}$, i.e.,

$$\delta_{\hat{\mathcal{F}}}(x) := \begin{cases} 0, & \text{if}\, x \in \hat{\mathcal{F}}, \\ \infty, & \text{if}\, x \notin \hat{\mathcal{F}}. \end{cases}$$

Note that the epigraph of $\delta_{\hat{\mathcal{F}}}(\cdot)$ is defined as

$$\text{epi}\, \delta_{\hat{\mathcal{F}}}(\cdot) := \{(x, \alpha) : \delta_{\hat{\mathcal{F}}}(x) \leq \alpha\} = \hat{\mathcal{F}} \times \mathbb{R}_+.$$

The convergence of $\hat{\mathcal{F}}_N$ to $\hat{\mathcal{F}}$ implies

$$\lim_{N\to\infty} \text{epi}\, \delta_{\hat{\mathcal{F}}_N}(\cdot) = \text{epi}\, \delta_{\hat{\mathcal{F}}}(\cdot),$$

and through [27, Definition 7.39] that $\delta_{\hat{\mathcal{F}}_N}(\cdot)$ epiconverges to $\delta_{\hat{\mathcal{F}}}(\cdot)$. Furthermore, it follows from [27, Theorem 7.46] that $\tilde{f}_N$ epiconverges to $\tilde{f}$. Since f is continuous and $\hat{\mathcal{F}}$ and $\hat{\mathcal{F}}_N$ are compact set, then $\{x_N\}$ has a subsequence converging to $\bar{x}$. By [5, Proposition 4.6], $\lim\limits_{N\to\infty} \hat{\vartheta}_N = \hat{\vartheta}$ and $\bar{x} \in \hat{S}$.

In what follows, we show Part (iii). Let $x_N \in \hat{S}_N$ and $x^* \in \hat{S}$. By the definition of $\mathbb{D}(\hat{\mathcal{F}}_N, \hat{\mathcal{F}})$, there exists $x' \in \hat{\mathcal{F}}$ such that $d(x_N, x') \leq \mathbb{D}(\hat{\mathcal{F}}_N, \hat{\mathcal{F}})$. Moreover, by the Lipschitz continuity of f, we have

$$\begin{array}{rcl} f(x^*) \leq f(x') & \leq & f(x_N) + |f(x_N) - f(x')| \leq f(x_N) + \beta|x_N - x'| \\ & \leq & f(x_N) + \beta\mathbb{D}(\hat{\mathcal{F}}_N, \hat{\mathcal{F}}). \end{array}$$

Exchanging the role of x_N and x^*, we have

$$f(x_N) \leq f(x^*) + \beta\mathbb{D}(\hat{\mathcal{F}}, \hat{\mathcal{F}}_N).$$

A combination of the two inequalities yields (4.61).

Next, we show (4.62). Let $x_N \in \hat{S}_N$ and $\overline{x} \in \hat{S}$. By the second order growth condition,

$$\begin{array}{rcl} f(x_N) - f(\Pi_{\hat{\mathcal{F}}}(x_N)) & = & f(x_N) - f(\overline{x}) - (f(\Pi_{\hat{\mathcal{F}}}(x_N)) - f(\overline{x})) \\ & \leq & f(\Pi_{\hat{\mathcal{F}}_N}(\overline{x})) - f(\overline{x}) - \alpha d(\Pi_{\hat{\mathcal{F}}}(x_N), \hat{S})^2, \end{array}$$

where $\Pi_{\hat{S}}(a)$ denotes the orthogonal projection of vector a on set $\hat{S}$, that is, $\Pi_{\hat{S}}(a) \in \arg\min_{s\in\hat{S}} \|s-a\|$. By the Lipschitz continuity of f, the inequality implies

$$d(\Pi_{\hat{\mathcal{F}}}(x_N), \hat{S}) \leq \sqrt{\beta/\alpha|\Pi_{\hat{\mathcal{F}}_N}(\overline{x}) - \overline{x}| + \beta/\alpha|\Pi_{\hat{\mathcal{F}}}(x_N) - x_N|}.$$

Therefore,

$$\begin{array}{rcl} d(x_N, \hat{S}) & \leq & |x_N - \Pi_{\hat{\mathcal{F}}}(x_N)| + d(\Pi_{\hat{\mathcal{F}}}(x_N), \hat{S}) \\ & \leq & |x_N - \Pi_{\hat{\mathcal{F}}}(x_N)| + \sqrt{\beta/\alpha|\Pi_{\hat{\mathcal{F}}_N}(\overline{x}) - \overline{x}| + \beta/\alpha|\Pi_{\hat{\mathcal{F}}}(x_N) - x_N|}. \end{array} \quad (4.65)$$

Since

$$\max\left\{\max_{x_N\in\hat{S}_N} \left|x_N - \Pi_{\hat{\mathcal{F}}}(x_N)\right|, \max_{\overline{x}\in\hat{S}} \left|\Pi_{\hat{\mathcal{F}}_N}(\overline{x}) - \overline{x}\right|\right\} \leq \mathbb{H}(\hat{\mathcal{F}}_N, \hat{\mathcal{F}}),$$

we have from inequality (4.65) and Part (i),

$$d(x_N, \hat{S}) \leq \max\left\{C, \sqrt{2C\beta/\alpha}\right\}\left[\mathbb{H}_K(\mathcal{P}_N, \mathcal{P}) + \sqrt{\mathbb{H}_K(\mathcal{P}_N, \mathcal{P})}\right].$$

The last inequality implies (4.65) in that x_N is arbitrarily chosen from $\hat{S}_N$ and $\mathbb{H}_K(\mathcal{P}_N, \mathcal{P}) \leq \sqrt{\mathbb{H}_K(\mathcal{P}_N, \mathcal{P})}$ when N sufficiently large. The proof is complete. □

Analogous to Corollary 4.1, we can derive confidence intervals and regions for the optimal value and the optimal solutions with different $\mathcal{P}_N$.

5 Application in portfolio optimization

In this section, we apply the (DRSRP) model to decision-making problems in portfolio optimization. Let ξ_i denote the rate of return from investment on stock i and x_i denote the capital invested in the stock i for $i = 1, \ldots, d$. The total return from the investment of the d stocks is $x^T\xi$, where we write ξ for $(\xi_1, \xi_2, \ldots, \xi_d)^T$ and x for $(x_1, x_2, \ldots, x_d)^T$. We consider a situation where the investor's decision on allocation of the capital is based on minimization of the distributionally robust shortfall of $x^T\xi$, that is, $\mathrm{SR}^{\mathcal{P}}_{l,\lambda}(x^T\xi)$ for some specified l, λ and $\mathcal{P}$, that is, the investor finds an optimal decision x^* by solving

$$\begin{array}{cl} \min\limits_{x\in X, t\in \mathbb{R}} & t \\ \text{s.t.} & \sup\limits_{P\in\mathcal{P}} \mathbb{E}_P[l(-x^T\xi - t)] \leq \lambda. \end{array} \tag{5.66}$$

We have undertaken numerical experiments on problem (5.66) from different perspectives ranging from efficiency of computational schemes applied to problem (5.66) as we discussed in Section 3, the out-of-sample performance of the optimal portfolio and the growth of the total portfolio value over a specified time horizon using different optimal strategies.

We begin by looking into problem (5.66) with the ambiguity set being constructed through moment conditions. Our focus is on the performance of the inner and outer approximation schemes discussed in Section 3.1.

Example 5.1 Consider problem (5.66) with the ambiguity set being defined as

$$\mathcal{P} := \left\{ P \in \mathscr{P}(\Xi) : \begin{array}{l} \mathbb{E}_P[\xi] = \mu, \\ \mathbb{E}_P[(\xi-\mu)(\xi-\mu)^T] = \Sigma \end{array} \right\},$$

where μ and Σ are the true mean and variance. We assume that both quantities are known with $\mu_i = 0.04 + 0.46(i-1)/(d-1)$ for $i = 1, \ldots, d$ and $\Sigma_{ii} = (\mu_i + 0.05)^2$ and $\Sigma_{ij} = 0.35(\mu_i + 0.05)(\mu_j + 0.05)$ for $i \neq j$.

We set the loss function $l(z) := \exp(z)$, $X := \{x \in \mathbb{R}^d : \sum_{i=1}^d x_i = 1, x \geq 0\}$ and $\lambda = 0.1$. Obviously $l(\cdot)$ is a monotonically increasing and strictly convex function, the total capital is normalized to one and no short selling is allowed. This experiment is to examine the inner and outer approximation schemes discussed in Section 3.1. The loss function is approximated by a piecewise affine function described as in (3.12). The resulting optimization problem is described as

$$\begin{array}{cl} \min\limits_{x,t,y_0,y_1,Y_2} & t \\ \text{s.t.} & y_0 + y_1^T\mu + \langle Y_2, \mu\mu^T + \Sigma\rangle \leq \lambda, \\ & \begin{bmatrix} -Y_2 & (-a_jx - y_1)/2 \\ (-a_jx - y_1)^T/2 & -a_jt + b_j - y_0 \end{bmatrix} \preceq 0,\ j = 1, \ldots, K, \\ & x \in X. \end{array} \tag{5.67}$$

and we solve the latter using toolbox CVX installed in MATLAB. Figure 1 depicts change of CPU time and the gap against increment of the number of linear pieces (denoted by K). It shows that the CPU time increases at a linear rate w.r.t. increase of K. Looking at problem (5.67), the approximation scheme affects the problem through the K semidefinite constraints and the trends that we have observed from Figure 1 shows roughly that the CPU time is linearly dependent on the number of the semidefinite constraints. The figure also shows that the gap between the inner and outer approximations tends to be very close when $K \geq 60$.

Figure 2 displays change of CPU time against variation of the number of stocks (d) for fixed $K = 60$. The CPU time increases rapidly w.r.t. increment of d. This may be partly explained by the fact that the size of the matrix in each of the semidefinite constraints depends on d (which is $(d+1) \times (d+1)$).

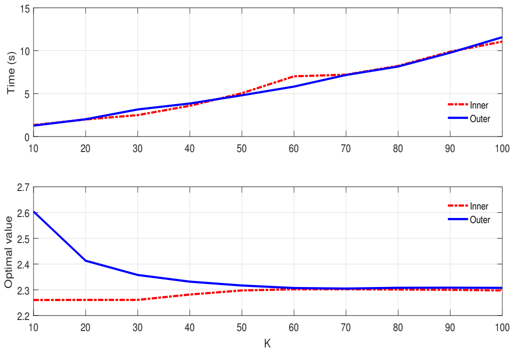

Figure 1: CPU time and optimal value w.r.t K.

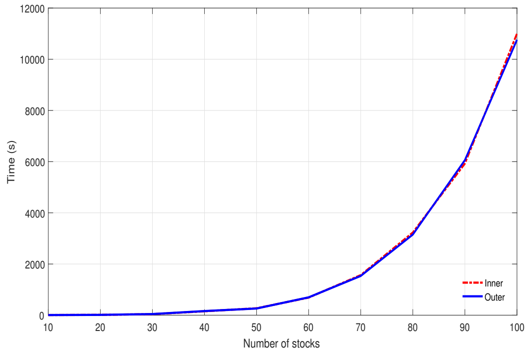

Figure 2: CPU time w.r.t d.

The following numerical experiments focus on problem (5.66) with the ambiguity set being defined through the Kantorovich ball. We report the details in Example 5.2.

Example 5.2 Let $\xi^1, \ldots, \xi^N$ be iid samples of ξ and P_N be the nominal distribution constructed through the samples, that is, $P_N(\cdot) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\xi^i}(\cdot)$. The ambiguity set is defined respectively as

$$\mathcal{P}_N = \{P \in \mathscr{P}(\mathbb{R}^d) : \mathsf{dl}_K(P, P_N) \le r\}. \tag{5.68}$$

To simplify the tests, we consider a specific piecewise affine loss function $l(t) = \max\{0.05t+1, t+0.1, 4t+2\}$. We set $\lambda = 1$ and let the total number of stocks d be fixed at 10. We follow Esfahani and Kuhn [8] to generate the iid samples by assuming that the rate of return ξ_i is decomposable into a systematic risk factor $\psi \sim \mathcal{N}(0, 2\%)$ common to all stocks and an unsystematic risk factor $\zeta_i \sim \mathcal{N}(i \times 3\%, i \times 2.5\%)$ specific to stock i, that is, $\xi_i = \psi + \zeta_i$, for $i = 1, \ldots, d$. Based on the discussions in Sections 3.3, problem (5.66) can be reformulated through dual formulation as

$$\begin{array}{lll} J_N(r) := & \displaystyle\min_{x\in X, t\in T, \eta, s} & t \\ & \text{s.t.} & \eta r + \dfrac{1}{N}\displaystyle\sum_{i=1}^N s_i \le 1, \\ & & b_j - a_j t - \langle a_j x, \xi^i\rangle \le s_i, \text{ for } i = 1, \ldots, N, j = 1, 2, 3, \\ & & \|a_j x\|_\infty \le \eta, \text{ for } j = 1, 2, 3. \end{array} \tag{5.69}$$

Following the terminology of Esfahani and Kuhn [8], we call $J_N(r)$ the certificate.

In the first set of experiments, we investigate the impact of the radius of the Kantorovich ball r on the out-of-sample performance of the optimal portfolio. For any fixed portfolio $x_N(r)$ obtained from problem (5.69), the out-of-sample performance is defined as $J(x_N(r)) := \mathrm{SR}_{l,\lambda}^{P^*}(x_N(r)^T\xi)$, which can be computed from theoretical point of view since the true probability distribution P^* is known by design although in the experiment we will generate a set of validation samples of size 2×10^5 to do the evaluation. Following the same strategy as in [8], we generate the training datasets of cardinality $N \in \{30, 300, 3000\}$ to solve problem (5.69) and then use the same validation samples to evaluate $J(x_N(r))$. Each of the experiments is carried out through 200 simulation runs.

Figures 3-5 depict the tubes between the 20% and 80% quanrtiles (shaded areas) and the means (solid lines) of the out-of-sample performance $J(x_N(r))$ as a function of radius r, the

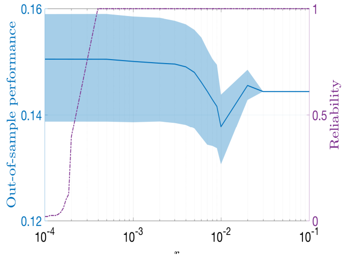

Figure 3: $N = 30$

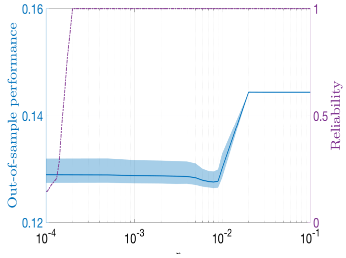

Figure 4: $N = 300$

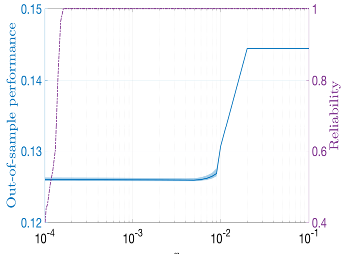

Figure 5: $N = 3000$

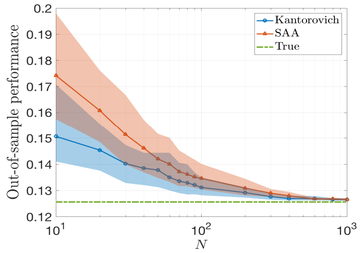

Figure 6:

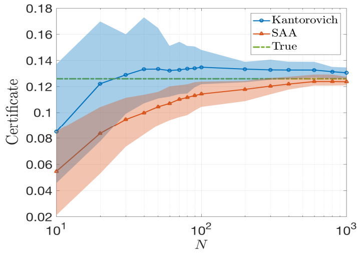

Figure 7:

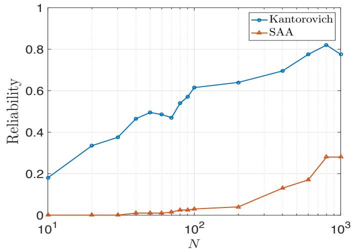

Figure 8:

dashed lines represent the empirical probability of the event $J(x_N(r)) \leq J_N(r)$ with respect to 200 independent runs which is called reliability in Esfahani and Kuhn [8]. It is clear that the reliability is nondecreasing in r and this is because the true probability distribution P^* is located in $\mathcal{P}_N$ more likely as r grows and hence the event $J(x_N(r)) \leq J_N(r)$ happens more likely. The out-of-sample performance of the portfolio improves (decreases) first and then deteriorates (increases).

In the second set of experiments, we investigate convergence of the out-of-sample performance, the certificate and the reliability of the distributionally robust approach and the sample average approximation (SAA) as the size of sample increases. Note that SAA corresponds to the case when the radius r of the Kantorovich ball is zero. In all of the tests we use cross validation method in [8] to select the Kantorovich radius from the discrete set $\{\{5, 6, 7, 8, 9\} \times 10^{-3}, \{0, 1, 2, \ldots, 9\} \times 10^{-2}, \{0, 1, 2, \ldots, 9\} \times 10^{-1}\}$. We have verified that refining or extending the above discrete set has only a marginal impact on the results.

Figure 6 shows the tubes between the 20% and 80% quantiles (shaded areas) and the means (solid lines) of the out-of-sample performance $J(x_N)$ as a function of the sample size N based on 200 independent simulation runs, where x_N is the minimizer of (5.69) and its SAA counterpart ($r = 0$). The constant dashed line represents the optimal value of the SAA problem with $N = 10^6$ samples which is regarded as the optimal value of the original problem with the true probability distribution. It is observed that the robust model outperforms the SAA model in terms of out-of-sample performance. Figure 7 depicts the optimal values of the DRO model and the SAA counterpart, which is the in-sample estimate of the obtained portfolio performance. Both of the approaches display asymptotic consistency, which is consistent with the out-of-sample and in-sample results. Figure 8 describes the empirical probability of the event $J(x_N) \leq J_N$ with respect to 200 independent runs, where x_N is the optimal value of the DRO model or SAA model, and J_N are the optimal value of the corresponding problem. It is clear that the performance of the DRO model is better than that of the SAA model.

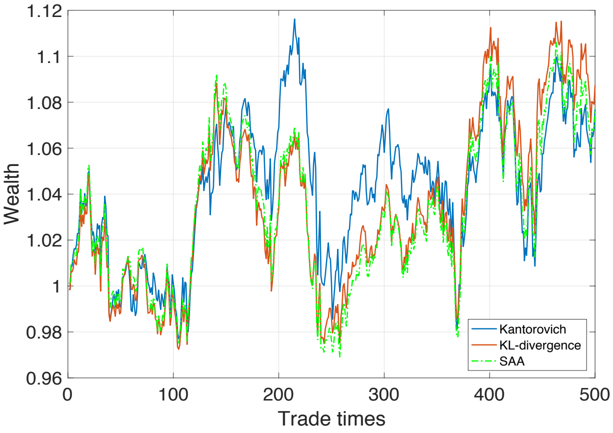

Figure 9: Wealth evolution with the trading times.

Example 5.3 In this set of experiments, we examine the performance of problem (5.66) with real data for the ambiguity set being constructed through the KL-divergence and the Kantorovich ball, specifically we have undertaken tests on problem (5.66) with 10 stocks (Apple Inc., Amazon.com, Inc., Baidu Inc., Costco Wholesale Corporation, DISH Network Corp., eBay Inc., Fox Inc., Alphabet Inc Class A, Marriott International Inc., QUALCOMM Inc.) where their historical data are collected from National Association of Securities Deal Automated Quotations (NASDAQ) index over 4 years (from 3rd May 2011 to 23rd April 2015) with total of 1000 records on the historical stock returns.

We have carried out out-of-sample tests with a rolling window of 500 days, that is, we use the first 500 data to calculate the optimal portfolio strategy for day 501 and then move on a rolling basis. The radiuses in the two ambiguity sets are selected through the cross validation method. Figure 5 depicts the performance of three models over 500 trading days. It seems that the KL-divergence model and SAA model perform similarly, whereas the Kantorovich model outperforms the both over most of the time period.

We have also carried out tests on problem (5.66) with the ambiguity set being constructed through mixture distribution.

Example 5.4 The example is varied from test examples in Zhu and Fukushima [36, Section 3.2.1]. Consider four different assets $A1, A2, A3$ and $A4$, a total of 2700 samples of daily returns of these assets are known. The samples are divided evenly into three groups with 900 samples specified by three different time periods and then we calculate empirical mean and covariance for each group (see Table 2). We can see that the difference of these quantities are significant and this motivates us to consider different distribution for each of the groups.

In the tests, we consider the loss function $l(z) = \exp(z)$ and set $\lambda = 0.1$. We also impose an additional constraint

$$\min_{i=1,2,3} \mathbb{E}_{P^i_{900}}[x^T \xi] \geq u \tag{5.70}$$

to the (DRSRP) model. Our focus is on comparison of the (DRSRP) model with the SAA model. Note that in the (DRSRP) model, we use 900 samples in each group to construct three

Group	Mean (10^{-3})				Variance (10^{-3})			
	A1	A2	A3	A4	A1	A2	A3	A4
Group 1	1.9294	0.8325	1.4316	0.5413	0.2192	0.2042	0.2376	0.2146
Group 2	0.5027	0.4965	0.3648	0.8378	0.1886	0.1779	0.2399	0.2082
Group 3	-0.2762	-0.1785	-0.3036	0.8226	0.5438	0.4539	0.7790	0.7392

Table 1: Emprical mean and variance of returns of four assets in each group

empirical distributions denoted by $P^1_{900}, P^2_{900}, P^3_{900}$ and in the SAA model, we use all 2700 samples to construct one empirical distribution. The additional constraint in the SAA model is

$$\mathbb{E}_{P_{2700}}[x^T\xi] \geq u. \tag{5.71}$$

u (10^{-3})	DRSRP (I) SAA (II)	Mean (10^{-3})			Optimal value
		Group 1	Group 2	Group 3	
0	I	0.5553	0.8345	0.8116	2.3021
	II	1.2722	0.6614	0.2441	2.3019
0.30	I	0.5553	0.8345	0.8116	2.3021
	II	1.2722	0.6614	0.2441	2.3019
0.56	I	0.5600	0.8333	0.8078	2.3021
	II	1.2722	0.6614	0.2441	2.3019
0.60	I	0.6000	0.8237	0.7762	2.3021
	II	1.2722	0.6614	0.2441	2.3019
0.73	I	—	—	—	—
	II	0.8984	0.7516	0.5400	2.3020

Table 2: Comparison of performance of DRSRP and SAA models

u (10^{-3})	DRSRP portfolio		SAA portfolio	
	mean (10^{-3})	variance (10^{-3})	mean (10^{-3})	variance (10^{-3})
0	0.7338	0.3791	0.7259	0.1693
0.30	0.7338	0.3791	0.7259	0.1693
0.56	0.7337	0.3765	0.7259	0.1693
0.60	0.7333	0.3647	0.7259	0.1693
0.73	—	—	0.7300	0.2303

Table 3: Mean and variance of the portfolio returns

Table 3 lists the five u values that are used in the test, the mean return in each group with (DRSRP) based optimal strategy and the SAA based optimal strategy, and the optimal value. The constraint (5.71) in the SAA model is inactive at the optimal solution for $u = 0, 0.00030, 0.00056,\ 0.00060$ and it becomes active for $u = 0.00073$. On the other hand, the constraint (5.70) is inactive for $u = 0, 0.00030$, active for $u = 0.00056, 0.00060$, and infeasible for $u = 0.00073$. For the same u, the optimal value of (DRSRP) model is slightly larger than that of the SAA model because the feasible set of the former is smaller. Table 4 illustrates the means and variances of the two portfolio returns computed by the total 2700 samples. We see that the mean and variance of the (DRSRP) portfolio return are both larger than that of the SAA portfolio return, which means the robust strategy brings higher return on average with larger variances. This finding is consistent with the observations made by Zhu and Fukushima [36].

Acknowledgements. We would like to thank Peyman M. Esfahani for sharing with us some programmes for generating Figures 3-8 and instrumental discussions about implementation of the numerical experiments. We would also like to thank three anonymous referees for insightful comments which help us significantly strengthen the paper.

References

[1] P. Artzner, F. Delbaen, J. M. Eber and D. Health, Coherent measures of risk, *Math. Finance*, 9: 203-228, 1999.

[2] Basel Committee on Banking Supervision, Fundamental review of the trading book: A revised market risk framework, Bank for International Settlements 2013, `http://www.bis.org/publ/bcbs265.htm`.

[3] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg and G. Rennen, Robust solutions of optimization problems affected by uncertain probabilities, *Management Sci.*, 59: 341-357, 2013.

[4] P. Billingsley, *Convergence of probability measures*, John Wiley, New York, 1968.

[5] J. F. Bonnans and A. Shapiro, *Perturbation analysis of optimization problems*, Springer, New York, 2000.

[6] E. Delage and Y. Ye, Distributionally robust optimization under moment uncertainty with application to data-driven problems, *Oper. Res.*, 58: 595-612, 2010.

[7] J. Dunkel and S. Weber, Stochastic root finding and efficient estimation of convex risk measures, *Oper. Res.*, 58: 1505-1521, 2010.

[8] P. M. Esfahani and D. Kuhn, Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations, *Math. Program.,* DOI 10.1007/s10107-017-1172-1, 2017.

[9] H. Föllmer and A. Schied, Convex measures of risk and trading constraints, *Finance Stochast.*, 6: 429-447, 2002.

[10] H. Föllmer and A. Schied, *Stochastic Finance-An Introduction in Discrete Time*, Walter de Gruyter, Berlin, 2011.

[11] N. Fournier and A. Guilline, On the rate of convergence in Wasserstein distance of the empirical measure, *Probab. Theory Relat. Fields*, 162: 707-738, 2015.

[12] M. Frittelli and E. Rosazza Gianin, Putting order in risk measures, *J. Banking and Finance*, 26: 1473-1486, 2002

[13] R. Gao and A.J. Kleywegt, Distributionally robust stochastic optimization with Wasserstein distance, arXiv preprint arXiv:1604.02199, 2016.

[14] A. L. Gibbs and F. E. Su, On choosing and bounding probability metrics, *International statistical review*, 70: 419-435, 2002.

[15] K. Giesecke, T. Schmidt and S. Weber, Measuring the risk of large losses, *J. Investment Management*, 6: 1-15, 2008.

[16] J. A. Hall, B. W. Brorsen and S. H. Irwin, The distribution of futures prices: a test of stable Paretian and mixture of normals hypotheses, *Journal of Financial and Quantitative Analysis*, 24: 105-116, 1989.

[17] D. Heath, *Back to the future*, Plenary Lecture at the First World Congress of the Bachelier Society, Paris, 2000.

[18] Z. Hu and D. Zhang, Convex risk Measures: efficient computations via monte carlo, manuscript, 2016.

[19] Y. Liu, A. Pichler and H. Xu, Discrete approximation and quantification in distributionally robust Optimization, to appear in *Math. Oper. Res.*, 2017.

[20] D. Love and G. Bayrakcan, Phi-divergence constrained ambiguous stochastic programs for data-driven optimization, available on Optimization Online, 2016.

[21] J. Moulton, Robust fragmentation: a data-driven approach to decision-making under distributional ambiguity, Ph.D. Dissertation, University of Minnesota, 2016.

[22] L. Pardo, *Statistical Inference Based on Divergence Measures*, Chapman and Hall/CRC, Boca Raton, FL, 2005.

[23] D. Peel and G. J. McLachlan, Robust mixture modeling using t distribution, *Stat. Comput.*, 10: 339-348, 2000.

[24] G. C. Pflug and A. Pichler, *Multistage stochastic optimization*, Springer International Publishing Switzerland, 2014.

[25] I. Pólik and T. Terlaky, A survey of the S-lemma, *SIAM Rev.*, 49: 371-418, 2007.

[26] S. M. Robinson, An application of error bounds for convex programming in a linear space, *SIAM J. Control*, 13: 271-273, 1975.

[27] R. T. Rockafellar and R. J. B. Wets, *Variational analysis*, Springer, New York, 1998.

[28] H. Scarf, A min-max solution of an inventory problem. K.J. Arrow, S. Karlin, H. Scarf, eds., Studies in the Mathematical Theory of Inventory and Production. Stanford University Press, Stanford, CA, 201-209, 1958.

[29] J. Shawe-Taylor and N. Cristianini, Estimating the moments of a random vector with applications, Proc. GRETSI 2003 Conf., 47-52, 2003.

[30] A. M. C. So, Moment inequalities for sums of random matrices and their applications in optimization, *Math. Program.*, 130: 125-151, 2011.

[31] H. Sun and H. Xu, Convergence analysis for distributionally robust optimization and equilibrium problems, *Math. Oper. Res.*, 41: 377-401, 2016.

[32] S. Weber, Distribution-invariant risk measures, information, and dynamic consistency, *Math. Finance*, 16: 419-442, 2006.

[33] W. Wiesemann, D. Kuhn and M. Sim, Distributionally robust convex optimization, *Oper. Res.*, 62: 1358-1376, 2014.

[34] H. Xu, Y. Liu and H. Sun, Distributionally robust optimization with matrix moment constraints: lagrange duality and cutting-plane methods, *Math. Program.*, DOI 10.1007/s10107-017-1143-6, 2017.

[35] C. Zhao anf Y. Guan, Data-driven risk-averse stochastic optimization with Wasserstein metric, available on Optimization Online, 2015.

[36] S. Zhu and M. Fukushima, Distributionally robust conditional Value-at-Risk with application to robust portfolio management, *Oper. Res.*, 57: 1155-1156, 2009.

[37] J. Zhang, H. Xu and L.W. Zhang, Quantitative Stability Analysis for Distributionally Robust Optimization with Moment Constraints, *SIAM J. Optimization*, 26: 1855-1882, 2016.