

Distributionally Robust Optimization with Confidence Bands for Probability Density Functions

Xi Chen^{*} Qihang Lin[†] Guanglin Xu[‡]

Abstract

Distributionally robust optimization (DRO) has been introduced for solving stochastic programs where the distribution of the random parameters is unknown and must be estimated by samples from that distribution. A key element of DRO is the construction of the ambiguity set, which is a set of distributions that covers the true distribution with a high probability. Assuming that the true distribution has a probability density function, we propose a class of ambiguity sets based on confidence bands of the true density function. The use of the confidence band enables us to take the prior knowledge of the shape of the underlying density function into consideration (e.g., unimodality or monotonicity). Using the confidence band constructed by density estimation techniques as the ambiguity set, we establish the convergence of the optimal value of DRO to that of the stochastic program as the sample size increases. However, the resulting DRO problem is computationally intractable, as it involves functional decision variables as well as infinitely many constraints. To address this challenge, using the duality theory, we reformulate it into a finite-dimensional stochastic program, which is amenable to a stochastic subgradient scheme as a solution method. We compare our approach with existing state-of-the-art DRO methods on the newsvendor problem and the portfolio management problem, and the numerical results showcase the advantage of our approach.

Keywords: Distributionally robust optimization; first-order method; confidence band; data-driven ambiguity sets

1 Introduction

The goal of *stochastic programming* (SP) is to minimize the expectation of an objective function that depends on both decision variables and some random parameters. Assuming that the random parameters follow a distribution denoted by P^* , a stochastic program can be formulated as follows:

$$v^* := \inf_{x \in \mathcal{X}} \left\{ \mathbb{E}_{P^*}[f(x, \xi)] := \int_{\mathbb{R}^m} f(x, \xi) P^*(d\xi) \right\}, \quad (1)$$

where $x \in \mathbb{R}^n$ is the vector of decision variables, $\mathcal{X} \subseteq \mathbb{R}^n$ is the feasible set, ξ is the vector of random variables taking values in \mathbb{R}^m with the distribution P^* , and $f(x, \xi) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ is the objective function. SP has been actively studied for several decades: see [5], [41] and references thereafter.

^{*}Stein School of Business, New York University, New York City, NY, 10012, USA. Email: xichen@nyu.edu

[†]Tippie College of Business, University of Iowa, Iowa City, IA 52245, USA. Email: qihang-lin@uiowa.edu

[‡]Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN, 55455, USA. Email: gxu@umn.edu.

Suppose ξ is *absolutely continuous* and thus has a probability density function p^* with respect to the Lebesgue measure. In other words, we have $p^* : \mathbb{R}^m \rightarrow [0, +\infty)$ such that $P^*(A) = \int_A p^*(\xi) d\xi$ for any Borel set $A \subset \mathbb{R}^m$. We can further write (1) as follows:

$$v^* = \inf_{x \in \mathcal{X}} \left\{ \mathbb{E}_{p^*}[f(x, \xi)] := \int_{\mathbb{R}^m} f(x, \xi) p^*(\xi) d\xi \right\}. \quad (2)$$

Despite of their popularity, (2) is often challenging to solve since the distribution P^* or the density p^* of ξ is rarely known in real-life applications. When a set of historical observations of ξ is collected, one may solve the approximation of (1) by replacing P^* with an estimated distribution from the dataset, e.g., the empirical distribution. However, due to the approximation error, the decision obtained from the approximate problem may be of inferior quality and thus may have an undesirable out-of-sample performance; see [2, 13, 29]. An alternative approach for solving (2) with an unknown distribution is *distributionally robust optimization* (DRO), in which one constructs an *ambiguity set* consisting of all distributions that are likely to be P^* and then minimizes the expectation of the objective function over the worst-case distribution from the ambiguity set. In particular, letting \mathcal{D} be the ambiguity set, we can formulate DRO as follows:

$$\inf_{x \in \mathcal{X}} \sup_{P \in \mathcal{D}} \left\{ \mathbb{E}_P[f(x, \xi)] := \int_{\mathbb{R}^m} f(x, \xi) P(d\xi) \right\}. \quad (3)$$

Scarf first proposed this model for a newsvendor problem [39], and it has been extensively studied over the past in operations research and operations management community since then.

The ambiguity sets constructed by most of the aforementioned approaches contain distributions that are not absolutely continuous. In fact, as shown in many of the works, the worst-case distribution in their ambiguity set corresponding to the optimal decision of (3) is discrete. However, in some applications, the vector of random parameter ξ is known to be absolutely continuous (e.g., when ξ models the price of electricity or the return of securities). In this case, by solving (3) with these ambiguity sets, one may obtain a solution that is hedging against a discrete distribution that will never be the true distribution. This phenomenon potentially leads to over-conservative decisions in the DRO problems.

In this paper, we consider the situation where P^* is absolutely continuous and propose a family of ambiguity sets \mathcal{D} consisting of only absolutely continuous distributions, or equivalently, distributions with density functions. With such an ambiguity set, the DRO model corresponding to (2) is given as:

$$\inf_{x \in \mathcal{X}} \sup_{p \in \mathcal{D}} \left\{ \mathbb{E}_p[f(x, \xi)] := \int_{\mathbb{R}^m} f(x, \xi) p(\xi) d\xi \right\}. \quad (4)$$

Thus, the worst-case distribution corresponding to the optimal solution of (4) will be absolutely continuous.

We note that the ambiguity set of density functions has also been considered by [30] and [8].

While their ambiguity sets contain only polynomial density functions, our approach does not require for similar restrictions. We also note that the ambiguity set considered in [30] utilizes the kernel density estimation, as does one of our ambiguity sets. While their method must specify the Legendre polynomial series density estimator, our method allows for using a broader family of kernel density estimations. Moreover, our ambiguity set is constructed with data samples and fully utilizes *shape information* of the density function p^* (e.g., unimodality or monotonicity), while the ambiguity set constructed in [8] is not based on a data-driven approach. Indeed, their ambiguity set is constructed using other prior knowledge on p^* (e.g., moment information).

Other works closely related to our method include [25, 27]. Li and Jiang consider an ambiguity set with moment and generalized unimodal constraints [27]. Lam imposes convexity constraints on the tail of the density function [25]. The main difference between our approach and theirs is that our methods use the shape information on p^* and a dataset generated by p^* to construct the ambiguity set, while theirs does not use data samples and directly impose shape information as constraints in their optimization problems.

In the rest of the paper, we first propose the generic DRO in Section 2, followed by the construction of our data-driven ambiguity sets by using density estimation techniques from the statistics literature. In particular, we will present two classes of ambiguity sets and showcase their convergence to the true density function and further prove the convergence of the optimal value of (4) to the optimal objective value of the SP in (2) as the sample size increases to infinity; see detail in Section 3. The setting of our ambiguity set gives rise to a challenging problem to solve, as the resulting optimization problem (4) involves functional decision variables (i.e., the density p) and infinitely many constraints. In Section 4, taking the exploitation of the special structure of our ambiguity set, we show that (4) can be reformulated into a finite-dimensional convex stochastic program, which is amenable to an efficient stochastic first-order method approach as the solution method. Finally, we validate our approach with a newsvendor problem and a portfolio management problem in Section 5. The numerical results demonstrate that our approach can generate decisions with superior out-of-sample performances, especially when the number of observations is limited.

1.1 Literature review

In the existing literature, different approaches have been utilized to construct ambiguity sets. We briefly review some popular approaches as follows:

- A moment-based ambiguity set is often constructed to consist of all distributions that share common marginal or cross moments; see [4, 6, 7, 9, 8, 11, 12, 16, 17, 27, 31, 43, 46, 47]. DRO problems with moment-based ambiguity sets in the format of (4) can usually be reformulated into tractable conic programs, e.g., second-order cone programming problems or semidefinite programming problems. However, the constructed ambiguity sets \mathcal{D} are not guaranteed to converge to the true distribution P^* , as the size of the historical observations increases to infinity, although the estimations of the moments of the random variables are guaranteed to converge to their true values.

- A distance-based ambiguity set is constructed by using some distance function to measure the distance between two distributions in the probability space. In fact, such an ambiguity set can be considered as a ball centered at a reference distribution, e.g., the empirical distribution, in the space of probability distributions. The distance functions considered in the literature include Kullback-Leibler divergence [20, 22], ϕ -divergence [1, 10, 24], Prohorov metric [12], empirical Burg-entropy divergence balls [25], and Wasserstein metric [34, 45, 13, 15, 14]. Many distance-based ambiguity sets have both asymptotic and finite-sample convergences. However, there is evidence that the resulting DRO problems have the tendency to be more challenging to solve compared to their counterparts with moment-based ambiguity sets; see [13].
- More recently, hypothesis-test-based ambiguity sets have been proposed; see [2, 3]. Based on a hypothesis test, e.g., χ^2 -test, G -test, etc., and a confidence level, these approaches construct ambiguity sets consisting of the distributions that pass the hypothesis test with a given set of historical data. The methods that we use in this paper belong to this category.
- Likelihood approaches are also considered to construct ambiguity sets in the literature; see, e.g., [10, 26, 35, 44]. The likelihood approaches construct ambiguity sets consisting of all distributions that make a set of observations to achieve a certain level of likelihood.

1.2 Notation and terminology

Let $\text{Proj}_{\mathcal{V}}(\cdot)$ denote the Euclidean projection operator on to the set \mathcal{V} , i.e., $\text{Proj}_{\mathcal{V}}(u) = \arg\min_{v \in \mathcal{V}} \|v - u\|_2^2$. Let $\mathbb{I}_E(\xi)$ be the indicator function that equals one when $\xi \in E$ and zero when $\xi \notin E$. Unless stated otherwise, the terms “almost every”, “measurable”, and “integrable” are defined with respect to the Lebesgue sense. For an extended-real valued function g on \mathbb{R}^n , let $\text{epi}(g)$ be its epigraph, $\text{dom}(g)$ be its domain, ∂g be its subdifferential, and $g' \in \partial g$ be its any subgradient. We call g a multifunction if it maps a point in \mathbb{R}^m to a subset of \mathbb{R}^n . A multifunction g is closed valued if $g(\xi)$ is a closed subset of \mathbb{R}^n for every $\xi \in \mathbb{R}^m$ and is measurable if, for every closed set $\Xi \subset \mathbb{R}^n$, the set $g^{-1}(\Xi) := \{\xi \in \mathbb{R}^m | g(\xi) \cap \Xi \neq \emptyset\}$ is measurable. We define $\text{dom}(g) := g^{-1}(\mathbb{R}^n)$. A mapping $G : \text{dom}(g) \rightarrow \mathbb{R}^n$ is called a measurable (integrable) selection of g if it is measurable (integrable) and $G(\xi) \in g(\xi)$ for every $\xi \in \text{dom}(g)$. We use $\int \partial g(\xi) d\xi$ to represent the set $\{\int \partial G(\xi) d\xi | G \text{ is an integrable selection of } g\}$.

2 Data-Driven Distributionally Robust Optimization

In this paper, we consider an ambiguity set that consists of the density functions whose value is between two known non-negative functions. To construct such a set, we assume that there exists a set of N independent realizations of the random variable ξ (i.e., samples from p^*) denoted by

$$\hat{\Xi}_N := \{\hat{\xi}^1, \dots, \hat{\xi}^N\} \subseteq \Xi.$$

Then, for a given $\alpha \in (0, 1)$, we construct two functions $l_\alpha : \mathbb{R}^m \rightarrow [0, +\infty]$ and $u_\alpha : \mathbb{R}^m \rightarrow [0, +\infty]$ based on $\widehat{\Xi}_N$, α , and some prior knowledge on p^* (e.g., its shape property) such that

$$\mathbb{P} \{l_\alpha(\xi) \leq p^*(\xi) \leq u_\alpha(\xi), \forall \xi \in [a, b]\} \geq 1 - \alpha. \quad (5)$$

We call the pair (l_α, u_α) the *confidence bands* for the density functions p^* at a confidence level of $1 - \alpha$ and α is called the significance level. We will introduce two methods to construct such a ambiguity set in Section 3.

Using (l_α, u_α) , we can construct an ambiguity set that contains p^* with a confidence level of $1 - \alpha$. More specifically, let \mathcal{L} be the space of all non-negative Lebesgue-measurable functions on \mathbb{R}^m . We consider the following ambiguity set:

$$\mathcal{D}(\widehat{\Xi}_N, \alpha) := \left\{ p \in \mathcal{L} \mid l_\alpha(\xi) \leq p(\xi) \leq u_\alpha(\xi), \forall \xi \in \Xi, \int_\Xi p(\xi) d\xi = 1 \right\}, \quad (6)$$

which satisfies $\mathbb{P} \left(p^* \in \mathcal{D}(\widehat{\Xi}_N, \alpha) \right) \geq 1 - \alpha$ according to (5).

With $\mathcal{D}(\widehat{\Xi}_N, \alpha)$, we can specify the distributionally robust optimization problem in (4) as

$$v_{\mathcal{D}(\widehat{\Xi}_N, \alpha)}^* := \inf_{x \in \mathcal{X}} \sup_{p \in \mathcal{D}(\widehat{\Xi}_N, \alpha)} \int_\Xi f(x, \xi) p(\xi) d\xi. \quad (7)$$

Immediately, we have the following result for the optimal value of (7).

Proposition 1. *Suppose the minimal objective value of (7) is finite and achieved at $\widehat{x}_N \in \mathcal{X}$. Let $\widehat{v}_N := \mathbb{E}_{p^*}[f(\widehat{x}_N, \xi)]$. We have $\mathbb{P} \left(v_{\mathcal{D}(\widehat{\Xi}_N, \alpha)}^* \geq \widehat{v}_N \right) \geq 1 - \alpha$.*

Proof. Proof. Whenever $p^* \in \mathcal{D}(\widehat{\Xi}_N, \alpha)$, we have $v_{\mathcal{D}(\widehat{\Xi}_N, \alpha)}^* = \sup_{p \in \mathcal{D}(\widehat{\Xi}_N, \alpha)} \mathbb{E}_p[f(\widehat{x}_N, \xi)] \geq \mathbb{E}_{p^*}[f(\widehat{x}_N, \xi)] =$

\widehat{v}_N . By the construction of the ambiguity set $\mathcal{D}(\widehat{\Xi}_N, \alpha)$, we have $\mathbb{P} \left(v_{\mathcal{D}(\widehat{\Xi}_N, \alpha)}^* \geq \widehat{v}_N \right) \geq \mathbb{P} \left(p^* \in \mathcal{D}(\widehat{\Xi}_N, \alpha) \right) \geq 1 - \alpha$. \square

3 Data-Driven Ambiguity Sets

In this section, we present two existing methods from the statistics literature to construct a confidence band for a density function based on observed data. The first method is applicable to only univariate distributions while the second one is applicable to multivariate distributions. Although we only described two confidence bands as the specific instances for constructing the ambiguity set in (6), the optimization method we propose in Section 4 can be applied to DRO with any ambiguity set in the form of (6) with other constructions of confidence bands.

3.1 Shape-restricted confidence bands

In this subsection, we present the method by [19] to construct a confidence band (l_α, u_α) for p^* with a confidence level of $1 - \alpha$ and use it to build an ambiguity set like (6). Although this method only

applies to a univariate density function, it is able to incorporate some shape information about p^* (e.g., unimodality and monotonicity) into the construction of the ambiguity set which improves the theoretical convergence rate of the set to the true density p^* .

We need the following assumptions in this subsection.

Assumption 1 (For shape-restricted confidence bands.). *We assume:*

- A1. Ξ (the support of p^*) is connected and contained in $[a, b]$ with known a and b satisfying $-\infty < a < b < +\infty$.
- A2. p^* is unimodal with a known mode $\mu \in [a, b]$, meaning that p^* is monotonically increasing on $[a, \mu]$ and decreasing on $[\mu, b]$.
- A3. There exists a known constant U such that $p(\mu) \leq U$ for any $\xi \in \Xi$.

In Assumption 1 [A1.], we assume that a and b are finite for the simplicity of the notations in the derivation below. In fact, our results can be generalized when $a = -\infty$ and $b = +\infty$. Moreover, for most applications, a conservative estimation of the range of ξ is usually available, which can be directly used as $[a, b]$. We also note that Assumption 1 [A2.] also covers the case where p^* is known to be monotonically increasing ($\mu = b$) or decreasing ($\mu = a$). In some real applications, the unimodality or monotonicity of a random parameter is a well-known fact and thus should be incorporated into the problem formulation. Note that the method by [19] can also be used to construct a confidence band even when the mode is only known to be in an interval $[\mu^+, \mu^-] \subset [a, b]$.

First, let $(\hat{\xi}_{(1)}, \dots, \hat{\xi}_{(N)})$ be the order statistics of ξ constructed from $\hat{\Xi}_N$ satisfying $\hat{\xi}_{(1)} < \dots < \hat{\xi}_{(N)}$. We choose a group size K satisfying $0 < K < N$ and define $M' := \lfloor N/K \rfloor$ and $M := \lceil N/K \rceil$. Then we partition the sorted sequence $(\hat{\xi}_{(1)}, \dots, \hat{\xi}_{(N)})$ into M groups with the first M' groups of size K and the M th group (if $M' = M - 1$) of size $N - KM'$. We then define $k_i := (i - 1)K + 1$ for $i = 1, 2, \dots, M'$ and $k_i = N$ for $i = M$ if $M \neq M'$. Let $F^* : \mathbb{R} \rightarrow [0, 1]$ be the cumulative density function of ξ . It is well-known that (citation?) the random variable $\Delta_i := F^*(\hat{\xi}_{(k_i)}) - F^*(\hat{\xi}_{(k_{i-1})})$ has the same distribution as

$$\tilde{\Delta}_i := \begin{cases} \frac{\Gamma(K, 1)}{\sum_{j=1}^{M'} \Gamma(K, 1) + \Gamma(M - M' + 1, 1)} & \text{if } i = 1, 2, \dots, M' \\ \frac{\Gamma(N - KM', 1)}{\sum_{j=1}^{M'} \Gamma(K, 1) + \Gamma(M - M' + 1, 1)} & \text{if } i = M \neq M' \end{cases}, \quad (8)$$

where $\Gamma(A, B)$ represents a gamma random variable with a shape parameter A and a rate parameter B . Let $c^-(\alpha)$ and $c^+(\alpha)$ be two constants that satisfy

$$\mathbb{P}\{c^-(\alpha) \leq \Delta_i \leq c^+(\alpha)\} = \mathbb{P}\{c^-(\alpha) \leq \tilde{\Delta}_i \leq c^+(\alpha)\} \geq 1 - \alpha.$$

Both $c^-(\alpha)$ and $c^+(\alpha)$ can be estimated to arbitrary precision by sampling according to (8). Let \mathcal{L}_μ be the set of all density functions on \mathbb{R} with mode at μ , i.e.

$$\mathcal{L}_\mu := \left\{ p \in \mathcal{L} \mid p(\xi) \geq 0, \int_{\mathbb{R}} p(\xi) d\xi = 1, \text{ the mode of } p \text{ is } \mu. \right\}$$

and

$$\mathcal{D}_\mu(\widehat{\Xi}, \alpha) := \left\{ p \in \mathcal{L}_\mu \left| c^-(\alpha) \leq \int_{\hat{\xi}_{(k_{i-1})}}^{\hat{\xi}_{(k_i)}} p(\xi) d\xi \leq c^+(\alpha), i = 2, 3, \dots, M \right. \right\}.$$

Then, by the definitions of $c^-(\alpha)$ and $c^+(\alpha)$, we have

$$\mathbb{P} \left\{ p^* \in \mathcal{D}_\mu(\widehat{\Xi}, \alpha) \right\} \geq 1 - \alpha.$$

Given the property above, $\mathcal{D}_\mu(\widehat{\Xi}, \alpha)$ can be used as an uncertainty set $\mathcal{D}(\widehat{\Xi}, \alpha)$ in (7). However, this uncertainty set may be too large to ensure a good solution from solving (7). Therefore, we need to further refine it using the prior knowledge about the shape of p^* given in Assumption 1. Next, we show how to construct a shape-constrained confidence band using the technique from in [19].

Given any $\xi \in [a, b]$, let \widehat{N} ($M \leq \widehat{N} \leq M + 4$) be the number of distinct elements in the set $\{a, b, \mu, \xi, \hat{\xi}_{(k_1)}, \dots, \hat{\xi}_{(k_M)}\}$ and z_j be the j th smallest element in this set. Then, we consider the following two sets of non-negative step functions on \mathbb{R} defined as

$$\begin{aligned} \mathcal{D}_\mu^-(\widehat{\Xi}, \xi) &:= \left\{ p \in \mathcal{L} \left| \begin{aligned} p(\cdot) &= \sum_{\{j: z_{j+1} < \mu\}} \beta_j \mathbb{I}_{(z_j, z_{j+1}]}(\cdot) + \max_{\{j: \mu \in [z_j, z_{j+1}]\}} \beta_j \mathbb{I}_\mu(\cdot) \\ &+ \sum_{\{j: \mu \in [z_j, z_{j+1}]\}} \beta_j \mathbb{I}_{(z_j, z_{j+1})}(\cdot) + \sum_{\{j: z_j > \mu\}} \beta_j \mathbb{I}_{[z_j, z_{j+1})}(\cdot), \\ &\text{where } \beta_j \in [0, U] \text{ for } j = 1, 2, \dots, \widehat{N} - 1. \end{aligned} \right. \right\} \quad (9) \\ \mathcal{D}_\mu^+(\widehat{\Xi}, \xi) &:= \left\{ p \in \mathcal{L} \left| \begin{aligned} p(\cdot) &= \sum_{\{j: z_{j+1} \leq \mu\}} \beta_j \mathbb{I}_{[z_j, z_{j+1})}(\cdot) + \infty \cdot \mathbb{I}_\mu(\xi) + \sum_{\{j: z_j > \mu\}} \beta_j \mathbb{I}_{(z_j, z_{j+1}]}(\cdot), \\ &\text{where } \beta_j \in [0, U] \text{ for } j = 1, 2, \dots, \widehat{N} - 1. \end{aligned} \right. \right\} \quad (10) \end{aligned}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Both $\mathcal{D}_\mu^-(\widehat{\Xi}, \xi)$ and $\mathcal{D}_\mu^+(\widehat{\Xi}, \xi)$ consist of piecewise constant density functions on $[a, b]$ with break points on $\{a, b, \mu, \xi, \hat{\xi}_{(k_1)}, \dots, \hat{\xi}_{(k_M)}\}$ but with different continuous properties (i.e., left-continuity or right-continuity) on these break points. Then, the confidence band for p^* given by [19] is a pair of functions (l_α, u_α) defined as

$$l_\alpha^{SR}(\xi) := \inf_{p \in \mathcal{D}_\mu^-(\widehat{\Xi}, \xi) \cap \mathcal{D}_\mu(\widehat{\Xi}, \alpha)} p(\xi) \quad \text{and} \quad u_\alpha^{SR}(\xi) := \sup_{p \in \mathcal{D}_\mu^+(\widehat{\Xi}, \xi) \cap \mathcal{D}_\mu(\widehat{\Xi}, \alpha)} p(\xi). \quad (11)$$

Here, the superscript “SR” refers to “shape restricted”. By its definition, the value of function $l_\alpha^{SR}(\xi)$ ($u_\alpha^{SR}(\xi)$) equals the smallest (largest) value at ξ of all density functions which have the form in (9) ((10)) and satisfy $c^-(\alpha) \leq \int_{\hat{\xi}_{(k_{i-1})}}^{\hat{\xi}_{(k_i)}} p(\xi) d\xi \leq c^+(\alpha)$ for all i . Then, our shape-restricted uncertainty set are defined as

$$\mathcal{D}^{SR}(\widehat{\Xi}_N, \alpha) := \left\{ p \in \mathcal{L} \left| l_\alpha^{SR}(\xi) \leq p(\xi) \leq u_\alpha^{SR}(\xi), \forall \xi \in \Xi, \int_\Xi p(\xi) d\xi = 1 \right. \right\}. \quad (12)$$

The following theorem is given by [19].

Theorem 1. $\mathbb{P} \left\{ p^* \in \mathcal{D}^{SR}(\widehat{\Xi}_N, \alpha) \right\} \geq 1 - \alpha.$

We then describe the method for compute $l_\alpha^{SR}(\xi)$ and $u_\alpha^{SR}(\xi)$ for any ξ . Given ξ , recall that z_j denotes the j th smallest element in $\{a, b, \mu, \xi, \hat{\xi}_{(k_1)}, \dots, \hat{\xi}_{(k_M)}\}$. Let \hat{j} and \tilde{j} be the indexes in $\{1, 2, \dots, \widehat{N}\}$ such that $z_{\hat{j}} = \xi$ and $z_{\tilde{j}} = \mu$. Using these indexes, we define a polyhedron

$$\mathcal{H}(\widehat{\Xi}_N, \alpha) = \left\{ \beta \in \mathbb{R}^{\widehat{N}} \left| \begin{array}{l} \beta_1 \leq \beta_2 \leq \dots \leq \beta_{\tilde{j}-1}, \quad \beta_{\tilde{j}} \geq \beta_{\tilde{j}+1} \geq \dots \geq \beta_{\widehat{N}-1} \\ c^-(\alpha) \leq \sum_{j: \xi_{(k_{i-1})} \leq z_j < \xi_{(k_i)}} \beta_j (z_{j+1} - z_j) \leq c^+(\alpha), \quad \forall i = 2, \dots, M \\ \sum_{j=1}^{\widehat{N}-1} \beta_j (z_{j+1} - z_j) = 1, \\ 0 \leq \beta_j \leq U, \quad \forall 1 \leq j \leq \widehat{N} - 1, \end{array} \right. \right\}. \quad (13)$$

According to (9), (10) and (11), for any ξ , $l_\alpha^{SR}(\xi)$ and $u_\alpha^{SR}(\xi)$ can be computed respectively by solving the following linear programs

$$l_\alpha^{SR}(\xi) = \min_{\beta \in \mathcal{H}(\widehat{\Xi}_N, \alpha)} \beta_{\hat{j}} \quad \text{and} \quad u_\alpha^{SR}(\xi) = \min_{\beta \in \mathcal{H}(\widehat{\Xi}_N, \alpha)} \beta_{\tilde{j}}. \quad (14)$$

Note that the first line of constraints in (13) is because μ is the mode of any p in $\mathcal{D}_\mu^-(\widehat{\Xi}, \xi) \cap \mathcal{D}_\mu(\widehat{\Xi}, \alpha)$ and $\mathcal{D}_\mu^+(\widehat{\Xi}, \xi) \cap \mathcal{D}_\mu(\widehat{\Xi}, \alpha)$. The second line of the constraints in (13) corresponds to the condition $c^-(\alpha) \leq \int_{\hat{\xi}_{(k_{i-1})}}^{\hat{\xi}_{(k_i)}} p(\xi) d\xi \leq c^+(\alpha)$ when p is the step function in (9) and (10). The third line requires that p must a density function while the last line requires p is non-negative and no more than U according the prior information of p^* .

We summary this procedure by [19] for constructing a confidence band for a unimodal density function in Algorithm 1.

It is worthwhile to note that if the density is monotone, for example, non-decreasing, we have $\mu = b$ and the first line of the constraints in (13) becomes $\beta_1 \leq \beta_2 \leq \dots \leq \beta_{\widehat{N}-1}$. Note also that, in the original work [19], no upper bound U for p^* is needed to construct this uncertainty set $\mathcal{D}^{SR}(\widehat{\Xi}_N, \alpha)$. In this paper, we require knowing U so that we can include a constraint $\beta_j \leq U$ in (13), this modification does not change the statistical property of $\mathcal{D}^{SR}(\widehat{\Xi}_N, \alpha)$ (i.e., Theorem 1) once we assume $p^* \leq U$. We need the constraint $\beta_j \leq U$ only to ensure $l_\alpha^{SR}(\xi) \leq U$ and $u_\alpha^{SR}(\xi) \leq U$ which are needed in our theoretical analysis.

[19] established the convergence rate of the constructed confidence band in the following proposition.¹

Proposition 2 (Corollary 7.2 [19]). *Let $\xi \in \Xi$ but $\xi \neq \mu$. Suppose p^* is (C, ρ) -Holder continuous for constants $C > 0$ and $\rho > 0$, i.e., $|p^*(\xi') - p^*(\xi)| \leq C |\xi' - \xi|^\rho$ for any ξ and ξ' . Suppose*

¹The original Corollary 7.2 [19] was stated a little differently form from Proposition 2. In fact, the conclusion there replaces \lim in (15) by \liminf and equality in (15) by > 0 . Their results were obtained by choosing the parameter $\tau = 2 + 2\rho$ appearing in the proof of Corollary 7.2 [19]. However, the same proof works for Proposition 2 and thus implies (15) if we choose $\tau = 3 + 2\rho$.

Algorithm 1 Shape-restricted confidence band $(l_\alpha^{SR}(\xi), u_\alpha^{SR}(\xi))$ at $\xi \in \Xi$

Input: Data $\hat{\Xi}_N = \{\hat{\xi}^1, \dots, \hat{\xi}^N\}$ sampled from p^* , an interval $[a, b]$ containing Ξ , the mode μ of p^* , a constant $U \geq p^*$ on Ξ , a significance level $\alpha \in (0, 1)$, a group size K with $0 < K < N$ and a point $\xi \in \Xi$.

- 1: Let $M' := \lfloor N/K \rfloor$ and $M := \lceil N/K \rceil$
- 2: Let $\hat{\xi}_{(1)}, \dots, \hat{\xi}_{(N)}$ be the order statistics of $\hat{\Xi}_N$ with $\hat{\xi}_{(1)} < \dots < \hat{\xi}_{(N)}$.
- 3: Let $k_i := \begin{cases} (i-1)K+1 & \text{if } i = 1, 2, \dots, M' \\ N & \text{if } i = M \neq M' \end{cases}$
- 4: Define random variables

$$\tilde{\Delta}_i := \begin{cases} \frac{\Gamma(K, 1)}{\sum_{j=1}^{M'} \Gamma(K, 1) + \Gamma(M - M' + 1, 1)} & \text{if } i = 1, 2, \dots, M' \\ \frac{\Gamma(N - KM', 1)}{\sum_{j=1}^{M'} \Gamma(K, 1) + \Gamma(M - M' + 1, 1)} & \text{if } i = M \neq M' \end{cases}$$

where $\Gamma(\cdot, \cdot)$ is a gamma random variable parameterized by a shape parameter and a scale parameter.

- 5: Find constants $c^-(\alpha)$ and $c^+(\alpha)$ that satisfy $\mathbb{P}\left\{c^-(\alpha) \leq \tilde{\Delta}_i \leq c^+(\alpha), i = 1, \dots, M\right\} \geq 1 - \alpha$.
- 6: Let z_j be the j th smallest value in the set $\{a, \mu, \xi, \hat{\xi}_{(k_1)}, \dots, \hat{\xi}_{(k_M)}, b\}$ for $j = 1, \dots, \hat{N}$ where \hat{N} represents the number of distinct elements in that set.
- 7: Let \hat{j} and \tilde{j} be the indexes in $\{1, 2, \dots, \hat{N}\}$ such that $z_{\hat{j}} = \xi$ and $z_{\tilde{j}} = \mu$. Define the polyhedron $\mathcal{H}(\hat{\Xi}_N, \alpha)$ in (13).
- 8: Compute $(l_\alpha^{SR}(\xi), u_\alpha^{SR}(\xi))$ by solving the two linear programs in (14).

Output: $(l_\alpha^{SR}(\xi), u_\alpha^{SR}(\xi))$

$K = \left\lceil B (N^{2\rho} \log N)^{1/(1+2\rho)} \right\rceil$ for a constant $B > 0$ and sufficiently large N . We have

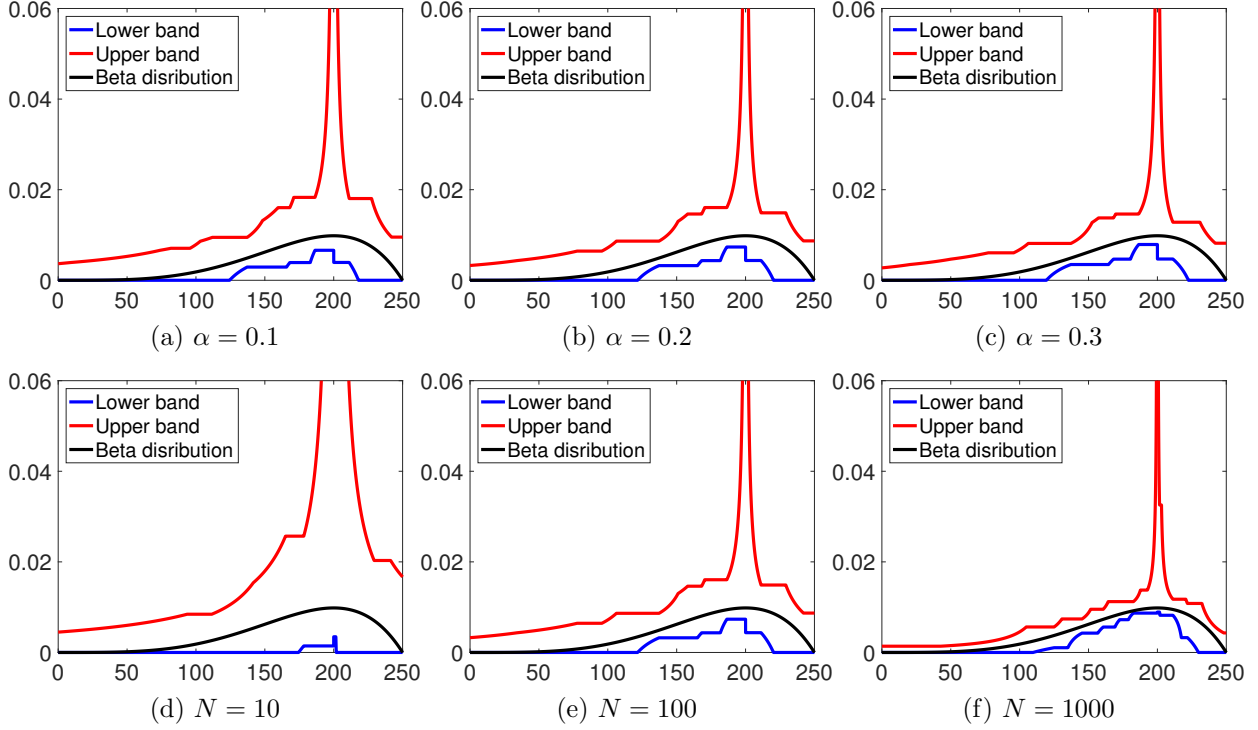
$$\lim_{N \rightarrow \infty} \mathbb{P} \left\{ |u_\alpha^{SR}(\xi) - l_\alpha^{SR}(\xi)| \leq 4p^*(\xi) \left(\sqrt{\frac{3+2\rho}{1+2\rho}} B^{-1/2} + C(p^*(\xi))^{-(\rho+1)} B^\rho \right) \left(\frac{\log N}{N} \right)^{\rho/(1+2\rho)} \right\} = 1. \quad (15)$$

According to the lower bound in [23], the convergence rate $O\left(\left(\frac{\log N}{N}\right)^{\rho/(1+2\rho)}\right)$ in (15) attains the minimax rate and thus is optimal (upto a constant factor) for the confidence band for a unimodal density. It is also worthwhile to note that the rate of convergence from Kolmogorov-Smirnov distance is slower as compared to this approach. As shown in [18], the confidence band formed by the Kolmogorov-Smirnov distance is only $O\left((\frac{1}{N})^{1/4}\right)$, which is slower than the rate of $O\left((\frac{\log N}{N})^{1/3}\right)$ in Proposition 2 when $\rho = 1$.

Proposition 2 shows the pointwise convergence of the confidence band to the true distribution p^* . In the next theorem, we show that, under additional assumption, we can characterize the convergence in probability of $v_{\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)}^*$ define in (7) to v^* in (2).

Theorem 2. Suppose p^* is (C, ρ) -Holder continuous for constants $C > 0$ and $\rho > 0$, i.e., $|p^*(\xi') - p^*(\xi)| \leq C |\xi' - \xi|^\rho$ for any ξ and ξ' . Moreover, suppose $\max_{x \in \mathcal{X}, \xi \in \Xi} |f(x, \xi)| < +\infty$. For any $\epsilon > 0$ and

Figure 1: The data-generating distribution is $250 \times \text{Beta}(5, 2)$. The first row contains three subfigures with the same sample size ($N = 100$) but different significance levels α . The second row contains three subfigures with the same significance level ($\alpha = 0.2$) but different sample sizes.



$\theta \in (0, 1)$, there exists an $N_{\epsilon, \theta}$ such that for any $N \geq N_{\epsilon, \theta}$, we have

$$\mathbb{P} \left(\sup_{x \in \mathcal{X}} \left| \int_{\Xi} f(x, \xi) p^*(\xi) d\xi - \sup_{p \in \mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)} \int_{\Xi} f(x, \xi) p(\xi) d\xi \right| \leq \epsilon \right) \geq 1 - \theta$$

and $\mathbb{P} \left(\left| v_{\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)}^* - v^* \right| \leq \epsilon \right) \geq 1 - \theta$.

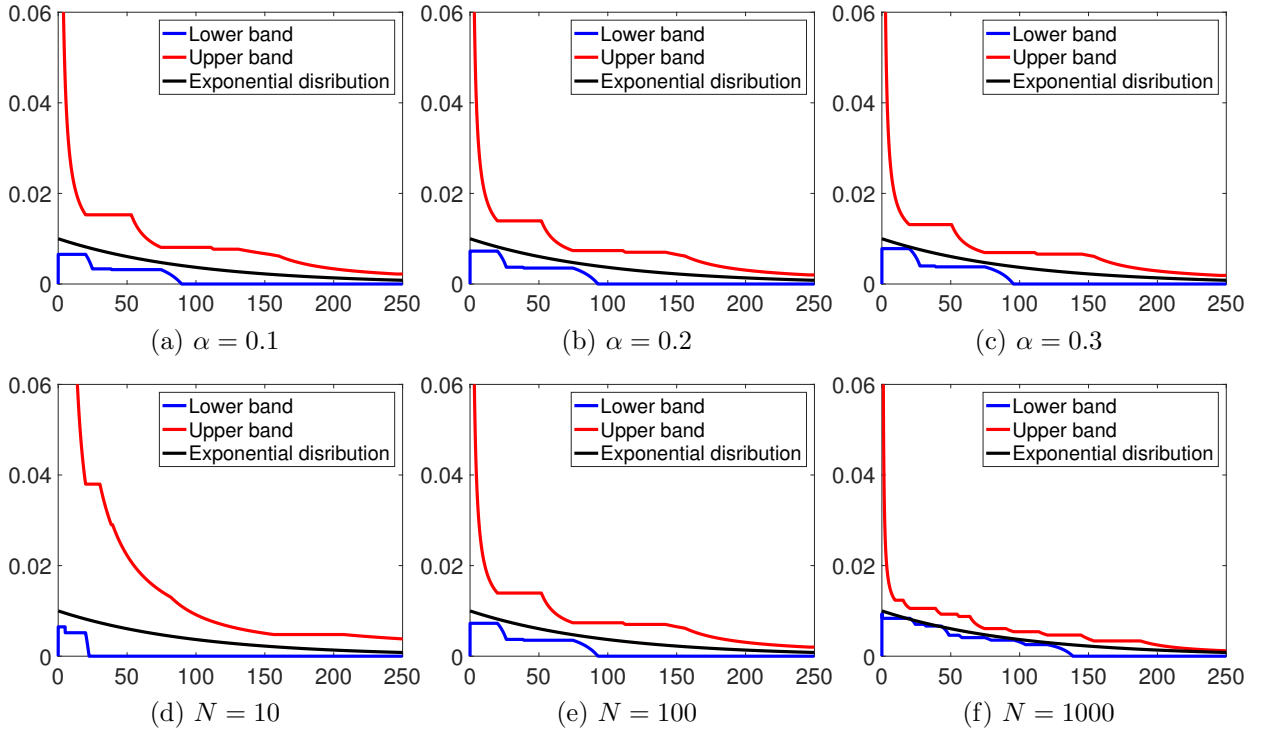
Proof. Proof. See Appendix 6. □

Finally, illustrations of this method are given in Figure 1 for a beta distribution and in Figure 2 for a truncated exponential distribution. (What exact parameters?)

3.2 Kernel-density-estimation confidence bands

The shape-restricted confidence bands described in the previous section can only be applied to univariate density function. In this subsection, we describe a method to construct confidence bands for multivariate density function based on the classical *kernel density estimation* (KDE) [38, 33]. This method requires a *kernel function* which is a mapping $\mathcal{K} : \mathbb{R}^m \rightarrow [0, +\infty)$ satisfying $\int_{\mathbb{R}^m} \mathcal{K}(\xi) d\xi = 1$. The commonly use kernel functions include uniform kernel $\mathcal{K}(\xi) = \frac{1}{2^m} \mathbb{I}_{\|\xi\|_{\infty} \leq 1}(\xi)$ and Gaussian kernel $\mathcal{K}(\xi) = \frac{1}{(2\pi)^{m/2}} \exp(-\|\xi\|_2^2/2)$. Let $h > 0$ be a *bandwidth parameter*. Recall that

Figure 2: The data-generating distribution is a truncated exponential distribution, $\text{Exp}(1/100)$ with the support $[0, 250]$. The first row contains three subfigures with the same sample size ($N = 100$) but different significance levels α . The second row contains three subfigures with the same significance level ($\alpha = 0.2$) but different sample sizes.



$\widehat{\Xi}_N := \{\hat{\xi}_1, \dots, \hat{\xi}_N\} \subseteq \mathbb{R}^m$ is N i.i.d. samples drawn from p^* . The KDE of p^* based on $\widehat{\Xi}_N$, \mathcal{K} and h is

$$\widehat{p}_h(\xi) := \frac{1}{N} \sum_{i=1}^N \frac{1}{h^m} \mathcal{K} \left(\frac{\xi - \hat{\xi}_i}{h} \right). \quad (16)$$

The convergence of $\widehat{p}_h(\xi)$ to the true density $p^*(\xi)$ have been studied for a long time (see e.g. [42]) with most of existing works focusing the asymptotic convergence property. Recently, the finite-sample non-asymptotic convergence property of KDE is characterized by [36] and [21]. The confidence band we construct based on KDE utilize the non-asymptotic convergence bound by [21].

We need the following assumptions in this subsection.

Assumption 2 (For KDE confidence bands.). *We assume:*

- A1. *There exists a constant U such that $p(\mu) \leq U$ for any $\xi \in \Xi$.*
- A2. *There exists a non-increasing function $\kappa : [0, +\infty) \rightarrow [0, +\infty)$ such that $\mathcal{K}(\xi) = \kappa(\|\xi\|_2)$.*
- A3. *There exists $r, C_r > 0$ and $\tau > 0$ such that for $t > \tau$, $\kappa(t) \leq C_r \cdot \exp(-t^r)$.*

A2 and A3 of Assumption 2 hold if κ is one of the popular kernel densities (up to scaling) in \mathbb{R} , including the two mentioned above as well as Exponential, Tricube, triangular, and Epanechnikov kernels. Under these assumptions, the following finite-sample convergence result is established by [21].

Proposition 3. *[Theorem 2. [21]] Suppose p^* is (C, ρ) -Holder continuous for constants $C > 0$ and $\rho \in (0, 1]$, i.e., $|p^*(\xi') - p^*(\xi)| \leq C \|\xi' - \xi\|_2^\rho$ for any ξ and ξ' . Let $\alpha \in (0, 1)$ and V_m be the volume of the unit ball in \mathbb{R}^m . Suppose $h > \left(\frac{\log(N/\alpha)}{N}\right)^{1/m}$. We have*

$$\mathbb{P} \left\{ \sup_{\xi \in \Xi} |\widehat{p}_h(\xi) - p^*(\xi)| \leq C_1 h^\rho + C_2 \sqrt{\frac{\log(N/\alpha)}{N h^m}} \right\} \geq 1 - \alpha.$$

where² $C_1 = V_m C \int_0^\infty \kappa(t) t^{m+\rho} dt$ and $C_2 = 8m\sqrt{V_m U} \left(\int_0^\infty \kappa(t) t^{m/2} dt + 1 \right) + 64m^2 \kappa(0)$. As a consequence, if $h = \left(\frac{\log(N/\alpha)}{N}\right)^{1/(2\rho+m)}$, we have

$$\mathbb{P} \left\{ \sup_{\xi \in \Xi} |\widehat{p}_h(\xi) - p^*(\xi)| \leq (C_1 + C_2) \left(\frac{\log(N/\alpha)}{N}\right)^{\rho/(2\rho+m)} \right\} \geq 1 - \alpha.$$

Based on the convergence property in Proposition 3, with $h > \left(\frac{\log(N/\alpha)}{N}\right)^{1/m}$, we construct the KDE confidence band of a significance level of α for $p^*(\xi)$ as follows

$$l_\alpha^{KDE}(\xi) = \max\{0, \widehat{p}_h(\xi) - \delta\}, \quad u_\alpha^{KDE}(\xi) = \widehat{p}_h(\xi) + \delta \quad (17)$$

²Note that $\int_0^\infty \kappa(t) t^{m+\rho} dt < +\infty$ because of A4 of Assumption 2.

where

$$\delta = C_1 h^\rho + C_2 \sqrt{\frac{\log(N/\alpha)}{N h^m}}. \quad (18)$$

The corresponding uncertainty set is

$$\mathcal{D}^{KDE}(\widehat{\Xi}_N, \alpha) := \left\{ p \in \mathcal{L} \mid l_\alpha^{KDE}(\xi) \leq p(\xi) \leq u_\alpha^{KDE}(\xi), \forall \xi \in \Xi, \int_\Xi p(\xi) d\xi = 1 \right\}. \quad (19)$$

The following property is a direct consequence of Proposition 3.

Theorem 3. $\mathbb{P} \left\{ p^* \in \mathcal{D}^{KDE}(\widehat{\Xi}_N, \alpha) \right\} \geq 1 - \alpha.$

We summary this procedure for constructing a KDE confidence band in Algorithm 2.

Algorithm 2 KDE-based confidence band $(l_\alpha^{KDE}(\xi), u_\alpha^{KDE}(\xi))$ at $\xi \in \Xi$

Input: Data $\widehat{\Xi}_N = \{\hat{\xi}^1, \dots, \hat{\xi}^N\}$ sampled from p^* , a constant $U \geq p^*$ on Ξ , a significance level $\alpha \in (0, 1)$, a kernel function $\mathcal{K}(\xi) = \kappa(\|\xi\|_2)$ with κ satisfying Assumption 2, a bandwidth $h > \left(\frac{\log(N/\alpha)}{N} \right)^{1/m}$ and a point $\xi \in \Xi$.

- 1: Compute $C_1 = V_m C \int_0^\infty \kappa(t) t^{m+\rho} dt$ and $C_2 = 8m\sqrt{V_m U} \left(\int_0^\infty \kappa(t) t^{m/2} dt + 1 \right) + 64m^2 \kappa(0)$.
- 2: Let $\widehat{p}_h(\xi)$ and δ defined as in (16) and (18), respectively.
- 3: Compute $(l_\alpha^{KDE}(\xi), u_\alpha^{KDE}(\xi))$ as in (17).

Output: $(l_\alpha^{KDE}(\xi), u_\alpha^{KDE}(\xi))$

The following theorem show that, under additional assumption, $v_{\mathcal{D}^{KDE}(\widehat{\Xi}_N, \alpha)}^*$ define in (7) converges to v^* in (2).

Theorem 4. Suppose p^* is (C, ρ) -Holder continuous for constants $C > 0$ and $\rho > 0$, i.e., $|p^*(\xi') - p^*(\xi)| \leq C \|\xi' - \xi\|_2^\rho$ for any ξ and ξ' . Moreover, suppose $\max_{x \in \mathcal{X}, \xi \in \Xi} |f(x, \xi)| < +\infty$ and $h > \left(\frac{\log(N/\alpha)}{N} \right)^{1/m}$. For any $\epsilon > 0$ and $\theta \in (0, 1)$, there exists an $N_{\epsilon, \theta}$ such that for any $N \geq N_{\epsilon, \theta}$, we have

$$\mathbb{P} \left(\sup_{x \in \mathcal{X}} \left| \int_\Xi f(x, \xi) p^*(\xi) d\xi - \sup_{p \in \mathcal{D}^{KDE}(\widehat{\Xi}_N, \alpha)} \int_\Xi f(x, \xi) p(\xi) d\xi \right| \leq \epsilon \right) \geq 1 - \theta$$

and $\mathbb{P} \left(\left| v_{\mathcal{D}^{KDE}(\widehat{\Xi}_N, \alpha)}^* - v^* \right| \leq \epsilon \right) \geq 1 - \theta.$

Proof. Proof. See Appendix 6. □

4 An Numerical Method for DRO

In this section, we discuss a numerical scheme for solving the DRO problem in (7) with an ambiguity set $\mathcal{D}(\widehat{\Xi}_N, \alpha)$ in the form of (6). Here, the ambiguity can be one of the two ambiguity sets $\mathcal{D}^{SR}(\widehat{\Xi}_N, \alpha)$ and $\mathcal{D}^{KDE}(\widehat{\Xi}_N, \alpha)$ introduced in Section 3 as long as Assumption 1 or 2 is satisfied. However, our numerical methods can be applied to a general ambiguity set like (6) as long as the following assumptions are satisfied. We make the following assumption regarding to $\mathcal{D}_\alpha(\widehat{\Xi}_N)$:

Assumption 3 (On ambiguity set $\mathcal{D}(\widehat{\Xi}_N, \alpha)$).

a. $\int |f(x, \xi)|u(\xi)d\xi < \infty, \quad \forall x \in \mathcal{X}.$

b. $\mathcal{D}(\widehat{\Xi}_N, \alpha) \neq \emptyset.$

Note that, under Assumption 1, $\mathcal{D}^{SR}(\widehat{\Xi}_N, \alpha)$ satisfies Assumption 3 because $u_\alpha^{SR}(\xi) \leq U$ for an upper bound U of p^* according to its definition (14) and the constraints $0 \leq \beta_j \leq U$ in (13). Under Assumption 2, with the appropriate bandwidth h (See Theorem 3), we have $u_\alpha^{KDE}(\xi) = \widehat{p}_h(\xi) + \delta \leq p^*(\xi) + 2\delta \leq U + 2\delta$ with a high probability.

We define $v(x)$ as the optimal value of the inner maximization problem of (6), i.e.,

$$v(x) := \sup_{p \in \mathcal{D}(\widehat{\Xi}_N, \alpha)} \int f(x, \xi)p(\xi)d\xi. \quad (20)$$

By Assumption 3, for every $x \in \mathcal{X}$,

$$|v(x)| \leq \sup_{p \in \mathcal{D}(\widehat{\Xi}_N, \alpha)} \int |f(x, \xi)|p(\xi)d\xi \leq \int |f(x, \xi)|u(\xi)d\xi < \infty. \quad (21)$$

Note that (20) is a *continuous linear program* which are formulated with continuously many decision variable (or a functional decision variable) and continuously many constraints. In general, (20) cannot be reformulated as a convex optimization problem of finite dimension and solved by off-the-shelf optimization techniques as in most of the works in distributionally robust optimization. In this section, we propose a stochastic *subgradient descent* (SGD) method for solving (20) and present its convergence rate.

The dual problem of (20) is given as follows:

$$\begin{aligned} \inf \quad & \lambda - \int l(\xi)\alpha(\xi)d\xi + \int u(\xi)\beta(\xi)d\xi \\ \text{s. t.} \quad & \lambda - \alpha(\xi) + \beta(\xi) \geq f(x, \xi) \quad \forall \xi \\ & \alpha(\xi) \geq 0, \beta(\xi) \geq 0 \quad \forall \xi. \end{aligned} \quad (22)$$

Weak duality always hold between (20) and (22). Furthermore, by the result of [40], strong duality holds between (20) and (22). Since $u(\xi) \geq l(\xi) \geq 0$, it is easy to eliminate $\alpha(\xi)$ and $\beta(\xi)$ in (22) and write it equivalently as:

$$v(x) = \inf_{\lambda} \lambda - \int l(\xi)(f(x, \xi) - \lambda)_- d\xi + \int u(\xi)(f(x, \xi) - \lambda)_+ d\xi, \quad (23)$$

where $(z)_- := \max\{-z, 0\}$ and $(z)_+ := \max\{z, 0\}$. Thus, we have

$$v_{\mathcal{D}(\widehat{\Xi}_N, \alpha)}^* := \inf_{x \in \mathcal{X}, \lambda} \left\{ F(x, \lambda) := \lambda - \int l(\xi)(f(x, \xi) - \lambda)_- d\xi + \int u(\xi)(f(x, \xi) - \lambda)_+ d\xi \right\}. \quad (24)$$

In general, the two integrals appearing in (24) do not have an analytical form, raising a challenging for finding the optimal solution. In this section, we provide a stochastic gradient (SGD) method for

solving (24) by approximating these integrals with random samples of ξ . To apply gradient-based algorithm like SGD, we need to be able to compute the subgradient of $F(x, \lambda)$ in (24) with respect to both x and λ . For doing that, we assume there exists a measurable mapping $f'(x, \xi) : \mathcal{X} \times \Xi \rightarrow \mathbb{R}^d$ such that $f'(x, \xi) \in \partial f(x, \xi)$ for any $x \in \mathcal{X}$ and $\xi \in \Xi$, where $\partial f(x, \xi)$ is the subdifferential of f with respect to x . Then, under mild regularity conditions, the subgradients of $F(x, \lambda)$ with respect to x and λ are, respectively,

$$\int l(\xi) f'(x, \xi) \mathbb{I}_{f(x, \xi) < \lambda}(\xi) d\xi + \int u(\xi) f'(x, \xi) \mathbb{I}_{f(x, \xi) \geq \lambda}(\xi) d\xi \in \partial_x F(x, \lambda) \quad (25)$$

$$1 - \int l(\xi) \mathbb{I}_{f(x, \xi) < \lambda}(\xi) d\xi - \int u(\xi) \mathbb{I}_{f(x, \xi) \geq \lambda}(\xi) d\xi \in \partial_\lambda F(x, \lambda) \quad (26)$$

where \mathbb{I} is an indicator function. Then, we can use Monte Carlo method to approximate the integral in the subgradients above.

In particular, let $I \subset \mathbb{R}^d$ be a set that contains the support of $l(\xi)$ and $u(\xi)$. Note that such a set always exists because Ξ is assumed to be compact. We denote the volume of the box I as $|I|$ and assume it can be computed. In fact, we can choose $I \subset \mathbb{R}^d$ to be a box so that $|I|$ is the product of the lengths of all edges. Suppose ξ is sampled from a uniform distribution on I . We can show that (25) is the expectation of $|I| \mathbb{I}_{f(x, \xi) < \lambda}(\xi) l(\xi) f'(x, \xi) + |I| \mathbb{I}_{f(x, \xi) \geq \lambda}(\xi) u(\xi) f'(x, \xi)$ and (26) is the expectation of $1 - |I| \mathbb{I}_{f(x, \xi) < \lambda}(\xi) l(\xi) + |I| \mathbb{I}_{f(x, \xi) \geq \lambda}(\xi) u(\xi)$. Hence, we can use these two as the stochastic subgradient of F . Although the algorithm converges with any number of samples, we can apply the mini-batch techniques by generating B i.i.d. samples from the uniform distribution on I and constructing such a stochastic subgradient using each sample and then take the average of all samples. This mini-batch approach reduces the approximation noise of the Monte Carlo method and accelerate the algorithm in practice.

Based on this idea, we proposed the SGD method for (24) in Algorithm 3. Note that g_x and g_λ are the mini-batch stochastic subgradients of F with respect to x and λ , respectively, with a batch size of B . The convergence analysis of Algorithm 3 is standard and well-known (see e.g. [32]) so we present the theorem below but omit its proof.

Theorem 5. *Suppose there exists a constant M such that $\mathbb{E} \|(g_x, g_\lambda)\|_2^2 \leq M^2$. Algorithm 3 ensures*

$$\begin{aligned} \mathbb{E} [v(\bar{x}_k) - v_{\mathcal{D}(\hat{\Xi}_N, \alpha)}^*] &\leq \mathbb{E} [F(\bar{x}_k, \bar{\lambda}_k) - v_{\mathcal{D}(\hat{\Xi}_N, \alpha)}^*] \leq \frac{\|(x_0, \lambda_0) - (x_*, \lambda_*)\|_2^2 + \sum_{i=0}^k \gamma_i^2 M^2}{\sum_{i=0}^k \gamma_i} \\ &\leq \frac{\|(x_0, \lambda_0) - (x_*, \lambda_*)\|_2^2 + \eta^2 M^2 (1 + \ln(k+1))}{2\eta(\sqrt{k+2} - 1)} \end{aligned} \quad (27)$$

Note that the first inequality in (27) is because of (23) which indicates $v(x) \leq F(x, \lambda)$ for any λ .

Algorithm 3 SGD for (24)

Input: An initial solution $(x_0, \lambda_0) \in \mathcal{X} \times \mathbb{R}$, batch size $B \geq 1$, step length $\eta_k = \eta/\sqrt{k+1}$ with $\eta > 0$, a set $I \subset \mathbb{R}^d$ containing the support of l and u , and the volume of I , denoted by $|I|$.

- 1: **for** $k = 0, 1, \dots$, **do**
- 2: $(\bar{x}_k, \bar{\lambda}_k) = \frac{\sum_{i=0}^k \eta_i(x_i, \lambda_i)}{\sum_{i=0}^k \eta_i}$
- 3: Sample $\{\xi_1, \xi_2, \dots, \xi_B\}$ from a uniform distribution over I .
- 4: Construct the stochastic gradients

$$\begin{aligned} g_x &= \frac{|I|}{B} \sum_{i: f(x_k, \xi_i) < \lambda_k} l(\xi_i) f'(x_k, \xi_i) + \frac{|I|}{B} \sum_{i: f(x_k, \xi_i) \geq \lambda_k} u(\xi_i) f'(x_k, \xi_i) \\ g_\lambda &= 1 - \frac{|I|}{B} \sum_{i: f(x_k, \xi_i) < \lambda_k} l(\xi_i) - \frac{|I|}{B} \sum_{i: f(x_k, \xi_i) \geq \lambda_k} u(\xi_i) \end{aligned}$$

- 5: $x_{k+1} = \operatorname{argmin}_{x \in \mathcal{X}} \frac{1}{2} \|x - x_k + \eta_k g_x\|_2^2$ and $\lambda_{k+1} = \lambda_k - \eta_k g_\lambda$

- 6: **end for**

Output: $(\bar{x}_k, \bar{\lambda}_k)$

5 Computational Results

In this section, we validate our approach on two examples: a single-item newsvendor problem and a portfolio selection problem. Particularly, for the newsvendor example, we compare our approach with that in [2], which applies hypothesis tests to construct ambiguity sets, and for the portfolio selection example, we compare our approach with that in [13], which applies the Wasserstein metric to construct ambiguity sets. We implement Algorithm 3 in MATLAB (R2014a) version 8.3.0.532. The linear programs from the approaches in the literature are solved by CPLEX 12.4 on an Intel Core i3 2.93 GHz Windows computer with 4GB of RAM and the mathematical models are implemented by using the modeling language YALMIP [28] in MATLAB (R2014a) version 8.3.0.532.

5.1 Single-item newsvendor

We consider a classic single-item newsvendor problem in which we assume the demand $\xi \in \mathbb{R}$ of an item follows a continuous distribution with a bounded support set $[a, b] \subseteq \mathbb{R}$ with $0 \leq a < b$ and a bounded density function. An order of $x \geq 0$ units must be placed before demand occurs. After the demand occurs, each unit of unmet demand incurs a shortage cost denoted by $c_s > 0$ and each unit of surplus inventory incurs a holding cost denoted by $c_h > 0$. Hence, the cost function is defined as $f(x, \xi) = \max\{c_s(\xi - x), c_h(x - \xi)\}$, which represents the cost of mismatch between supply and demand. In a classical newsvendor problem, the goal is to determine the order size x to minimize the expected cost. When the demand's distribution is unknown, a corresponding DRO approach can be considered. Assuming a set of historical demand data is available, we construct an ambiguity set in the form of (6), or more specifically, $\mathcal{D}^{SR}(\hat{\Pi}_N, \alpha)$ in (12) and solve the DRO

(7). We compare the optimal order obtained with the one found by the DRO model in [2] where the ambiguity set is built using Kolmogorov-Smirnov test.

In our numerical experiments, we choose $c_s = 19$ and $c_h = 1$ and consider three different ground true distributions for the demand:

1. A truncated normal distribution created by truncating a normal distribution with mean 100 and standard deviation 50 on $[0, 250]$.
2. A beta distribution rescaled onto $[0, 250]$ with parameters $\alpha = 5$ and $\beta = 2$.
3. A truncated exponential distribution created by truncating an exponential distribution with mean 100 on $[0, 250]$.

For each distribution, we consider eight different sample sizes, i.e., $N \in \{10, 20, 40, 80\}$. For each size, we randomly generate a dataset $\hat{\Xi}_N$ by i.i.d. sampling from the demand distribution. Using $\hat{\Xi}_N$, we apply our approach and the method by [2] to construct the ambiguity sets and then solve the corresponding DRO problems to obtain an order size \hat{x} from each approach. To evaluate the out-of-sample performance of \hat{x} , we sample another i.i.d. dataset $\{\xi'_i\}_{i=1}^{N_{\text{large}}}$ with $N_{\text{large}} = 100,000$ from the true distribution and calculate the sample average approximation of the expected cost, i.e., $\frac{1}{N_{\text{large}}} \sum_{i=1}^{N_{\text{large}}} f(\hat{x}, \xi'_i)$ with \hat{x} from each approach. We repeat this procedure 100 times to show the mean and variation of the out-of-sample performance.

When constructing the ambiguity set $\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)$ in our method, we need to provide a significance level α and a group size K (see Algorithm 1). Although a theoretical value of K is suggested in Proposition 2, it involves quantities which are hard to estimate (e.g. B and ρ). When we construct $\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)$, we set $K = K(N, c) = \min \left\{ \left\lceil c (N^2 \log N)^{1/3} \right\rceil, N - 1 \right\}$ and select c from $\{0.5, 0.75, 1, 1.25, 1.5\}$ based on the holdout validation method. The value significance level α is chosen from $\{0.75, 0.8, 0.85, 0.95\}$ based on the same validation method. In particular, we randomly partition $\hat{\Xi}_N$ into $\hat{\Xi}_{\text{train}}$ and $\hat{\Xi}_{\text{test}}$ with $|\hat{\Xi}_{\text{train}}| = 0.7N$ and $|\hat{\Xi}_{\text{test}}| = 0.3N$. Given a combination of c and α , we first construct $\mathcal{D}^{SR}(\hat{\Xi}_{\text{train}}, \alpha)$ using $K = K(0.7N, c)$, and then we solve $x_{c,\alpha}^* \in \operatorname{argmin}_{x \in \mathcal{X}} \sup_{p \in \mathcal{D}^{SR}(\hat{\Xi}_{\text{train}}, \alpha)} \int_{\Xi} f(x, \xi) p(\xi) d\xi$ using Algorithm 3. Then, we select the combination of c and α with the largest $\frac{1}{0.3N} \sum_{\hat{\xi} \in \hat{\Xi}_{\text{test}}} f(x_{c,\alpha}^*, \hat{\xi})$. For each of the 100 independent trial, we repeat this process to select c and α . Note that the entire procedure is simulating how a decision maker select K in practice when only $\hat{\Xi}_N$ is available. For a fair comparison, we also apply the same validation scheme to choose the significance level α used in the method by [2].

We denote our approach by CLX and the method in [2] by BGK. Figure 3 illustrates the performances of CLX and BGK for each of the three distributions of demand. For each sample size N , we plot the 20th percentile, the mean, and 80th percentile of the out-of-sample performances in the 100 trials. The blue lines show the results from CLX, while the red lines show the results from BGK. Figure 3 indicates that CLX has a better out-of-sample performances than BGK when the sample size is small. As the sample size increases, the performances of both approaches become similar as the ambiguity sets in both approaches converge to the true demand distribution.

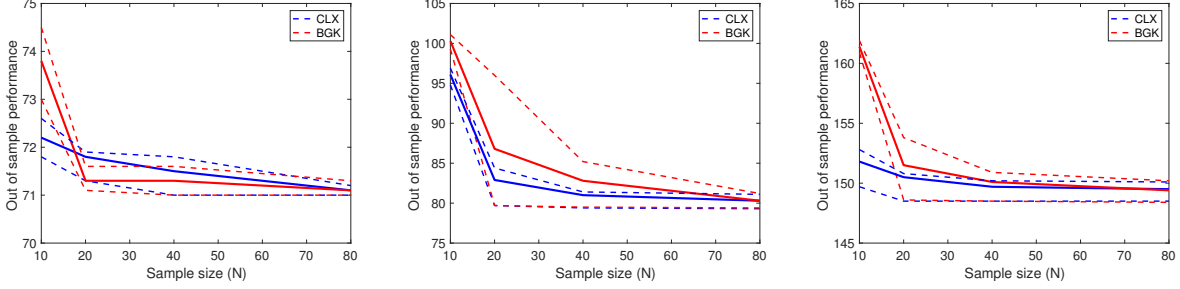


Figure 3: The out-of-sample performance of CLX and BGK on a newsvendor problem. The dash lines represent the 20th and 80th percentiles and the solid line represents the mean from 100 independent trials. Demand distribution: truncated normal (left), rescaled beta (middle), and truncated exponential (right).

5.2 Portofolio management

In this example, we consider the classical portfolio selection problem consisting of n assets in which the investor must divide the total budget to fractions $w = (w_1, w_2, \dots, w_n)$ with $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$, and invest w_i of the budget in the i^{th} security. We assume that the i^{th} security has a random future return ξ_i . The return from each unit of budget is thus $w^\top \xi$. We assume that the investor is risk-averse and measures the investment risk by the *conditional value at risk* (CVaR) of the return $w^\top \xi$; see [37]. Suppose the joint distribution of the return $\xi = (\xi_1, \xi_2, \dots, \xi_n)$ has a density function $p^*(\xi)$. The CVaR at level $\epsilon \in (0, 1)$ of the return of a portfolio with respect to a probability distribution p^* is defined as

$$\text{CVaR}_{p^*, \epsilon}(-w^\top \xi) \equiv \inf_{\beta \in \mathbb{R}} \mathbb{E}_{p^*} \left[\beta + \frac{1}{\epsilon} \left(-w^\top \xi - \beta \right)_+ \right] = \inf_{\beta \in \mathbb{R}} \int \left[\beta + \frac{1}{\epsilon} \left(-w^\top \xi - \beta \right)_+ \right] p^*(\xi) d\xi,$$

which represents the average of the $\epsilon \times 100\%$ worst portfolio losses (negative return) under distribution p^* . When p^* is known, we consider the case where the investor wants to minimize a weighted sum of the mean and the CVaR of the portfolio loss $-w^\top \xi$, which is formulated as the following stochastic optimization:

$$\begin{aligned} & \inf_{w_i \geq 0, \sum_{i=1}^n w_i = 1} \left\{ \mathbb{E}_{p^*}[-w^\top \xi] + \gamma \text{CVaR}_{p^*, \epsilon}(-w^\top \xi) \right\} \\ &= \inf_{\beta, w_i \geq 0, \sum_{i=1}^n w_i = 1} \int \left[\max \left\{ -w^\top \xi + \gamma\beta, -(1 + \gamma/\epsilon)w^\top \xi + \gamma(1 - 1/\epsilon) \right\} \right] p^*(\xi) d\xi \\ &= \inf_{x \in \mathcal{X}} \int f(x, \xi) p(\xi) d\xi. \end{aligned}$$

where $\gamma > 0$ indicates the investor's risk-aversion level, $\mathcal{X} = \{x = (w, \beta) \in \mathbb{R}^{n+1} | w_i \geq 0, \sum_{i=1}^n w_i = 1\}$, and $f(x, \xi) = \max \left\{ -w^\top \xi + \gamma\beta, -(1 + \gamma/\epsilon)w^\top \xi + \gamma(1 - 1/\epsilon) \right\}$.

If the joint distribution p^* of ξ is unknown but a collection of historical data return is collected, the investor can construct a data-driven ambiguity set of p^* and solve the DRO problem correspond-

ing to the stochastic optimization problem above. We construct the ambiguity set $\mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)$ in (19) and solve the DRO (7) to construct an portfolio. Then, we compare our solution with the one obtained by the DRO model in [13] where the ambiguity set is constructed using Wasserstein metric.

Following the numerical experiments by [13], we consider $n = 10$ assets with decomposable returns $\xi_i = \phi + \zeta_i$ for $i = 1, 2, \dots, 10$ where $\phi \sim \text{normal}(0, 2\%)$ is a systematic risk factor shared by all assets and $\zeta_i \sim \text{normal}(i \times 3\%, i \times 2.5\%)$ is an unsystematic risk factor associated with specific assets. By the construction, assets with higher indices promise higher mean returns at a higher risk. We set $\epsilon = 20\%$ and $\gamma = 10$ in our all experiments. We consider 6 different sample sizes, i.e., $N \in \{30, 60, 120, 240, 480, 960\}$. For each size, we randomly generate a dataset $\hat{\Xi}_N$ by i.i.d. sampling returns from the aforementioned distribution of ξ . Using $\hat{\Xi}_N$, we apply our approach and the method by [13] to construct the ambiguity sets and then solve the DRO to obtain an portfolio \hat{x} from each approach. To evaluate the out-of-sample performance of \hat{x} , we sample another i.i.d. dataset $\{\xi'_i\}_{i=1}^{N_{\text{large}}}$ with $N_{\text{large}} = 100,000$ from the true distribution and calculate the sample average approximation of the expected cost, i.e., $\frac{1}{N_{\text{large}}} \sum_{i=1}^{N_{\text{large}}} f(\hat{x}, \xi'_i)$ with \hat{x} from each approach. We repeat this procedure 100 times to show the mean and variation of the out-of-sample performance.

When constructing $\mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)$, we choose $\mathcal{K}(\xi) = \kappa(\|\xi\|_2)$ with κ being the boxcar kernel, namely, $\kappa(z) = Q$ if $z \in [0, 1]$ and $\kappa(z) = 0$ otherwise. Here, Q is a normalization constant that ensures $\int_{\mathbb{R}^m} \mathcal{K}(\xi) d\xi = 1$. Similar to $\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)$, the construction of $\mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)$ in our method requires some quantities which are hard to estimate (e.g. C and ρ). Therefore, we construct $\mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)$ using l_α^{KDE} and u_α^{KDE} in the form (17) with the parameters δ and h selected by the holdout validation method rather than their theoretical values in (18) and Proposition 3. In particular, we set $h = c \left(\frac{\log(N)}{N} \right)^{1/(2+m)}$ (see Proposition 3) and then select c from $\{0.02, 0.04, 0.06, 0.08, 0.1\}$ and δ from $\{0.02, 0.04, 0.06, 0.08, 0.1\}$. We randomly partition $\hat{\Xi}_N$ into $\hat{\Xi}_{\text{train}}$ and $\hat{\Xi}_{\text{test}}$ with $|\hat{\Xi}_{\text{train}}| = 0.7N$ and $|\hat{\Xi}_{\text{test}}| = 0.3N$. Given a combination of c and δ , we first construct $\mathcal{D}^{KDE}(\hat{\Xi}_{\text{train}})$ using $K = K(0.7N, c)$, and then solve $x_{c,\delta}^* \in \arg\min_{x \in \mathcal{X}} \sup_{p \in \mathcal{D}^{KDE}(\hat{\Xi}_{\text{train}})} \int_{\Xi} f(x, \xi) p(\xi) d\xi$ using Algorithm 3. Then, we select the combination of c and δ with the largest $\frac{1}{0.3N} \sum_{\hat{\xi} \in \hat{\Xi}_{\text{test}}} f(x_{c,\delta}^*, \hat{\xi})$. For each of the 100 independent trial, we repeat this process to select c and α . For a fair comparison, we also apply the same validation scheme to choose the radius of the Wasserstein ball used to construct the ambiguity set in [13].

We denote our approach by CLX and the method in [13] by WASS and plot the numerical results in Figure 4. In particular, for each of the 100 datasets in each sample-size scenario, we evaluate the solutions from both CLX and Wasserstein by testing their out-of-sample performances over the large dataset. We then plot the 20th percentile, the mean, and 80th of the out-of-sample performances of both approaches over the 6 sample-size scenarios. In figure 4, the blue lines show the results from CLX, while the red lines show the results from Wasserstein. Figure 4 indicates that both approaches converge to the true expectation with the increase in sample size.

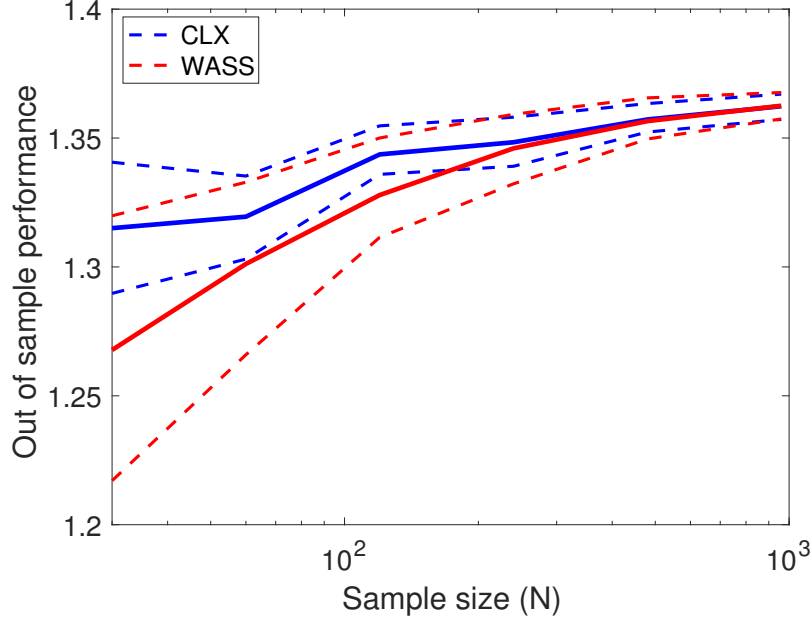


Figure 4: The out-of-sample performance of CLX and WASS on a portfolio selection problem. The dash lines represent the 20th and 80th percentiles and the solid line represents the mean from 100 independent trials.

6 Conclusions

In this paper, we proposed data-driven approaches to construct ambiguity sets that consist of continuous probability density functions. The ambiguity sets enjoy both finite and asymptotic convergences. The resulting distributionally robust optimization problem has infinite many variables and constraints. We then proposed a stochastic gradient decent method to solve the optimization problems. Numerical experiments in newsvendor problem and portfolio selection problem verify the effectiveness of our approach.

Appendix

Proof of Theorem 2

For $\delta > 0$, we define a subset of Ξ as $\Xi_\delta := \{\xi \in \Xi | p^*(\xi) > \delta\}$. We first analyze the approximation error between u_α , l_α and p^* at a given point $\xi \in \Xi_\delta$. We assume $\xi > \mu$ first and the proof for $\xi < \mu$ is similar. For simplicity of notations, we use c^+ and c^- to represent $c^-(\alpha)$ and $c^+(\alpha)$ in this proof.

There exists an index i such that $\hat{\xi}_{(k_i)} < \xi < \hat{\xi}_{(k_{i+1})}$. Since $\xi > \mu$ and p^* is (C, ρ) -Holder continuous, for a sufficiently large sample size N , we will further have $p^*(\hat{\xi}_{(k_{i+1})}) \geq \frac{\delta}{2}$ (as $\hat{\xi}_{(k_{i+1})}$ is close enough to ξ) and $\mu < \hat{\xi}_{(k_{i-1})} < \hat{\xi}_{(k_i)} < \xi < \hat{\xi}_{(k_{i+1})}$. Note that p^* is monotonically decreasing over $[\hat{\xi}_{(k_i)}, \hat{\xi}_{(k_{i+1})}]$. By the definitions of $l_\alpha^{SR}(\xi)$ and $u_\alpha^{SR}(\xi)$ as in (14) and the constraints in (13),

we must have $u \leq \frac{c^+}{\hat{\xi}_{(k_i)} - \hat{\xi}_{(k_{i-1})}}$ and $l \geq \frac{c^-}{\hat{\xi}_{(k_{i+1})} - \hat{\xi}_{(k_i)}}$, which implies

$$u_\alpha^{SR}(\xi) - l_\alpha^{SR}(\xi) \leq D_\xi \equiv \frac{c^+}{\hat{\xi}_{(k_i)} - \hat{\xi}_{(k_{i-1})}} - \frac{c^-}{\hat{\xi}_{(k_{i+1})} - \hat{\xi}_{(k_i)}} \quad (28)$$

Using the monotonicity of p^* , we can show that

$$(\hat{\xi}_{(k_i)} - \hat{\xi}_{(k_{i-1})})p^*(\hat{\xi}_{(k_i)}) \leq \Delta_i = F^*(\hat{\xi}_{(k_i)}) - F^*(\hat{\xi}_{(k_{i-1})}) \leq (\hat{\xi}_{(k_i)} - \hat{\xi}_{(k_{i-1})})p^*(\hat{\xi}_{(k_{i-1})}) \quad (29)$$

$$(\hat{\xi}_{(k_{i+1})} - \hat{\xi}_{(k_i)})p^*(\hat{\xi}_{(k_{i+1})}) \leq \Delta_{i+1} = F^*(\hat{\xi}_{(k_{i+1})}) - F^*(\hat{\xi}_{(k_i)}) \leq (\hat{\xi}_{(k_{i+1})} - \hat{\xi}_{(k_i)})p^*(\hat{\xi}_{(k_i)}). \quad (30)$$

Applying (29) and (30) to (28), we obtain

$$\begin{aligned} D_\xi &\leq \left| \frac{c^+ p^*(\hat{\xi}_{(k_{i-1})})}{\Delta_i} - \frac{c^- p^*(\hat{\xi}_{(k_{i+1})})}{\Delta_{i+1}} \right| \\ &\leq p^*(\hat{\xi}_{(k_{i+1})}) \left| \frac{c^+}{\Delta_i} - \frac{c^-}{\Delta_{i+1}} \right| + \frac{c^+}{\Delta_i} \left| p^*(\hat{\xi}_{(k_{i+1})}) - p^*(\hat{\xi}_{(k_{i-1})}) \right| \\ &\leq U \left| \frac{c^+}{\Delta_i} - \frac{c^-}{\Delta_{i+1}} \right| + \frac{Cc^+}{\Delta_i} \left| \hat{\xi}_{(k_{i+1})} - \hat{\xi}_{(k_{i-1})} \right|^\rho \\ &\leq U \left| \frac{c^+}{\Delta_i} - \frac{c^-}{\Delta_{i+1}} \right| + \frac{Cc^+}{\Delta_i} \left| \frac{\Delta_i + \Delta_{i+1}}{p^*(\hat{\xi}_{(k_{i+1})})} \right|^\rho \\ &\leq U \left| \frac{c^+}{\Delta_i} - \frac{c^-}{\Delta_{i+1}} \right| + \frac{Cc^+}{\Delta_i} \left| \frac{\Delta_i + \Delta_{i+1}}{\delta/2} \right|^\rho \end{aligned} \quad (31)$$

According to equation (98) and (101) with $(\tau = 2)$ in [19], we have

$$\lim_{N \rightarrow +\infty} \mathbb{P} \left(\left| \frac{c^+}{\Delta_j} - \frac{c^-}{\Delta_{j+1}} \right| \leq 4\sqrt{\frac{\log(N/K)}{K}} \right) \geq 1 - 2\theta \quad (32)$$

and

$$\lim_{N \rightarrow +\infty} \mathbb{P} \left(\frac{c^+}{\Delta_j} (\Delta_j + \Delta_{j+1})^\rho \leq 2 \left(\frac{K}{N} \right)^\rho \right) \geq 1 - 2\theta \quad (33)$$

for any $\theta \in (0, 1)$ and any $j = 0, 1, \dots, M$ (not necessarily i). Recall that $\xi \in \Xi_\delta$ so that $p^*(\xi) > \delta$.

We apply (32) and (33) to (31) to achieve

$$\lim_{N \rightarrow +\infty} \mathbb{P} \left(\max_{\xi \in \Xi_\delta} D_\xi \leq 4U\sqrt{\frac{\log(N/K)}{K}} + \frac{2^{\rho+1}C}{\delta^\rho} \left(\frac{K}{N} \right)^\rho \right) \geq 1 - 4\theta \quad (34)$$

Let $\mathcal{M}(\cdot)$ represent the Lebesgue measure on Ξ . We then choose δ to be small enough such that

$2 \max_{x \in \mathcal{X}, \xi \in \Xi} |f(x, \xi)| U\mathcal{M}(\Xi \setminus \Xi_\delta) \leq \frac{\epsilon}{2}$. Consider a fixed $x \in \mathcal{X}$, we have

$$\begin{aligned} & \max_{x \in \mathcal{X}} \left| \int_{\Xi} f(x, \xi) p^*(\xi) d\xi - \int_{\Xi} f(x, \xi) u_\alpha^{SR}(\xi) d\xi \right| \\ & \leq \max_{x \in \mathcal{X}} \left| \int_{\Xi_\delta} f(x, \xi) p^*(\xi) d\xi - \int_{\Xi_\delta} f(x, \xi) u_\alpha^{SR}(\xi) d\xi \right| + \max_{x \in \mathcal{X}} \left| \int_{\Xi \setminus \Xi_\delta} f(x, \xi) p^*(\xi) d\xi - \int_{\Xi \setminus \Xi_\delta} f(x, \xi) u_\alpha^{SR}(\xi) d\xi \right| \\ & \leq \max_{x \in \mathcal{X}, \xi \in \Xi} |f(x, \xi)| \max_{\xi \in \Xi_\delta} D_\xi + 2 \max_{x \in \mathcal{X}, \xi \in \Xi} |f(x, \xi)| U\mathcal{M}(\Xi \setminus \Xi_\delta) \end{aligned}$$

which, according to (34), implies

$$\lim_{N \rightarrow +\infty} \mathbb{P} \left(\begin{array}{l} \max_{x \in \mathcal{X}} \left| \int_{\Xi} f(x, \xi) p^*(\xi) d\xi - \int_{\Xi} f(x, \xi) u_\alpha(\xi) d\xi \right| \\ \leq \max_{x \in \mathcal{X}, \xi \in \Xi} |f(x, \xi)| \left[4U \sqrt{\frac{\log(N/K)}{K}} + \frac{2^{\rho+1}C}{\delta^\rho} \left(\frac{K}{N}\right)^\rho \right] + \frac{\epsilon}{2} \end{array} \right) \geq 1 - 4\theta$$

Similarly, we can also show that

$$\lim_{N \rightarrow +\infty} \mathbb{P} \left(\begin{array}{l} \max_{x \in \mathcal{X}} \left| \int_{\Xi} f(x, \xi) p^*(\xi) d\xi - \int_{\Xi} f(x, \xi) l_\alpha(\xi) d\xi \right| \\ \leq \max_{x \in \mathcal{X}, \xi \in \Xi} |f(x, \xi)| \left[4U \sqrt{\frac{\log(N/K)}{K}} + \frac{2^{\rho+1}C}{\delta^\rho} \left(\frac{K}{N}\right)^\rho \right] + \frac{\epsilon}{2} \end{array} \right) \geq 1 - 4\theta$$

Therefore, for any ϵ and θ , there exists an $N_{\epsilon, \theta}$ such that for any $N \geq N_{\epsilon, \theta}$,

$$\mathbb{P} \left(\max_{x \in \mathcal{X}} \left| \int_{\Xi} f(x, \xi) p^*(\xi) d\xi - \sup_{p \in \mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)} \int_{\Xi} f(x, \xi) p(\xi) d\xi \right| \leq \epsilon \right) \geq 1 - 10\theta.$$

Then, we have proof the first conclusion. The second conclusion can be easily implied from the first conclusion.

Proof of Theorem 4

Suppose $p^* \in \mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)$. We have

$$\begin{aligned} & \max_{x \in \mathcal{X}} \left| \int_{\Xi} f(x, \xi) p^*(\xi) d\xi - \sup_{p \in \mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)} \int_{\Xi} f(x, \xi) p(\xi) d\xi \right| \\ & \leq \max_{x \in \mathcal{X}} \int_{\Xi} |f(x, \xi)| |u_\alpha^{KDE}(\xi) - l_\alpha^{KDE}(\xi)| d\xi \\ & \leq \max_{x \in \mathcal{X}} \int_{\Xi} |f(x, \xi)| 2\delta d\xi \leq 2\delta \max_{x \in \mathcal{X}, \xi \in \Xi} |f(x, \xi)| \mathcal{M}(\Xi) \end{aligned}$$

where δ is defined in (18). By the definition of δ , there exists $N_{\epsilon, \alpha}$ such that for $N > N_{\epsilon, \alpha}$, we have $2\delta \max_{x \in \mathcal{X}, \xi \in \Xi} |f(x, \xi)| \mathcal{M}(\Xi) \leq \epsilon$. As a result of Theorem 3, for $N > N_{\epsilon, \alpha}$, we have

$$\begin{aligned} & \mathbb{P} \left(\max_{x \in \mathcal{X}} \left| \int_{\Xi} f(x, \xi) p^*(\xi) d\xi - \sup_{p \in \mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)} \int_{\Xi} f(x, \xi) p(\xi) d\xi \right| \leq \epsilon \right) \\ & \geq \mathbb{P} \left(p^* \in \mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha) \right) \geq 1 - \alpha \end{aligned}$$

The second conclusion can be easily implied from the first conclusion.

Acknowledgements

The authors would like to thank Aditya Guntuboyina for referring us to the papers about confidence band constructions.

References

- [1] A. Ben-Tal, D. Den Hertog, A. De Waegenare, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.
- [2] D. Bertsimas, V. Gupta, and N. Kallus. Robust sample average approximation. *Mathematical Programming*, pages 1–66, 2014.
- [3] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.
- [4] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.
- [5] J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.
- [6] G. C. Calafiore and L. El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.
- [7] Z. Chen, M. Sim, and H. Xu. Distributionally robust optimization with infinitely constrained ambiguity sets. Working Paper, 2016.
- [8] E. de Klerk, D. Kuhn, and K. Postek. Distributionally robust optimization with polynomial densities: theory, models and algorithms. *arXiv preprint arXiv:1805.03588*, 2018.
- [9] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.
- [10] J. Duchi, P. Glynn, and H. Namkoong. Statistics of robust optimization: a generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.
- [11] J. Dupačová. The minimax approach to stochastic programming and an illustrative application. *Stochastics: An International Journal of Probability and Stochastic Processes*, 20(1):73–88, 1987.
- [12] E. Erdoğan and G. Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2006.
- [13] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

- [14] R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *ArXiv preprint arXiv:1712.06050*, 2017.
- [15] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.
- [16] L. E. Ghaoui, M. Oks, and F. Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations research*, 51(4):543–556, 2003.
- [17] G. A. Hanasusanto, D. Kuhn, S. W. Wallace, and S. Zymler. Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Mathematical Programming*, 152(1):1–32, 2015.
- [18] J. A. Hartigan and P. M. Hartigan. The dip test of unimodality. *The Annals of Statistics*, 13(1):70–84, 1985.
- [19] N. W. Hengartner and P. B. Stark. Finite-sample confidence envelopes for shape-restricted densities. *The Annals of Statistics*, 23(2):525–550, 1995.
- [20] Z. Hu and L. J. Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.
- [21] H. Jiang. Uniform convergence rates for kernel density estimation. In *International Conference on Machine Learning*, pages 1694–1703, 2017.
- [22] R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, pages 1–37, 2015.
- [23] R. Z. Khas’minskii. A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability and Its Applications*, 23(4):794–798, 1976.
- [24] D. Klabjan, D. Simchi-Levi, and M. Song. Robust stochastic lot-sizing by means of histograms. *Production and Operations Management*, 22(3):691–710, 2013.
- [25] H. Lam and C. Mottet. Tail analysis without parametric models: A worst-case perspective. *Operations Research*, 65(6):1696–1711, 2017.
- [26] H. Lam and E. Zhou. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, page To appear., 2018.
- [27] B. Li, R. Jiang, , and J. L. Mathieu. Ambiguous risk constraints with moment and unimodality information. *Mathematical Programming*, page To appear., 2018.
- [28] J. Lofberg. Yalmip: A toolbox for modeling and optimization in matlab. In *Computer Aided Control Systems Design, 2004 IEEE International Symposium on*, pages 284–289. IEEE, 2004.
- [29] W.-K. Mak, D. P. Morton, and R. K. Wood. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations research letters*, 24(1-2):47–56, 1999.
- [30] M. Mevissen, E. Ragnoli, and J. Y. Yu. Data-driven distributionally robust polynomial optimization. In *Advances in Neural Information Processing Systems*, pages 37–45, 2013.
- [31] K. Natarajan and C. P. Teo. On reduced semidefinite programs for second order moment bounds with applications. *Mathematical Programming*, 161(1):487–518, 2017.

- [32] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.
- [33] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.
- [34] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.
- [35] P. Qian, Z. Wang, and Z. Wen. A composite risk measure framework for decision making under uncertainty. *Journal of the Operations Research Society of China*, pages 1–26, 2018.
- [36] A. Rinaldo, L. Wasserman, et al. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010.
- [37] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.
- [38] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.
- [39] H. Scarf, K. Arrow, and S. Karlin. A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production*, 10:201–209, 1958.
- [40] A. Shapiro. On duality theory of conic linear problems. In *Semi-infinite programming*, pages 135–165. Springer, 2001.
- [41] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*, volume 16. SIAM, 2014.
- [42] A. B. Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.
- [43] L. Vandenberghe, S. Boyd, and K. Comanor. Generalized chebyshev bounds via semidefinite programming. *SIAM review*, 49(1):52–64, 2007.
- [44] Z. Wang, P. W. Glynn, and Y. Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261, 2016.
- [45] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.
- [46] S. Zymmler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, 2013.
- [47] S. Zymmler, D. Kuhn, and B. Rustem. Worst-case value at risk of nonlinear portfolios. *Management Science*, 59(1):172–188, 2013.