CrossMark

# Distributionally robust shortfall risk optimization model and its approximation

**Shaoyan Guo[1] · Huifu Xu[2]**

*Dedicated to Professor Roger J-B. Wets on the occasion of his 80th birthday*

**Abstract** Utility-based shortfall risk measures (SR) have received increasing attention over the past few years for their 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 measure (DRSR) 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 through $\phi$-divergence ball and Kantorovich ball. In the case when the nominal distribution in the balls is an empirical distribution constructed through iid samples, we quantify convergence of the ambiguity sets to the true probability distribution as the sample size increases under the Kantorovich metric and consequently the optimal values of the corresponding DRSR problems. 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 prelim-

✉ Huifu Xu
  h.xu@soton.ac.uk

  Shaoyan Guo
  syguo@dlut.edu.cn

[1] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China

[2] School of Mathematics, University of Southampton, Southampton SO17 1BJ, UK

⌂ Springer

inary numerical test results are reported for the proposed modeling and computational schemes.

**Keywords** DRSR · Kantorovich metric · $\phi$-divergence ball · Kantorovich ball · Quantitative convergence analysis

**Mathematics Subject Classification** 90C15 · 90C47 · 90C31 · 91B30

## 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 $Z : (\Omega, \mathscr{F}, P) \to \mathbb{R}$, where $\Omega$ is a sample space with sigma algebra $\mathscr{F}$ and $P$ is a probability measure. A risk measure $\rho$ 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].

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 positive homogeneity and 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 $\lambda$, 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 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 [26, Example 3.4]. The other is that CVaR is not particularly sensitive to heavy tailed losses [15, Sect. 5]. Here, we illustrate this by a simple example. Let

$$X_1 := \begin{cases} 100, & p_1 = 98\% \\ -100, & p_2 = 1\% \\ -200, & p_3 = 1\% \end{cases},$$

$$X_2 := \begin{cases} 100, & p_1 = 98\% \\ -1, & p_2 = 1\% \\ -299, & p_3 = 1\% \end{cases},$$

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

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

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 $\lambda$ 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

$$(\text{SR}) \quad \text{SR}^P_{l,\lambda}(Z) := \inf\{t \in \mathbb{R} : t + Z \in \mathcal{A}_P\}, \tag{2}$$

where $\mathcal{A}_P := \{Z \in L^\infty : \mathbb{E}_P[l(-Z(\omega))] \leq \lambda\}$ is called the acceptance set and $L^\infty$ 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 $z \in (0, +\infty)$, and 0 otherwise, in such a case $\text{SR}^P_{l,\lambda}(Z)$ coincides with $\text{VaR}_\lambda(Z)$. Of course, here $l$ is nonconvex.

Compared to CVaR, SR not only satisfies convexity, but also satisfies invariance under randomization and can be used more appropriately for dynamic measurement of risks over time. To see invariance under randomization, we note that SR defined as in (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 [26, Remark 2.1]. In the latter case, the acceptance set can be characterized by $\mathcal{N} := \{\mu \in \mathscr{P}(C) : \int_C l(-x)\mu(dx) \leq \lambda\}$, where $\mathscr{P}(C)$ denotes the space of probability measures with support being contained in a compact set $C \subset \mathbb{R}$. If $\mu, \nu \in \mathcal{N}$, i.e., $\int_C l(-x)\mu(dx) \leq \lambda$, $\int_C l(-x)\nu(dx) \leq \lambda$, then for any $\alpha \in (0, 1)$, $\int_C l(-x)(\alpha\mu + (1 - \alpha)\nu)(dx) \leq \lambda$, which means $\alpha\mu + (1 - \alpha)\nu \in \mathcal{N}$. Moreover, the SR is found to be more sensitive to financial losses from extreme events with heavy tailed distributions, see [15, Sect. 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 (1) with $\text{SR}^P_{l,\lambda}(X_1) \approx 194$, $\text{SR}^P_{l,\lambda}(X_2) \approx 293$, and $\text{SR}^P_{l,\lambda}(X_3) \approx 393$. 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,

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

In the case when $l(z) = z^{\alpha}\mathbb{1}_{[0,+\infty)}(z)$ with $\alpha \geq 1$, 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 (2) in order to hedge the risk arising from ambiguity of the true probability distribution,

$$\text{(DRSR)} \quad \mathrm{SR}_{l,\lambda}^{\mathcal{P}}(Z) := \inf\{t \in \mathbb{R} : t + Z \in \mathcal{A}_{\mathcal{P}}\}, \tag{3}$$

where $\mathcal{A}_{\mathcal{P}} := \{Z \in L^{\infty} : \sup_{P \in \mathcal{P}} \mathbb{E}_P[l(-Z)] \leq \lambda\}$, 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 (DRSR). In [10, Corollary 4.119], they established a robust representation theorem for DRSR. More recently, Wiesemann et al. [27] demonstrated how a 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 [27, 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 $x \in X \subset \mathbb{R}^n$ and random vector $\xi \in \mathbb{R}^k$, we consider an optimization problem which aims to minimize the distributionally robust shortfall risk measure of the random loss:

$$\text{(DRSRP)} \quad \min_{x \in X} \mathrm{SR}_{l,\lambda}^{\mathcal{P}}(-c(x, \xi)), \tag{4}$$

where $\mathrm{SR}_{l,\lambda}^{\mathcal{P}}(\cdot)$ is defined as in (3). We present a detailed discussion on (DRSRP) including tractable reformulation for the problem when the ambiguity set has a specific structure.

As far as we are concerned, the main contribution of the paper can be summarized as follows. First, we demonstrate that DRSR is the worst-case SR (Proposition 1)

and hence it is a convex risk measure. Second, we investigate tractability of (DRSRP) by considering particular cases where the ambiguity set $\mathcal{P}$ is constructed respectively through $\phi$-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 (Propositions 3 and 5). To quantify how the errors arising from the ambiguity set propagate to the optimal value of (DRSRP), we then 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 finite sample guarantee (Theorem 1) and confidence intervals for the optimal value of (DRSRP) associated with the ambiguity sets (Theorem 2 and Corollary 1). Finally, as an application, we apply the (DRSRP) model to a portfolio management problem and carry out various out-of-sample tests on the numerical schemes for the (DRSRP) model with simulated data and real data (Sect. 5).

The rest of the paper is organised as follows. In Sect. 2, we present the properties of DRSR, that is, it is a convex risk measure and it is the worst-case SR. In Sect. 3, we derive the formulation of (DRSRP) when the ambiguity set is constructed through $\phi$-divergence ball and Kantorovich ball and then establish the convergence of ambiguity sets as sample size increases. In Sect. 4, the finite sample guarantees on the quality of the optimal solutions and convergence of the optimal values as the sample size increases are discussed. In Sect. 5, we report results of numerical experiments.

Throughout the paper, we use $\mathbb{R}^n$ to represent $n$ dimensional Euclidean space and $\mathbb{R}^n_+$ nonnegative orthant. Given a norm $\| \cdot \|$ in $\mathbb{R}^n$, the dual norm $\| \cdot \|_*$ is defined by $\|y\|_* := \sup_{\|z\| \leq 1} \langle y, z \rangle$. Let $d(x, A) := \inf_{x' \in A} \|x - x'\|$ be the distance from a point $x$ to a set $A \subset \mathbb{R}^n$. 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 $B$. We use $\mathbb{B}$ to denote the unit ball in a matrix or vector space. Finally, for a sequence of subsets $\{S_N\}$ 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 $SR^{\mathcal{P}}_{l,\lambda}(Z)$ is the optimal value of the following minimization problem:

$$\begin{aligned} &\min_{t \in \mathbb{R}} t \\ &\text{s.t. } \sup_{P \in \mathcal{P}} \mathbb{E}_P[l(-Z - t)] \leq \lambda. \end{aligned} \quad (5)$$

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

**Proposition 1** *Let $SR^{\mathcal{P}}_{l,\lambda}(Z)$ be defined as in (3), $Z \in L^\infty$ and $l : \mathbb{R} \to \mathbb{R}$ be a convex, increasing and non-constant function, let $\lambda$ be a pre-specified constant in the range of $l$. Then $SR^{\mathcal{P}}_{l,\lambda}(Z)$ is finite,*

$$SR_{l,\lambda}^{\mathcal{P}}(Z) = \sup_{P \in \mathcal{P}} SR_{l,\lambda}^{P}(Z), \qquad (6)$$

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

*Proof* Since $Z$ is bounded, then there exist constants $\alpha$, $\beta$ 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 problem (5) is bounded. To show equality (6), we note that

$$\hat{t} := \sup_{P \in \mathcal{P}} SR_{l,\lambda}^{P}(Z) \leq \inf_t \sup_{P \in \mathcal{P}} \{t \in \mathbb{R} : \mathbb{E}_P[l(-Z - t)] \leq \lambda\}$$

$$\leq \inf_t \{t \in \mathbb{R} : \sup_{P \in \mathcal{P}} \mathbb{E}_P[l(-Z - t)] \leq \lambda\} = SR_{l,\lambda}^{\mathcal{P}}(Z) =: t^*.$$

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

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

which implies $\hat{t}$ is a feasible solution of (5) and hence $t^* \leq \hat{t}$. □

*Remark 1* It may be helpful to make some comments on Proposition 1.

(i) The relationship established in (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. In this case, the SR gives rise to an expectile, see [3, Theorem 4.9]. Using this result, we can easily show through equation (6) that DRSR is a coherent risk measure when $l$ takes the specific form in that the operation $\sup_{P \in \mathcal{P}}$ preserves positive homogeneity and subadditivity.

(ii) The restriction of $Z$ to $L^\infty$ 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

---

[1] The support of the probability distribution $P$ is the smallest closed set $C \subset \mathbb{R}$ such that $P(C) = 1$.

Sects. 3 and 4. To this end, we need to make some assumptions on the random function $c(\cdot, \cdot)$ and the loss function $l(\cdot)$. Throughout this section, we use $\varXi$ to denote the image space of random variable $\xi(\omega)$ and $\mathscr{P}(\varXi)$ to denote the set of all probability measures defined on the measurable space $(\varXi, \mathscr{B})$ with Borel sigma algebra $\mathscr{B}$. To ease notation, we will use $\xi$ to denote either the random vector $\xi(\omega)$ or an element of $\mathbb{R}^k$ depending on the context.

**Assumption 1** Let $X$, $l(\cdot)$ and $c(\cdot, \cdot)$ be defined as in (DRSRP) (4). We assume the following. (a) $X$ is a convex and compact set and $\varXi$ 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 $\xi \in \varXi$ and there exists a positive constant $\kappa$ such that

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

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

**Proposition 2** Let $g(x, t, \xi) := l(c(x, \xi) - t)$ and $v(x, t) := \sup_{P \in \mathcal{P}} \mathbb{E}_P[g(x, t, \xi)] - \lambda$. The following assertions hold.

(i) *Under Assumption 1 (b) and (c), $g(\cdot, \cdot, \xi)$ is convex w.r.t. $(x, t)$ for each fixed $\xi \in \varXi$, $g(x, t, \cdot)$ is uniformly Lipschitz continuous w.r.t. $\xi$ with modulus $L\kappa$, and $v(x, t)$ is a convex function w.r.t. $(x, t)$.*

(ii) *If, in addition, Assumption 1 (a) holds and $\lambda$ 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*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[l(c(x_0, \xi) - t_0)] - \lambda < -\eta \tag{7}$$

*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.

Part (ii). Since $c(x, \xi)$ is finite valued and convex in $x$, it is continuous in $x$ for each fixed $\xi$. Together with its uniform continuity in $\xi$, we are able to show that $c(x, \xi)$ is continuous over $X \times \varXi$. By the boundedness of $X$ and $\varXi$, there is a positive constant $\alpha$ such that $c(x, \xi) \leq \alpha$ for all $(x, \xi) \in X \times \varXi$. 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 1. We omit the details. □

## 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 (5) for DRSR, we can reformulate (DRSRP) as

$$
\text{(DRSRP')} \quad \begin{array}{c} \min\limits_{x \in X, t \in T} t \\ \text{s.t.} \quad \sup\limits_{P \in \mathcal{P}} \mathbb{E}_P[l(c(x, \xi) - t)] \leq \lambda, \end{array} \tag{8}
$$

where $T$ is a compact set in $\mathbb{R}$ which contains $t_0$ defined as in (7) and its existence is ensured by Proposition 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 [27,28] and the references therein. Here we consider $\phi$-divergence ball and Kantorovich ball approaches and discuss tractable formulations of the corresponding (DRSRP').

### 3.1 Ambiguity set constructed through $\phi$-divergence

Let us now consider the case that the only available information about the random vector $\xi$ 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 multi-stage decision-making problems), then the quality of such approximation may be compromised. $\phi$-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 $\phi$-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 $\phi$-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) $\chi^2$-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 $\chi^2$-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 1** (Relationships between $\phi$-divergences) *For two probability vectors $p, q \in \mathbb{R}_+^M$, the following inequalities hold.*

(i) $I_{\phi_V}(p, q) \leq \min\left(\sqrt{2I_{\phi_{KL}}(p, q)}, \sqrt{2I_{\phi_B}(p, q)}, \sqrt{I_{\phi_J}(p, q)}, \sqrt{I_{\phi_{\chi^2}}(p, q)},\right.$

$\left.\sqrt{I_{\phi_{m\chi^2}}(p, q)}\right);$

(ii) $I_{\phi_H}(p, q) \leq I_{\phi_V}(p, q) \leq 2\sqrt{I_{\phi_H}(p, q)}.$

We omit the proof as the results can be easily derived by the divergence functions $\phi$.

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

$$P_N(\cdot) := \sum_{i=1}^M \frac{N_i}{N} \mathbb{1}_{\zeta^i}(\cdot), \tag{9}$$

and ambiguity set

$$\mathcal{P}_N^M := \left\{ \sum_{i=1}^M p_i \mathbb{1}_{\zeta^i}(\cdot) : I_\phi(p, p_N) \leq r, \sum_{i=1}^M p_i = 1, p_i \geq 0, \forall i = 1, \ldots, M \right\}, \tag{10}$$

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:

$$
\begin{aligned}
\min_{x \in X, t \in T, \tau, u} \quad & t \\
\text{s.t.} \quad & \tau + ru + u\sum_{i=1}^M [p_N]_i \phi^*(s_i) \leq \lambda, \\
& s_i \leq \lim_{t \to \infty} \frac{\phi(t)}{t}, \ i = 1, \ldots, M, \\
& s_i = [l(c(x, \zeta^i) - t) - \tau]/u, \ i = 1, \ldots, M, \\
& u \geq 0,
\end{aligned}
\tag{11}
$$

where $p_N$ is defined as in (10) and we write $[p_N]_i$ for the $i$th component of $p_N$, $\phi^*$ denotes the Fenchel conjugate of $\phi$, i.e., $\phi^*(s) = \sup_{t \geq 0}\{st - \phi(t)\}$, see similar formulation in [4]. Note that $u\sum_{i=1}^M [p_N]_i \phi^*([l(c(x, \zeta^i) - t) - \tau]/u)$ is a convex function of $x, u, \tau$ and $t$, see [19]. Thus, problem (11) is a convex program.

It is important to note that the reformulation (11) 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 (11) 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 SAA approach to deal with the expected value.

The reallocation of the probabilities through Voronoi partition provides an effective way to reduce the scenarios of the discretized problem and hence the size of problem (11). It remains to be explained how the ambiguity set approximates the true probability distribution.

Let $\mathscr{L}$ denote the set of functions $h : \Xi \to \mathbb{R}$ satisfying $|h(\xi_1) - h(\xi_2)| \leq \|\xi_1 - \xi_2\|$, 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 the deviation of a set of probability measures $\mathcal{P}$ from another set of probability measures $\mathcal{Q}$ by $\mathbb{D}_K(\mathcal{P}, \mathcal{Q}) := \sup_{P \in \mathcal{P}} \inf_{Q \in \mathcal{Q}} \mathsf{dl}_K(P, Q)$, and the Hausdorff distance between the two sets by $\mathbb{H}_K(\mathcal{P}, \mathcal{Q}) := \max \{\mathbb{D}_K(\mathcal{P}, \mathcal{Q}), \mathbb{D}_K(\mathcal{Q}, \mathcal{P})\}$. An important property of the Kantorovich metric is that it metrizes weak convergence of probability measures [5] 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 $\Xi$ is defined as $M$ subsets of $\Xi$, denoted by $\Xi_1, \ldots, \Xi_M$, with $\bigcup_{i=1,\ldots,M} \Xi_i = \Xi$ and $\Xi_i \subseteq \left\{ y \in \Xi : \|y - \zeta^i\| = \min_{j=1,\ldots,M} \|y - \zeta^j\| \right\}$. By [22, Lemma 4.9],

$$\mathsf{dl}_K \left( \sum_{i=1}^M P^*(\Xi_i) \mathbb{1}_{\zeta^i}(\cdot), P^* \right) = \int \min_{1 \leq i \leq M} d(\xi, \zeta^i) dP^*$$

$$= \sum_{i=1}^M \int_{\Xi_i} d(\xi, \zeta^i) dP^* \leq \beta_M, \qquad (12)$$

where

$$\beta_M := \max_{\xi \in \Xi} \min_{1 \leq i \leq M} d(\xi, \zeta^i). \qquad (13)$$

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

**Proposition 3** *Let $\mathcal{P}_N^M$ be defined as in (10) and $P^*$ be the true probability distribution of $\xi$. Let $\beta_M$ be defined as in (13) and $\delta$ be a positive number such that $M\delta < 1$. If $\phi$ is chosen from one of the functions listed in (a)-(g) preceding Lemma 1, then with probability at least $1 - M\delta$,*

$$\mathbb{H}_K(\mathcal{P}_N^M, P^*) \leq \beta_M + \frac{D}{2} \max\{2\sqrt{r}, r\} + \frac{D}{2} \Delta(M, N, \delta), \qquad (14)$$

*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 $\Xi$, that is, $\sup\{\|\xi' - \xi''\| : \xi', \xi'' \in \Xi\}$, and $r$ is defined as in ([10]). In the case when $\xi$ follows a discrete distribution with support $\{\zeta^1, \ldots, \zeta^M\}$, we have*

$$\mathbb{H}_K(\mathcal{P}_N^M, P^*) \leq \frac{D}{2}\max\{2\sqrt{r}, r\} + \frac{D}{2}\Delta(M, N, \delta) \tag{15}$$

*with probability at least $1 - M\delta$.*

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

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

By ([12]), $\mathsf{dl}_K\left(\sum_{i=1}^M P^*(\Xi_i)\mathbb{1}_{\zeta^i}(\cdot), P^*\right) \leq \beta_M$. Moreover, it follows by [14, Theorem 4], the Kantorovich distance is bounded by $D/2$ times the total variation distance, that is,

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

Observe that

$$\sum_{i=1}^M |p_i - P^*(\Xi_i)| \leq \sum_{i=1}^M \left(|p_i - [p_N]_i| + |[p_N]_i - P^*(\Xi_i)|\right)$$

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

By Lemma [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 ([14]), it suffices to show

$$\sum_{i=1}^M |[p_N]_i - P^*(\Xi_i)| \leq \Delta(M, N, \delta). \tag{16}$$

Let $a \in \mathbb{R}^M$ be a vector with $\|a\|_\infty := \max_{1 \le i \le M} |a_i| = 1$, and $\phi_a(\xi) := \sum_{i=1}^M a_i \mathbb{1}_{\Xi_i}(\xi)$. Then $\sup_{\xi \in \Xi} |\phi_a(\xi)| \le 1$ and it follows by [25, Theorem 3] that

$$\left| \frac{1}{N} \sum_{k=1}^N \phi_a(\xi^k) - \mathbb{E}_{P^*}[\phi_a(\xi)] \right| \le \frac{1}{\sqrt{N}} \left( 2 + \sqrt{2 \ln \frac{1}{\delta}} \right) \tag{17}$$

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

$$\left| [p_N]_i - P^*(\Xi_i) \right| = \left| \frac{1}{N} \sum_{k=1}^N \phi_{e_i}(\xi^k) - \mathbb{E}_{P^*}[\phi_{e_i}(\xi)] \right| \le \frac{1}{\sqrt{N}} \left( 2 + \sqrt{2 \ln \frac{1}{\delta}} \right) \tag{18}$$

with probability at least $1 - \delta$ for each $i = 1, \cdots, M$. By Bouferroni's inequality,

$$\sum_{i=1}^M \left| [p_N]_i - P^*(\Xi_i) \right| \le \frac{M}{\sqrt{N}} \left( 2 + \sqrt{2 \ln \frac{1}{\delta}} \right) \tag{19}$$

with probability at least $1 - M\delta$ and hence we have shown (16) 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 (16). Let $A := \{a \in \mathbb{R}^M : \|a\|_\infty = 1\}$. For a small positive number $v$ (less or equal to 2), let $A_k := \{a^1, \cdots, a^k\}$ be such that for any $a \in A$, there exists a point $a^i(a) \in A_k$ depending on $a$ such that $\|a - a^i(a)\|_\infty \le v$, i.e., $A_k = \{a^1, \cdots, a^k\}$ is a $v$-net of $A$. Observe that

$$\sum_{i=1}^M \left| [p_N]_i - P^*(\Xi_i) \right| = \sup_{\|a\|_\infty = 1} |p_N^T a - p^{*T} a|, \tag{20}$$

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

$$|p_N^T a - p^{*T} a| \le |p_N^T(a - a^i(a))| + |p_N^T a^i(a) - p^{*T} a^i(a)| + |p^{*T} a^i(a) - p^{*T} a|$$
$$\le 2v + |p_N^T a^i(a) - p^{*T} a^i(a)|.$$

By (17), for each $a^i, i = 1, \cdots, k$

$$|p_N^T a^i - p^{*T} a^i| \le \frac{1}{\sqrt{N}} \left( 2 + \sqrt{2 \ln \frac{1}{\delta}} \right) \tag{21}$$

with probability at least $1-\delta$, thus inequality (21) holds uniformly for all $i = 1, \cdots, k$ with probability at least $1 - k\delta$. This enables us to conclude that

$$\sum_{i=1}^{M} \left| [p_N]_i - P^*(\varXi_i) \right| = \sup_{\|a\|_\infty=1} |p_N^T a - p^{*T} a|$$

$$\leq 2\nu + \frac{1}{\sqrt{N}} \left( 2 + \sqrt{2 \ln \frac{1}{\delta}} \right) \qquad (22)$$

with probability at least $1 - k\delta$. Since when $\nu = 2$, $A_k$ will be a trivial $\nu$-net, then we can set $k = M$ and obtain from (22) that

$$\sum_{i=1}^{M} \left| [p_N]_i - P^*(\varXi_i) \right| \leq 4 + \frac{1}{\sqrt{N}} \left( 2 + \sqrt{2 \ln \frac{1}{\delta}} \right) \qquad (23)$$

with probability at least $1 - M\delta$. This completes the proof of (16) and hence inequality (14).

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

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

The rest follows from similar analysis for the proof of (14). □

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 (14) is

$$\frac{D}{2} \min \left( \frac{M}{\sqrt{N}} \left( 2 + \sqrt{2 \ln(10M)} \right), 4 + \frac{1}{\sqrt{N}} \left( 2 + \sqrt{2 \ln(10M)} \right) \right). \qquad (24)$$

In order for the first part of (24) 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 $\varXi$, 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 (24) 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 [20,27]. 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 [22, Algorithms 4.1–4.5]. Forth, the inequality (14) 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 [21] showed that the estimated $\phi$-divergence $\frac{2N}{\phi''(1)} I_\phi(p^*, p_N)$ asymptotically follows a $\chi_{M-1}^2$-distribution with $M - 1$ degrees of freedom, where $p^*$ denotes the probability vector corresponding to probability measure $P^*$ and $M$ is the cardinality of $\Xi$ (the support of $P^*$), which means if we set

$$r := \frac{\phi''(1)}{2N} \chi_{M-1,1-\delta}^2, \tag{25}$$

then with probability $1 - \delta$, $I_\phi(p^*, p_N) \leq r$. The latter indicates that the ambiguity set (10) lies in the $1 - \delta$ confidence region.

For general $\phi$-divergences, we are unable to establish the quantitative convergence as in Proposition 3. However, if $P^*$ follows a discrete distribution with support $\{\zeta^1, \ldots, \zeta^M\}$, the following qualitative convergence result holds.

**Proposition 4** [19, 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 $\mathbb{H}_K(\mathcal{P}_N^M, P^*) \to 0$, w.p.1, as $N \to \infty$, where $r$ is defined as in (25).*

Note that in [19, Proposition 2] the convergence is established under the total variation metric, since the probability distributions here are discrete, the convergence is equivalent to that under the Kantorovich metric. We refer readers to [19] for the details of the proof.

### 3.2 Kantorovich ball

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

$$\mathcal{P}_N = \{P \in \mathscr{P}(\Xi) : \mathsf{dl}_K(P, P_N) \leq r\}, \tag{26}$$

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

$$\theta := \int_\Xi \exp(\|\xi\|^a) P^*(d\xi) < \infty, \tag{27}$$

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

$$\text{Prob}(\mathsf{dl}_K(P^*, P_N) \geq r) \leq \begin{cases} C_1 \exp(-C_2 N r^{\max\{k,2\}}) & \text{if } r \leq 1, \\ \exp(-C_2 N r^a) & \text{if } r > 1, \end{cases} \tag{28}$$

for all $N \geq 1$, $k \neq 2$, where $C_1$ and $C_2$ are positive constants only depending on $a$, $\theta$ 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 $\xi$, see [11] for details.

By setting the right hand side of the above inequality to $\delta$ and solving for $r$, we may set

$$r_N(\delta) := \begin{cases} \left(\frac{\log(C_1\delta^{-1})}{C_2 N}\right)^{1/\max\{k,2\}} & \text{if } N \geq \frac{\log(C_1\delta^{-1})}{C_2}, \\ \left(\frac{\log(C_1\delta^{-1})}{C_2 N}\right)^{1/a} & \text{if } N < \frac{\log(C_1\delta^{-1})}{C_2}, \end{cases} \tag{29}$$

and consequently the ambiguity set (26) contains the true probability distribution $P^*$ with probability $1 - \delta$ when $r = r_N(\delta)$.

In [8,13,29], the dual formulation of distributionally robust optimization problem with the ambiguity set (26) has been established. Based on these results, the dual of (DRSRP') can be written as

$$\begin{aligned} \min_{x \in X, t \in T, \eta, s} \quad & t \\ \text{s.t.} \quad & \eta r + \frac{1}{N}\sum_{i=1}^{N} s_i \leq \lambda, \\ & \sup_{\xi \in \Xi}\left[l(c(x,\xi) - t) - \eta\|\xi - \xi^i\|\right] \leq s_i, i = 1, \ldots, N. \end{aligned} \tag{30}$$

In the case when $c(x,\xi) = -x^T\xi$, $\Xi = \{\xi \in \mathbb{R}^k : G\xi \leq d\}$ 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 (30) can be recast as

$$\begin{aligned} \min_{x \in X, t \in T, \eta, s, \gamma_{ij}} \quad & t \\ \text{s.t.} \quad & \eta r + \frac{1}{N}\sum_{i=1}^{N} s_i \leq \lambda, \\ & b_j - a_j t - \langle a_j x, \xi_i\rangle + \langle\gamma_{ij}, d - G\xi^i\rangle \leq s_i, \\ & i = 1, \ldots, N, j = 1, \ldots, K, \\ & \|G^T\gamma_{ij} + a_j x\|_* \leq \eta, i = 1, \ldots, N, j = 1, \ldots, K, \\ & \gamma_{ij} \geq 0, i = 1, \ldots, N, j = 1, \ldots, K. \end{aligned}$$

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

**Proposition 5** *Let $\mathcal{P}_N$ be defined as in* (26) *and $P^*$ denote the true probability distribution. Let $r_N(\delta)$ be defined as in* (29). *If the radius of the Kantorovich ball in* (26) *is equal to $r_N(\delta)$, then with probability at least $1 - \delta$,*

$$\mathbb{H}_K(\mathcal{P}_N, P^*) \leq 2r_N(\delta). \tag{31}$$

*Proof* We first prove that

$$\mathbb{H}_K(\mathcal{P}_N, P^*) \leq \mathsf{dl}_K(P_N, P^*) + r. \tag{32}$$

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

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

which implies $\mathbb{D}_K(\mathcal{P}_N, P^*) \leq r + \mathsf{dl}_K(P_N, P^*)$. On the other hand,

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

A combination of the last two inequalities yields (32).

Let us now estimate the first term in (32), i.e., $\mathsf{dl}_K(P_N, P^*)$. By the definition of $r_N(\delta)$, we have with probability $1 - \delta$, $\mathsf{dl}_K(P_N, P^*) \leq r_N(\delta)$. The conclusion follows. $\square$

In the case when the centre of the Kantorovich ball $P_N$ in (26) is replaced by that defined as in (9), we have

$$\mathbb{H}_K(\mathcal{P}_N, P^*) \leq \beta_M + r + \frac{D}{2}\Delta(M, N, \delta) \tag{33}$$

with probability at least $1 - M\delta$, where $\Delta(M, N, \delta)$ is defined as in Proposition 3. To see this, we can use the triangle inequality of the Hausdorff distance with the Kantorovich metric to derive

$$\mathbb{H}_K(\mathcal{P}_N, P^*)$$
$$\leq \sup_{P \in \mathcal{P}_N} \mathsf{dl}_K\left(P, \sum_{j=1}^{M} P^*(\varXi_j)\mathbb{1}_{\zeta^j}(\cdot)\right) + \mathsf{dl}_K\left(\sum_{j=1}^{M} P^*(\varXi_j)\mathbb{1}_{\zeta^j}(\cdot), P^*\right). \tag{34}$$

Since

$$\mathsf{dl}_K\left(P, \sum_{j=1}^{M} P^*(\varXi_j)\mathbb{1}_{\zeta^j}(\cdot)\right) \leq \mathsf{dl}_K\left(P, P_N\right) + \mathsf{dl}_K\left(P_N, \sum_{j=1}^{M} P^*(\varXi_j)\mathbb{1}_{\zeta^j}(\cdot)\right)$$
$$\leq r + \frac{D}{2}\sum_{j=1}^{M}\left|[p_N]_j - P^*(\varXi_j)\right|, \tag{35}$$

we establish (33) by combining (34), (35), (12), (19) and (23).

Before concluding this section, we note that it is possible to use other statistical methods for constructing the ambiguity sets such as moment conditions and mixture distribution, we omit them due to limitation of the length of the paper, interested readers may find them in [16] and references therein.

## 4 Convergence of (DRSRP')

In Sect. 3, we discussed two approaches for constructing the ambiguity set of the (DRSRP') model, each of which is defined through iid samples.

Let us rewrite the model with $\mathcal{P}$ being replaced by $\mathcal{P}_N$:

$$
\text{(DRSRP'-N)} \quad
\begin{cases}
\min\limits_{X \in X, t \in T} t \\
\text{s.t.} \quad \sup\limits_{P \in \mathcal{P}_N} \mathbb{E}_P[l(c(x,\xi) - t)] \leq \lambda,
\end{cases}
\tag{36}
$$

to explicitly indicate the dependence of the samples. In this section, we 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 as the sample size increases.

Let $x_N$ be a solution of distributionally robust shortfall risk minimization problem (DRSRP'N), let $\vartheta_N$ be the optimal value and $S_N$ the corresponding optimal solution set. The out-of-sample performance of $x_N$ is defined as $\mathrm{SR}_{l,\lambda}^{P^*}(-c(x_N, \xi))$, where $P^*$ is the true probability distribution. Since $P^*$ is unknown, the exact out-of-sample performance of $x_N$ cannot be computed, but we may seek its upper bound $\vartheta_N$ such that

$$
\mathrm{Prob}(\mathrm{SR}_{l,\lambda}^{P^*}(-c(x_N, \xi)) \leq \vartheta_N) \geq 1 - \delta,
\tag{37}
$$

where $\delta \in (0, 1)$. Following the terminology of Esfahani and Kuhn [8], we call $\delta$ a significance parameter and $\vartheta_N$ the certificate for the out-of-sample performance. The probability on the left-hand side of (37) indicates $\vartheta_N$'s reliability. The following theorem states that the finite sample guarantee condition is fulfilled for the ambiguity sets discussed in Sect. 3, that is, when the size of the ambiguity sets are chosen carefully, the certificate $\vartheta_N$ can provide a $1 - \delta$ confidence bound of the type (37) on the out-of-sample performance of $x_N$.

**Theorem 1** (Finite sample guarantee) *The following assertions hold:*

(i) *Suppose the true probability distribution $P^*$ is discrete, i.e., $\Xi = \{\zeta^1, \ldots, \zeta^M\}$. Let $\mathcal{P}_N^M$ be defined as (10) with $r$ being given as (25), then with $\mathcal{P}_N = \mathcal{P}_N^M$, the finite sample guarantee (37) holds.*

(ii) *Let $\mathcal{P}_N$ be defined as in (26) with $r = r_N(\delta)$ being given in (29). Under condition (27), the finite sample guarantee (37) holds.*

*Proof* The results follow straightforwardly from (25), (28), (29) and the definition of finite sample guarantee. □

We now move on to investigate convergence of $\vartheta_N$ and $S_N$. From the discussion in Sect. 3, we know that $\mathbb{H}_K(\mathcal{P}_N, P^*) \to 0$. However, to broaden the coverage of the convergence results, we present them by considering a slightly more general case with $P^*$ being replaced by a set $\mathcal{P}^*$.

**Theorem 2** (Convergence of the optimal values and optimal solutions) *Let $\mathcal{P}^* \subset \mathcal{P}(\Xi)$ be such that $\lim_{N\to\infty} \mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*) = 0$. Let $\vartheta^*$ denote the optimal value of (DRSRP') with $\mathcal{P}$ being replaced by $\mathcal{P}^*$. Let $S^*$ be the corresponding optimal solutions. Under Assumption 1,*

$$|\vartheta_N - \vartheta^*| \leq \frac{2D_X L\kappa}{\eta} \mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*) \tag{38}$$

*for N sufficiently large and*

$$\limsup_{N\to\infty} S_N = S^*, \tag{39}$$

*where $D_X$ denotes the diameter of X, $\eta$ is defined as in Proposition 2, and $L, \kappa$ are defined as in Assumption 1.*

*Proof* Let $v^*(x, t) := \sup_{P\in\mathcal{P}^*} \mathbb{E}_P[l(c(x, \xi) - t)] - \lambda$ and $v_N(x, t) := \sup_{P\in\mathcal{P}_N} \mathbb{E}_P[l(c(x, \xi) - t)] - \lambda$. Let $g(x, t, \xi) := l(c(x, \xi) - t)$. By the definition,

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

where the inequality is due to equi-Lipschitz continuity of $g$ in $\xi$ 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

$$\sup_{x\in X, t\in T} |v_N(x, t) - v^*(x, t)| \leq L\kappa \mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*). \tag{40}$$

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, $v^*$ and $v_N$ are convex on $X \times T$. Moreover, the Slater condition (7) allows us to apply Robinson's error bound for the convex inequality system (see [23]), 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

$$
\begin{aligned}
d((x, t), \mathcal{F}^*) &\leq C_1[v^*(x, t) - \lambda]_+ \\
&\leq C_1(|v^*(x, t) - v_N(x, t)| + [v_N(x, t) - \lambda]_+) \\
&= C_1|v^*(x, t) - v_N(x, t)| \leq C_1 L\kappa \mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*). \tag{41}
\end{aligned}
$$

The last inequality follows from (40) and Robinson's error bound [23] ensures that the constant $C_1$ is bounded by $D_X/\eta$, 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

$$d((x, t), \mathcal{F}_N) \leq C_2 |v^*(x, t) - v_N(x, t)| \leq C_2 L \kappa \mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*) \tag{42}$$

for $(x, t) \in \mathcal{F}^*$ and $N$ is sufficiently large, where $C_2$ is bounded by $2D_X/\eta$. Combining (41) and (42), we obtain

$$\mathbb{H}(\mathcal{F}_N, \mathcal{F}^*) \leq \frac{2D_X L \kappa}{\eta} \mathbb{H}_K(\mathcal{P}_N, \mathcal{P}^*). \tag{43}$$

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 (38) via (43).

Now, we move on to show (39). 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 (43) and (38) that $(\hat{x}, \hat{t}) \in \mathcal{F}^*$ and $\hat{t} = \vartheta^*$. This shows $(\hat{x}, \hat{t}) \in S^*$. □

Theorem 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 Sect. 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 $\vartheta^*$ in the following corollary.

**Corollary 1** *Under the assumptions in Theorem 2, the following assertions hold.*

(i) *If $\mathcal{P}^*$ comprises the true probability distribution only and $\mathcal{P}_N$ is defined by (10), then under conditions of Proposition 3, $\vartheta^* \in [\vartheta_N - \Theta, \vartheta_N + \Theta]$ with probability $1 - M\delta$, where*

$$\Theta := \frac{2D_X L \kappa}{\eta} [\beta_M + \frac{D}{2} (\max\{2\sqrt{r}, r\} + \Delta(M, N, \delta))]$$

*with* $\Delta(M, N, \delta) = \min\left(\frac{M}{\sqrt{N}}\left(2 + \sqrt{2\ln\frac{1}{\delta}}\right), 4 + \frac{1}{\sqrt{N}}\left(2 + \sqrt{2\ln\frac{1}{\delta}}\right)\right)$, $\beta_M$ *being defined as in* (13) *and D being the diameter of* $\Xi$.

(ii) *If* $\mathcal{P}^*$ *comprises the true probability distribution only and* $\mathcal{P}_N$ *is defined by* (26), *then under conditions of Proposition* 5,

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

*with probability* $1 - \delta$.

## 4.1 Extension

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

$$\text{(DRSRCP)} \quad \begin{array}{l} \min_{x \in X} f(x) \\ \text{s.t.} \ \ \text{SR}^{\mathcal{P}}_{l,\lambda}(-c(x, \xi)) \leq \gamma, \end{array} \tag{44}$$

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

$$\text{(DRSRCP-N)} \quad \begin{array}{l} \min_{x \in X} f(x) \\ \text{s.t.} \ \ \text{SR}^{\mathcal{P}_N}_{l,\lambda}(-c(x, \xi)) \leq \gamma. \end{array} \tag{45}$$

Tractable reformulation of problem (DRSRCP) or (DRSRCP-N) may be derived as we did in Sect. 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 3** *Let Assumption* 1 *hold. Suppose that there exists* $x_0 \in X$ *such that*

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

*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) $\lim_{N \to \infty} \hat{\vartheta}_N = \hat{\vartheta}$ *and* $\limsup_{N \to \infty} \hat{S}_N = \hat{S}$.

(iii) *If, in addition, $f$ is Lipschitz continuous with modulus $\beta$, then*

$$|\hat{\vartheta}_N - \hat{\vartheta}| \leq \beta \mathbb{H}(\hat{\mathcal{F}}_N, \hat{\mathcal{F}}). \tag{46}$$

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

$$\mathbb{D}(\hat{S}_N, \hat{S}) \leq \max \left\{ 2C, \sqrt{8C\beta/\alpha} \right\} \sqrt{\mathbb{H}_K(\mathcal{P}_N, \mathcal{P})} \tag{47}$$

*when $N$ is sufficiently large.*

*Proof* Part (i) can be established through an analogous proof of Theorem 2. We omit the details.

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

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

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 [24, Definition 7.39] that $\delta_{\hat{\mathcal{F}}_N}(\cdot)$ epiconverges to $\delta_{\hat{\mathcal{F}}}(\cdot)$. Furthermore, it follows from [24, 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 sets, then any sequence $\{x_N\}$ in $\hat{S}_N$ has a subsequence converging to $\overline{x}$. By [6, Proposition 4.6], $\lim_{N \to \infty} \hat{\vartheta}_N = \hat{\vartheta}$ and $\overline{x} \in \hat{S}$.

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 $x'_N \in \hat{\mathcal{F}}$ such that $d(x_N, x'_N) \leq \mathbb{D}(\hat{\mathcal{F}}_N, \hat{\mathcal{F}})$. Moreover, by the Lipschitz continuity of $f$, we have

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

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 (46).

Next, we show (47). Let $x_N \in \hat{S}_N$ and $\overline{x} \in \hat{S}$. By the second order growth condition,

$$f(x_N) - f(\Pi_{\hat{\mathcal{F}}}(x_N)) = f(x_N) - f(\overline{x}) - (f(\Pi_{\hat{\mathcal{F}}}(x_N)) - f(\overline{x}))$$

$$\leq f(\Pi_{\hat{\mathcal{F}}_N}(\overline{x})) - f(\overline{x}) - \alpha d(\Pi_{\hat{\mathcal{F}}}(x_N), \hat{S})^2,$$

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

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

Therefore,

$$\begin{aligned}
d(x_N, \hat{S}) &\leq \|x_N - \Pi_{\hat{\mathcal{F}}}(x_N)\| + d(\Pi_{\hat{\mathcal{F}}}(x_N), \hat{S}) \\
&\leq \|x_N - \Pi_{\hat{\mathcal{F}}}(x_N)\| + \sqrt{(\beta/\alpha)(\|\Pi_{\hat{\mathcal{F}}_N}(\overline{x}) - \overline{x}\| + \|\Pi_{\hat{\mathcal{F}}}(x_N) - x_N\|)}.
\end{aligned} \tag{48}$$

Since $\max\left\{\max_{x_N \in \hat{S}_N} \|x_N - \Pi_{\hat{\mathcal{F}}}(x_N)\|, \max_{\overline{x} \in \hat{S}} \|\Pi_{\hat{\mathcal{F}}_N}(\overline{x}) - \overline{x}\|\right\} \leq \mathbb{H}(\hat{\mathcal{F}}_N, \hat{\mathcal{F}})$, we have from inequality (48) 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 (47) 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. □

Analogous to Corollary 1, we can derive confidence intervals and regions for the optimal values 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 $\xi_i$ 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 $\xi$ 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 risk of $x^T \xi$, that is, $\text{SR}_{l,\lambda}^{\mathcal{P}}(x^T \xi)$ for some specified $l$, $\lambda$ and $\mathcal{P}$, that is, the investor finds an optimal decision $x^*$ by solving

$$\begin{aligned}
\min_{x \in X, t \in \mathbb{R}} \quad & t \\
\text{s.t.} \quad & \sup_{P \in \mathcal{P}} \mathbb{E}_P[l(-x^T \xi - t)] \leq \lambda.
\end{aligned} \tag{49}$$

We have undertaken numerical experiments on problem (49) from different perspectives ranging from efficiency of computational schemes as we discussed in Sect. 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.

Our main numerical experiments focus on problem (49) with the ambiguity set being defined through the Kantorovich ball. We report the details in Example 1.

*Example 1* Let $\xi^1, \ldots, \xi^N$ be iid samples of $\xi$ 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

$$\mathcal{P}_N = \{P \in \mathscr{P}(\mathbb{R}^d) : \mathsf{dl}_K(P, P_N) \leq r\}. \tag{50}$$

To simplify the tests, we consider a specific piecewise affine loss function $l(z) = \max\{0.05z + 1, z + 0.1, 4z + 2\}$. 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 $\xi_i$ 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 discussion in Sect. 3.2, problem (49) can be reformulated through dual formulation as

$$
\begin{aligned}
J_N(r) := \min_{x \in X, t \in \mathbb{R}, \eta, s} \quad & t \\
\text{s.t.} \quad & \eta r + \frac{1}{N} \sum_{i=1}^{N} s_i \leq 1, \\
& b_j - a_j t - \langle a_j x, \xi^i \rangle \leq s_i, \text{ for } i = 1, \ldots, N, j = 1, 2, 3, \\
& \|a_j x\|_* \leq \eta, \text{ for } j = 1, 2, 3.
\end{aligned}
\tag{51}
$$

We use $\|\cdot\|$ to denote 1-norm and thus $\|\cdot\|_*$ is the $\infty$-norm. 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 (51), 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 $2 \times 10^5$ 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 (51) and then use the same validation samples to evaluate $J(x_N(r))$. Each of the experiments is carried out through 200 simulation runs.

Figure 1 depicts the tubes between the 20 and 80% quantiles (shaded areas) and the means (solid lines) of the out-of-sample performance $J(x_N(r))$ as a function of radius $r$, the 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).
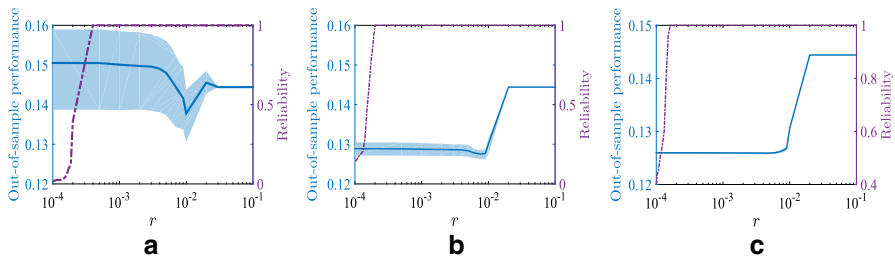
**Fig. 1** Out-of-sample performance $J(x_N(r))$ (left axis, solid line and shade area) and reliability Prob$(J(x_N(r)) \leq J_N(r))$ (right axis and dashed line) based on 200 independent runs. **a** $N = 30$ training samples, **b** $N = 300$ training samples, **c** $N = 3000$ training samples
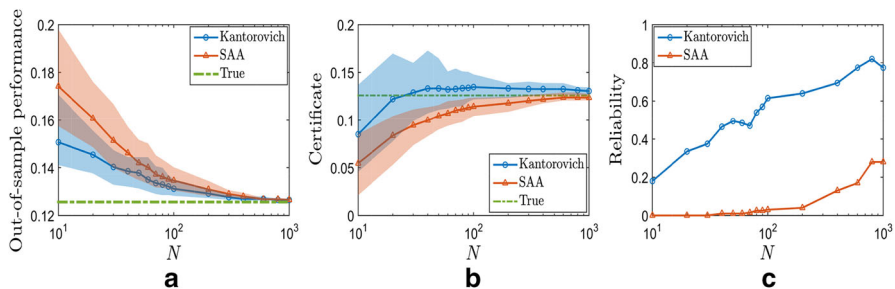


**Fig. 2** **a** Out-of-sample performance $J(x_N)$, **b** certificate $J_N$, and **c** reliability Prob$(J(x_N) \leq J_N)$ for the Kantorovich and SAA solutions of $N$

In the second set of experiments, we investigate convergence of the out-of-sample performance, the certificate and the reliability of the DRO approach (51) and the SAA approach 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 2a 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 (51) 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 DRO model (51) outperforms the SAA model in terms of out-of-sample performance. Figure 2b 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 2c 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

**Fig. 3** Wealth evolution with the trading times
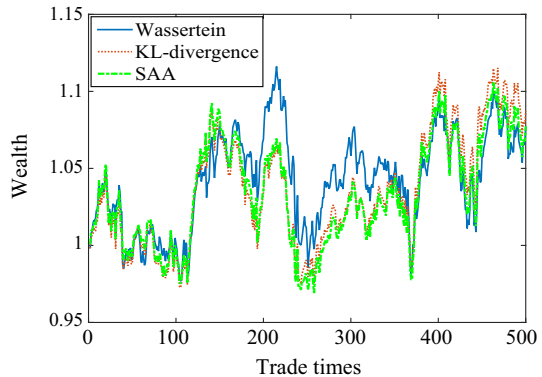


corresponding problems. It is clear that the performance of the DRO model is better than that of the SAA model.

*Example 2* In the last experiment, we evaluate the performance of problem (49) with the ambiguity set being constructed through the KL-divergence ball and the Kantorovich ball, we have also undertaken tests on problem (49) 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., QUAL-COMM 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 3 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.

## References

1. Artzner, P., Delbaen, F., Eber, J.M., Health, D.: 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. Bellini, F., Bignozzi, V.: On elicitable risk measures. Quant. Finance **15**, 725–733 (2015)
4. Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. Manag. Sci. **59**, 341–357 (2013)
5. Billingsley, P.: Convergence of Probability Measures. Wiley, New York (1968)
6. Bonnans, J.F., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer, New York (2000)
7. Dunkel, J., Weber, S.: Stochastic root finding and efficient estimation of convex risk measures. Oper. Res. **58**, 1505–1521 (2010)
8. Esfahani, P.M., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Math. Program. (2017). https://doi.org/10.1007/s10107-017-1172-1
9. Föllmer, H., Schied, A.: Convex measures of risk and trading constraints. Finance Stochast. **6**, 429–447 (2002)
10. Föllmer, H., Schied, A.: Stochastic Finance-An Introduction in Discrete Time. Walter de Gruyter, Berlin (2011)
11. Fournier, N., Guillin, A.: On the rate of convergence in Wasserstein distance of the empirical measure. Probab. Theory Relat. Fields **162**, 707–738 (2015)
12. Frittelli, M., Rosazza Gianin, E.: Putting order in risk measures. J. Bank. Finance **26**, 1473–1486 (2002)
13. Gao, R., Kleywegt, A.J.: Distributionally robust stochastic optimization with Wasserstein distance (2016). arXiv preprint arXiv:1604.02199
14. Gibbs, A.L., Su, F.E.: On choosing and bounding probability metrics. Int. Stat. Rev. **70**, 419–435 (2002)
15. Giesecke, K., Schmidt, T., Weber, S.: Measuring the risk of large losses. J. Invest. Manag. **6**, 1–15 (2008)
16. Guo, S., Xu, H.: Distributionally Robust Shortfall Risk Optimization Model and Its Approximation (2018). http://www.personal.soton.ac.uk/hx/research/Published/Manuscript/2018/Shaoyan/DRSR-20-Feb_2018_online.pdf
17. Heath, D.: Back to the future. Plenary Lecture at the First World Congress of the Bachelier Society, Paris (2000)
18. Hu, Z., Zhang, D.: Convex risk measures: efficient computations via Monte Carlo (2016). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2758713
19. Love, D., Bayraksan, G.: Phi-divergence constrained ambiguous stochastic programs for data-driven optimization, available on Optimization Online (2016)
20. Moulton, J.: Robust fragmentation: a data-driven approach to decision-making under distributional ambiguity. Ph.D. Dissertation, University of Minnesota (2016)
21. Pardo, L.: Statistical Inference Based on Divergence Measures. Chapman and Hall/CRC, Boca Raton (2005)
22. Pflug, G.C., Pichler, A.: Multistage Stochastic Optimization. Springer, Cham (2014)
23. Robinson, S.M.: An application of error bounds for convex programming in a linear space. SIAM J. Control **13**, 271–273 (1975)
24. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer, New York (1998)
25. Shawe-Taylor, J., Cristianini, N.: Estimating the moments of a random vector with applications. In: Proceedings of GRETSI 2003 Conference, pp. 47–52 (2003)
26. Weber, S.: Distribution-invariant risk measures, information, and dynamic consistency. Math. Finance **16**, 419–442 (2006)
27. Wiesemann, W., Kuhn, D., Sim, M.: Distributionally robust convex optimization. Oper. Res. **62**, 1358–1376 (2014)
28. Xu, H., Liu, Y., Sun, H.: Distributionally robust optimization with matrix moment constraints: lagrange duality and cutting-plane methods. Math. Program. (2017). https://doi.org/10.1007/s10107-017-1143-6
29. Zhao, C., Guan, Y.: Data-driven risk-averse stochastic optimization with Wasserstein metric. Available on Optimization Online (2015)