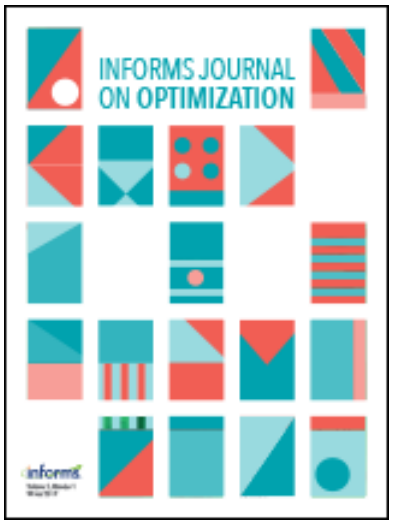


INFORMS Journal on Optimization

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Distributionally Robust Optimization with Confidence Bands for Probability Density Functions

Xi Chen, Qihang Lin, Guanglin Xu

To cite this article:
Xi Chen, Qihang Lin, Guanglin Xu (2021) Distributionally Robust Optimization with Confidence Bands for Probability Density Functions. INFORMS Journal on Optimization

Published online in Articles in Advance 06 Oct 2021

. https://doi.org/10.1287/ijoo.2021.0059

INFORMS JOURNAL ON OPTIMIZATION
Articles in Advance, pp. 1–25
ISSN 2575-1484 (print), ISSN 2575-1492 (online)
http://pubsonline.informs.org/journal/ijoo

Distributionally Robust Optimization with Confidence Bands for Probability Density Functions

Xi Chen,[a] **Qihang Lin,**[b] **Guanglin Xu**[c]

[a] Stein School of Business, New York University, New York City, New York 10012; [b] Tippie College of Business, University of Iowa, Iowa City, Iowa 52245; [c] Department of Systems Engineering and Engineering Management, University of North Carolina at Charlotte, Charlotte, North Carolina 28223

Contact: xichen@nyu.edu (XC); qihang-lin@uiowa.edu, https://orcid.org/0000-0003-2943-3267 (QL); guanglin.xu@uncc.edu, https://orcid.org/0000-0002-5786-4913 (GX)

Received: September 7, 2019
Revised: May 15, 2020; October 14, 2020; February 1, 2021
Accepted: March 18, 2021
Published Online in Articles in Advance: October 6, 2021

https://doi.org/10.1287/ijoo.2021.0059

Abstract. Distributionally robust optimization (DRO) has been introduced for solving stochastic programs in which the distribution of the random variables 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 contains 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. As examples, we consider the shape-restricted confidence bands and the confidence bands constructed with a kernel density estimation technique. The former allows us to incorporate the prior knowledge of the shape of the underlying density function (e.g., unimodality and monotonicity), and the latter enables us to handle multidimensional cases. Furthermore, we establish the convergence of the optimal value of DRO to that of the underlying stochastic program as the sample size increases. The DRO with our ambiguity set involves functional decision variables and infinitely many constraints. To address this challenge, we apply duality theory to reformulate the DRO to a finite-dimensional stochastic program, which is amenable to a stochastic subgradient scheme as a solution method.

Funding: X. Chen was supported by the National Science Foundation [Grant IIS-1845444].

Keywords: distributionally robust optimization • confidence band • data-driven ambiguity sets

1. Introduction

An important task in a *stochastic program* (SP) is to minimize the expectation of a cost function that depends on both decision variables and random variables. In particular, an SP is typically formulated as[1]

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

where $x \in \mathbb{R}^n$ is a decision variable, $\mathcal{X} \subseteq \mathbb{R}^n$, $\xi^\star \in \mathbb{R}^m$ is a random variable following distribution $P^\star$, and $f(x,\xi) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is a measurable cost function. SP has been actively studied in the past decades: see Birge and Louveaux (2011), Shapiro et al. (2014), and references therein. Suppose $\xi^\star$ is *absolutely continuous* so that it has a probability density function $p^\star : \mathbb{R}^m \to [0, +\infty)$ such that $P^\star(A) = \int_A p^\star(\xi) d\xi$ for any Borel set $A \subset \mathbb{R}^m$.[2] Then, we can rewrite (1) as

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

Despite their popularity, (1) and (2) are often challenging to solve becaue the distribution $P^\star$ or the density $p^\star$ is rarely known in real-life applications. When a set of historical data of $\xi^\star$ is collected, one may solve the approximation of (1) by replacing $P^\star$ with a distribution estimated from the data, for example, the empirical distribution. However, because of the approximation error, the decision made with the estimated distribution may be of inferior quality and, thus, may have an undesirable out-of-sample performance (Mak et al. 1999, Bertsimas et al. 2014, Mohajerin Esfahani and Kuhn 2018). An attractive alternative is the so-called *distributionally robust optimization* (DRO), in which one constructs an *ambiguity set* consisting of the distributions that are likely to be $P^\star$ and then minimizes the expected cost over the worst-case distribution from the

ambiguity set. In particular, letting $\mathcal{D}$ be an ambiguity set of probability measures, a DRO problem can be formulated as

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

Scarf (1958) first propose this model for a newsvendor problem, and it has been extensively studied since then by operations research and operations management communities.

In the literature, there exist several effective approaches to construct ambiguity sets, including the moment-, distance-, and hypothesis test–based approaches. We defer the detailed review of these approaches to Section 1.1. In many real-world applications, the random variable $\xi^\star$ is known to be absolutely continuous (e.g., when ξ models the prices of commodities or the returns of securities). However, the ambiguity sets constructed by most of the aforementioned approaches contain distributions that are not absolutely continuous. In fact, as shown in many papers (e.g., Bertsimas et al. 2014, Mohajerin Esfahani and Kuhn 2018), the worst-case distribution from those ambiguity sets is discrete. As a result, by solving (3) with those ambiguity sets, one may obtain a solution that is hedging against a discrete distribution which is never the true distribution. This phenomenon potentially leads to overconservative decisions made by DRO models.

In Section 3.3, we use a simple example to show that the ambiguity set constructed by the Wasserstein distance (Mohajerin Esfahani and Kuhn 2018) may lead to a worst-case expected cost *significantly higher* than the expected cost under the true distribution in the case where the true distribution is absolutely continuous. This follows from the fact that the ambiguity set constructed by the Wasserstein distance contains very conservative discrete distributions. To address such conservativeness, the *main goal* of this paper is to propose an approach to construct ambiguity sets that consist of only absolutely continuous distributions, which can potentially exclude those conservative discrete distributions and, thus, reduce the worst-case expected cost.

The ambiguity set we propose is constructed based on a *confidence band* of the true density function. We show in Section 3.3 with the aforementioned example that, using our ambiguity set, the worst-case expected cost can become *significantly closer* to the true expected cost than that using the ambiguity set constructed by the Wasserstein distance. Although what ambiguity set is the best in general remains an open question, this paper provides additional options for practitioners who rely on DRO to make decisions under uncertainty and need to estimate the worst-case cost under various perspectives to avoid overly conservative decisions.

With a little abuse of notation, in the rest of the paper, we still denote by $\mathcal{D}$ the ambiguity set that consists of density functions (instead of probability measures). With such an ambiguity set, the DRO model corresponding to the SP in (2) becomes

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

As a result, the worst-case distribution (if attainable) corresponding to the optimal solution of (4) is absolutely continuous.

The rest of the paper proceeds as follows. We review the related literature in Section 1.1 and summarize our contributions in Section 1.2. The mathematical notations needed are defined in Section 1.3. Then, we first propose the generic DRO in Section 2, followed by the construction of our data-driven ambiguity sets using density estimation techniques from the statistics literature. In particular, we 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 details 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 continuously many constraints. In Section 4, using 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 subgradient method approach as the solution method. Finally, we validate our approach with a newsvendor problem and a portfolio management problem in Section 5.

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 consists of all distributions that share the same (marginal and cross) moments; see Bertsimas and Popescu (2005), Calafiore and El Ghaoui (2006), Chen et al. (2019), de Klerk et al. (2019), Delage and Ye (2010), Dupačová (1987), Erdoğan and Iyengar (2006), El Ghaoui et al. (2003), Wiesemann et al. (2014), Hanasusanto et al. (2015), Li et al. (2018), Natarajan and Teo (2017), Vandenberghe et al. (2007), and Zymler et al. (2013a,b). A DRO problem in (3) with a moment-based ambiguity set can usually be reformulated into a tractable

conic program, for example, a second order cone program or a semidefinite program. However, the ambiguity set $\mathcal{D}$ is typically not guaranteed to converge to the true distribution $P^\star$ as the size of the historical data increases 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 metric to measure the distance between two probability distributions. In fact, such an ambiguity set can be considered as a ball centered at a reference distribution, for example, the empirical distribution, in the space of probability distributions. The distance metrics considered in the literature include Kullback–Leibler divergence (Hu and Hong 2013, Jiang and Guan 2015), ϕ-divergence (Ben-Tal et al. 2013, Klabjan et al. 2013, Duchi et al. 2021), the Prohorov metric (Erdoğan and Iyengar 2006), empirical Burg-entropy divergence (Lam and Mottet 2017, Lam 2019), and the Wasserstein metric (Pflug and Wozabal 2007, Gao and Kleywegt 2016, Gao et al. 2017, Mohajerin Esfahani and Kuhn 2018). Many distance-based ambiguity sets can guarantee asymptotic or finite-sample convergence to the true distribution.

• Hypothesis test–based ambiguity sets are proposed by Bertsimas et al. (2014, 2018). Based on a hypothesis test (e.g., a goodness-of-fit test), those approaches construct ambiguity sets consisting of the distributions that pass the hypothesis test with a given confidence level. According to Bertsimas et al. (2014), as the sample size increases, a hypothesis test–based ambiguity set can ensure that both the optimal objective value and the optimal solution of DRO asymptotically converge to those of the original SP if the hypothesis test has certain consistency property. The method that we propose in this paper belongs to this category.

• A likelihood-based approach is proposed in Wang et al. (2016) to construct an ambiguity set that consists of all distributions that make the likelihood of the historical data above a given threshold. It is shown by Wang et al. (2016) that, if such a threshold is appropriately chosen according to the data size, the ambiguity set converges to the true distribution as the data size increases to infinity.

Before our work, the ambiguity sets consisting of density functions were considered by Mevissen et al. (2013) and de Klerk et al. (2019). Their ambiguity sets contain only polynomial density functions, and our approach does not have this restriction. We also note that the ambiguity set considered in Mevissen et al. (2013) utilizes the kernel density estimation as does one of our ambiguity sets. Although their method must specify the Legendre polynomial series as density estimators, our method allows for using a broader family of kernel density estimations. Moreover, our ambiguity sets are constructed with data samples and, in the case of univariate $p^\star$, may integrate some shape information of the density function $p^\star$ (e.g., unimodality or monotonicity), and the ambiguity set in de Klerk et al. (2019) is not data-driven and integrates some moment information of $p^\star$. Other methods that integrate shape information on $p^\star$ include Lam and Mottet (2017) and Li et al. (2018). Li et al. (2018) consider an ambiguity set with moment and generalized unimodal constraints. Lam and Mottet (2017) impose convexity constraints on the tail of the density function. The main difference between our approach and theirs is that our method is data-driven and theirs do not require data samples and directly impose shape information as constraints in their optimization models.

1.2. Contributions

We summarize the main contributions of the paper as follows.

• We propose a class of ambiguity sets for DRO in which the true distribution of the random variables is known to be absolutely continuous. Our ambiguity set is constructed using a confidence band of the true density function and only contains absolutely continuous distributions. We use the shape-restricted confidence band by Hengartner and Stark (1995) and the confidence band constructed with kernel density estimators (Jiang 2017) as two examples. We further devise an example (see Section 3.3) to show that DRO with the proposed ambiguity set may lead to a less conservative worst-case expected loss by effectively excluding conservative discrete distributions. This suggests that the proposed ambiguity set can serve as an effective alternative to the existing ambiguity sets.

• It is shown by Hengartner and Stark (1995) and Jiang (2017), respectively, that the two confidence bands mentioned converge to the true density function as the data size increases to infinity (Theorem 1, Lemmas 1 and 2). Based on their results, we further show that the optimal values of the DRO problems using these two confidence bands converge to that of the original SP (Theorems 2 and 4). This new result may enrich the literature.

• As the DRO with the proposed ambiguity set contains a functional decision variable, it is challenging to optimize in general. Using the strong duality theory of conic programming on a Banach space (see, e.g., Shapiro 2001), we reformulate the DRO into a finite-dimensional convex program, which can be efficiently solved by using a stochastic subgradient method (Nemirovski et al. 2009).

1.3. Notation and Terminology

Let $\mathbb{Z}_+$ be the set of nonnegative integers. Let $\mathrm{Proj}_A(\cdot)$ denote the Euclidean projection operator on to the set A, that is, $\mathrm{Proj}_A(u) = \arg\min_{v\in A} \|v-u\|_2$. Given a set E, let $\mathbb{I}_E(\xi)$ be the indicator function that equals one when $\xi \in E$ and zero when $\xi \notin E$. Given an extended real-value function $g : \mathbb{R}^n \to [0, +\infty]$, we denote its epigraph, domain,

and subdifferential at x by $\mathrm{epi}(g)$, $\mathrm{dom}(g)$, and $\partial g(x)$, respectively, and use $g'(x)$ to denote a subgradient of g at x. In this paper, all random events to which $\mathbb{P}$ is applied are determined by independent and identically distributed (i.i.d.) random variables $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_N$, each of which has density $p^\star$. Hence, $\mathbb{P}$ is essentially a probability measure generated by the cross-product distribution $\underbrace{p^\star \times \cdots \times p^\star}_{N \text{ folds}}$.

2. Data-Driven Distributionally Robust Optimization

In this paper, we consider an ambiguity set for the true distribution $p^\star$ in (2) that consists of all density functions whose value is between two nonnegative functions constructed by historical data. We assume that there exists a set of N historical data points sampled from $p^\star$ and denoted by

$$\hat{\Xi}_N := \{\hat{\xi}_1, \ldots, \hat{\xi}_N\} \subseteq \Xi,$$

where $\Xi \subset \mathbb{R}^m$ is the support of $p^\star$, that is, $p^\star(\xi) = 0$ for any $\xi \notin \Xi$.[3] We assume Ξ is bounded. For a given $\alpha \in (0,1)$, we then construct two functions $l_\alpha : \Xi \to [0, +\infty]$ and $u_\alpha : \Xi \to [0, +\infty]$ based on $\hat{\Xi}_N$, α, and some shape information on $p^\star$ (e.g., unimodality and monotonicity) such that

$$\mathbb{P}\{l_\alpha(\xi) \le p^\star(\xi) \le u_\alpha(\xi), \ \forall \ \xi \in \Xi\} \ge 1 - \alpha. \tag{5}$$

Functions l_α and u_α are nonnegative but not necessarily density functions as they are not required to have an integral of one. Note that the randomness in (5) is due to $u_\alpha(\xi)$ and $l_\alpha(\xi)$, which depend on the random samples in $\hat{\Xi}_N$. We call the pair of functions (l_α, u_α) the confidence band for the density function $p^\star$ at a *confidence level* of $1 - \alpha$ and α is called the *significance level*. In Section 3, we introduce two different approaches from the literature on statistics to construct (l_α, u_α).

With (l_α, u_α) satisfying (5), we can construct an ambiguity set that contains $p^\star$ with a confidence level of $1 - \alpha$. More specifically, we define

$$\mathcal{L} := \{p | p : \Xi \to \mathbb{R} \text{ is a Lebesgue} - \text{measurable function on } \Xi\}$$

and consider the following ambiguity set:

$$\mathcal{D}(\hat{\Xi}_N, \alpha) := \left\{ p \in \mathcal{L} \middle| l_\alpha(\xi) \le p(\xi) \le u_\alpha(\xi), \ \forall \ \xi \in \Xi, \ \int_\Xi p(\xi)\, d\xi = 1 \right\}, \tag{6}$$

which satisfies $\mathbb{P}\left\{p^\star \in \mathcal{D}(\hat{\Xi}_N, \alpha)\right\} \ge 1 - \alpha$ according to (5). With $\mathcal{D}(\hat{\Xi}_N, \alpha)$ defined in (6), we can instantiate the DRO problem in (4) as

$$v^\star_{\mathcal{D}(\hat{\Xi}_N, \alpha)} := \inf_{x \in \mathcal{X}} \sup_{p \in \mathcal{D}(\hat{\Xi}_N, \alpha)} \int_\Xi f(x, \xi) p(\xi) d\xi. \tag{7}$$

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

Proposition 1. *Suppose that l_α and u_α in (6) satisfy (5) and that $v^\star_{\mathcal{D}(\hat{\Xi}_N, \alpha)}$ in (7) is finite and attained by $\hat{x}_N \in \mathcal{X}$. Let $\hat{v}_N := \mathbb{E}_{p^\star}[f(\hat{x}_N, \xi)]$. We have $\mathbb{P}\left\{v^\star_{\mathcal{D}(\hat{\Xi}_N, \alpha)} \ge \hat{v}_N\right\} \ge 1 - \alpha$.*

Proof. Whenever $p^\star \in \mathcal{D}(\hat{\Xi}_N, \alpha)$, we have

$$v^\star_{\mathcal{D}(\hat{\Xi}_N, \alpha)} = \sup_{p \in \mathcal{D}(\hat{\Xi}_N, \alpha)} \mathbb{E}_p[f(\hat{x}_N, \xi)] \ge \mathbb{E}_{p^\star}[f(\hat{x}_N, \xi)] = \hat{v}_N.$$

Hence, we have

$$\mathbb{P}\left\{v^\star_{\mathcal{D}(\hat{\Xi}_N, \alpha)} \ge \hat{v}_N\right\} \ge \mathbb{P}\left\{p^\star \in \mathcal{D}(\hat{\Xi}_N, \alpha)\right\} \ge 1 - \alpha$$

according to (5) and (6).

3. Data-Driven Ambiguity Sets

In this section, we present two existing methods from the literature on statistics to construct a confidence band for a density function based on observed data. The first method is applicable to only univariate distributions but can integrate some prior information on the shape of the true density function. The second one is applicable to multivariate distributions.

3.1. Shape-Restricted Confidence Bands

In this section, we assume $m = 1$, namely the random variable $\xi^\star$ is univariate, and present the method by Hengartner and Stark (1995) to construct a confidence band (l_α, u_α) for $p^\star$. Although this method only applies to a univariate density function, it is capable of incorporating some shape information about $p^\star$ (e.g., unimodality and monotonicity) into the construction of the ambiguity set that improves the convergence rate of the ambiguity set to the true density $p^\star$.

We need the following assumptions in this subsection.

Assumption 1 (For Shape-Restricted Confidence Bands).

The following statements hold:

A1. *Set $\Xi = [a, b]$ (the support of $p^\star$) with known a and b satisfying $-\infty < a < b < +\infty$.*

A2. *The true distribution $p^\star$ is unimodal with a known mode $\mu \in [a, b]$, meaning that $p^\star$ is monotonically increasing on $[a, \mu]$ and decreasing on $[\mu, b]$.*

A3. *There exists a known constant U such that $p(\xi) \le U$ 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 following derivation. In fact, the method by Hengartner and Stark (1995) can be generalized to the case in which $a = -\infty$ and/or $b = +\infty$. Moreover, for most applications, a conservative estimation of the range of ξ is usually available, which can be directly used as $[a, b]$. Similarly, any conservative estimation of the global upper bound of $p(\xi)$ can be used as U. We also note that Assumption 1(A2) covers the case in which $p^\star$ is known to be monotonically increasing ($\mu = b$) or decreasing ($\mu = a$). The statistical techniques for testing whether a set of data are sampled from a unimodal distribution are studied in literature (e.g., Hartigan and Hartigan 1985).

Next, we describe the method by Hengartner and Stark (1995) to construct a confidence band. Let $(\hat{\xi}_{(1)}, \ldots, \hat{\xi}_{(N)})$ be the order statistics of $\xi^\star$ constructed from $\hat{\Xi}_N$ so that $\hat{\xi}_{(1)} < \cdots < \hat{\xi}_{(N)}$. We choose a group size $K \in \mathbb{Z}_+$ satisfying $0 < K < N$ and partition the sequence $(\hat{\xi}_{(1)}, \ldots, \hat{\xi}_{(N)})$ into groups while keeping their order so that all groups have a size of K except the last group that might have a size less than K if N is not dividable by K. In fact, let

$$M' := \lfloor N/K \rfloor \text{ and } M := \lceil N/K \rceil$$

so that $M' = M$ if N is dividable by K and $M' = M - 1$ otherwise. Then, the size of the first M' groups is always K and only when $M' = M - 1$ is the size of the Mth (last) group $N - KM'$. Let $F^\star : \mathbb{R} \to [0, 1]$ be the cumulative density function of $\xi^\star$. We then define the following indexes:

$$k_i = \begin{cases} (i-1)K + 1, & i = 1, 2, \ldots, M' \\ N, & i = M \text{ if } M \neq M', \end{cases}$$

and the random variables

$$\Delta_i := F^\star(\hat{\xi}_{(k_i)}) - F^\star(\hat{\xi}_{(k_{(i-1)})}), \quad i = 2, 3, \ldots, M.$$

In Figure 1, we illustrate the construction of Δ_i corresponding to a beta distribution using some specific values of N, K, and i.

Given a significance level α, the construction of the confidence band by Hengartner and Stark (1995) requires computing two quantities, $c^-(\alpha)$ and $c^+(\alpha)$, that satisfy

$$\mathbb{P}\{c^-(\alpha) \le \Delta_i \le c^+(\alpha), i = 2, 3, \ldots, M\} \ge 1 - \alpha.$$

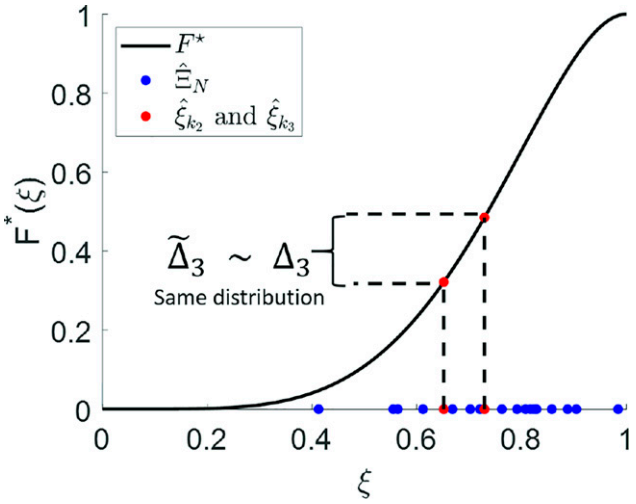
However, $c^-(\alpha)$ and $c^+(\alpha)$ are difficult to calculate directly using the definition of Δ_is because $F^\star$ is unknown. To address this issue, Hengartner and Stark (1995) construct random variable Δ_i that has the same distribution as Δ_i and can be easily simulated without any knowledge on $F^\star$. In fact, it is well known that $F^\star(\xi^\star)$ is a uniformly distributed random variable on $[0, 1]$. By theorem 6.6(b) and (c) in DasGupta (2019), the random vector

$$\left(F^\star(\hat{\xi}_{(1)}), F^\star(\hat{\xi}_{(2)}) - F^\star(\hat{\xi}_{(1)}), \ldots, F^\star(\hat{\xi}_{(N)}) - F^\star(\hat{\xi}_{(N-1)}), 1 - F^\star(\hat{\xi}_{(N)})\right)$$

has standard Dirichlet distribution, that is, uniform distribution in the N-dimensional simplex, and has the same distribution as

$$\left(\frac{X_1}{\sum_{j=1}^{N+1} X_j}, \frac{X_2}{\sum_{j=1}^{N+1} X_j}, \ldots, \frac{X_{N+1}}{\sum_{j=1}^{N+1} X_j}\right),$$

Figure 1. (Color online) We Generate $n = 20$ Samples from Beta Distribution Beta(5,2), Whose Cumulative Distribution Function $F^\star$ Is Plotted Here



Notes. We sort and split the samples into $M = M' = 5$ groups with four samples ($K = 4$) in each group. The value $\Delta_3 = F^\star(\hat{\xi}_{k_3}) - F^\star(\hat{\xi}_{k_2}) = F^\star(\hat{\xi}_9) - F^\star(\hat{\xi}_5)$ is chosen as an example and marked in the figure with samples $\hat{\xi}_{(k_2)} = \hat{\xi}_{(5)}$ and $\hat{\xi}_{(k_3)} = \hat{\xi}_{(9)}$ marked by the red dots. Here, Δ_3 has the same distribution as $\tilde{\Delta}_3$ defined in (8).

where $X_1, X_2, \ldots, X_{N+1}$ are i.i.d. exponential random variables with mean one. Because

$$\Delta_i = F^\star(\hat{\xi}_{(k_i)}) - F^\star(\hat{\xi}_{(k_{(i-1)})}) = \sum_{j=k_{(i-1)}}^{k_i-1} \left[F^\star(\hat{\xi}_{(j+1)}) - F^\star(\hat{\xi}_{(j)})\right],$$

it has the same distribution as

$$\tilde{\Delta}_i := \frac{\sum_{j=k_{(i-1)}+1}^{k_i} X_j}{\sum_{j=1}^{N+1} X_j}. \tag{8}$$

Different from Δ_is, the distributions of $\tilde{\Delta}_i$ s can be simulated by sampling X_is in (8). Through such simulation, we can estimate $c^-(\alpha)$ and $c^+(\alpha)$ with an arbitrary precision such that

$$\mathbb{P}\{c^-(\alpha) \le \Delta_i \le c^+(\alpha),\ i = 2,3,\ldots,M\} = \mathbb{P}\{c^-(\alpha) \le \tilde{\Delta}_i \le c^+(\alpha),\ i = 2,3,\ldots,M\} \ge 1 - \alpha.$$

Let $\mathcal{L}_\mu$ be the set of all density functions on Ξ with mode at μ, that is,

$$\mathcal{L}_\mu := \left\{p \in \mathcal{L} \middle| p(\xi) \ge 0,\ \forall\ \xi \in \Xi,\ \int_\Xi p(\xi)d\xi = 1,\ \text{the mode of } p \text{ is } \mu.\right\}$$

and let

$$\mathcal{D}_\mu(\hat{\Xi}, \alpha) := \left\{p \in \mathcal{L}_\mu \middle| c^-(\alpha) \le \int_{\hat{\xi}_{(k_{i-1})}}^{\hat{\xi}_{(k_i)}} p(\xi)d\xi \le c^+(\alpha),\ i = 2,3,\ldots,M\right\}. \tag{9}$$

Then, by the definitions of Δ_i, $\mathcal{D}_\mu(\hat{\Xi}, \alpha)$, $c^-(\alpha)$, and $c^+(\alpha)$, we have

$$\mathbb{P}\left\{p^\star \in \mathcal{D}_\mu(\hat{\Xi}, \alpha)\right\} \ge \mathbb{P}\{c^-(\alpha) \le \Delta_i \le c^+(\alpha),\ i = 2,3,\ldots,M\} \ge 1 - \alpha. \tag{10}$$

Given this property, $\mathcal{D}_\mu(\hat{\Xi}, \alpha)$ can be used as an ambiguity set $\mathcal{D}$ in (4). However, this ambiguity set may be too large to ensure a good solution from solving (4). Therefore, Hengartner and Stark (1995) further refine $\mathcal{D}_\mu(\hat{\Xi}, \alpha)$ with a confidence band based on the unmodality of $p^\star$.

The confidence band (l_α, u_α) by Hengartner and Stark (1995) has no analytical form in general, but $l_\alpha(\xi)$ and $u_\alpha(\xi)$ can be calculated numerically at any given ξ. Let $\xi \in [a,b]$ be the point at which we want to calculate $l_\alpha(\xi)$ and $u_\alpha(\xi)$. Let $\hat{N} \in \mathbb{Z}_+$ be the number of distinct elements in the set $\{a, b, \mu, \xi, \hat{\xi}_{(k_1)}, \ldots, \hat{\xi}_{(k_M)}\}$ and z_j be the jth smallest element in this set, namely

$$\{z_j\}_{j=1\hat{N}} = \{a, b, \mu, \xi, \hat{\xi}_{(k_1)}, \ldots, \hat{\xi}_{(k_M)}\} \text{ and } z_1 = a \le z_2 \le \cdots \le z_{\hat{N}} = b.$$

Note that $M \le \hat{N} \le M+4$ as a, b, μ, and ξ may coincide with each other and with some $\hat{\xi}_{(k_i)}$. Then, we consider the following two sets of nonnegative step functions on $[a,b]$:

$$p(\cdot) \in \mathcal{D}_\mu^-(\hat{\Xi}, \xi) := \left\{ p \in \mathcal{L} \,\middle|\, \begin{array}{rcl} p(\cdot) & = & \displaystyle\sum_{\{j: z_{j+1} < \mu\}} \beta_j \mathbb{I}_{(z_j, z_{j+1}]}(\cdot) + \Big(\max_{\{j: \mu \in [z_j, z_{j+1}]\}} \beta_j\Big) \mathbb{I}_\mu(\cdot) \\ & & \displaystyle + \sum_{\{j: \mu \in [z_j, z_{j+1}]\}} \beta_j \mathbb{I}_{(z_j, z_{j+1})}(\cdot) + \sum_{\{j: z_j > \mu\}} \beta_j \mathbb{I}_{[z_j, z_{j+1})}(\cdot), \\ & & \text{where } \beta_j \in [0, U] \text{ for } j = 1, 2, \ldots, \hat{N}-1. \end{array} \right\} \tag{11}$$

$$p(\cdot) \in \mathcal{D}_\mu^+(\hat{\Xi}, \xi) := \left\{ p \in \mathcal{L} \,\middle|\, \begin{array}{rcl} p(\cdot) & = & \displaystyle\sum_{\{j: z_{j+1} \le \mu\}} \beta_j \mathbb{I}_{[z_j, z_{j+1})}(\cdot) + U \cdot \mathbb{I}_\mu(\xi) + \sum_{\{j: z_j \ge \mu\}} \beta_j \mathbb{I}_{(z_j, z_{j+1}]}(\cdot), \\ & & \text{where } \beta_j \in [0, U] \text{ for } j = 1, 2, \ldots, \hat{N}-1. \end{array} \right\}, \tag{12}$$

where $\mathbb{I}_E(\cdot)$ denotes the indicator function of the set E. In spite of the sophisticated form of the functions they contain, the sets $\mathcal{D}_\mu^-(\hat{\Xi}, \xi)$ and $\mathcal{D}_\mu^+(\hat{\Xi}, \xi)$ essentially contain all nonnegative step functions on $[a,b]$ with $\hat{N}-1$ pieces and break points at $\{z_j\}_{j=1\hat{N}}$. Given any $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{\hat{N}-1}) \in \mathbb{R}_+^{\hat{N}-1}$, one can construct a step function by setting the function value to be β_i in the ith piece. Here, we require $\beta_j \le U$ because of the known upper bound U of $p^\star$ (see Assumption 1(A3)). The only difference between the functions in $\mathcal{D}_\mu^-(\hat{\Xi}, \xi)$ and $\mathcal{D}_\mu^+(\hat{\Xi}, \xi)$ is how their values are determined at the break points $\{z_j\}_{j=1\hat{N}}$. In particular, each function in $\mathcal{D}_\mu^-(\hat{\Xi}, \xi)$ is left-continuous on $[a, \mu)$ and right-continuous on $(\mu, b]$, and its value at μ is the larger one of the two pieces adjacent to μ. On the contrary, each function in $\mathcal{D}_\mu^+(\hat{\Xi}, \xi)$ is right-continuous on $[a, \mu)$ and left-continuous on $(\mu, b]$, and its value at μ is always U. In Figure 2, we show some examples of the curves of the functions from $\mathcal{D}_\mu^-(\hat{\Xi}, \xi)$ and $\mathcal{D}_\mu^+(\hat{\Xi}, \xi)$. In this figure, a solid break point means that the end point is included in the piece, and a hollow break point means not included. The height of the jth piece is β_j.

Then, the confidence band by Hengartner and Stark (1995) is calculated at ξ as

$$l_\alpha^{SR}(\xi) := \inf_{p \in \mathcal{D}_\mu^-(\hat{\Xi}, \xi) \cap \mathcal{D}_\mu(\hat{\Xi}, \alpha)} p(\xi) \quad \text{and} \quad u_\alpha^{SR}(\xi) := \sup_{p \in \mathcal{D}_\mu^+(\hat{\Xi}, \xi) \cap \mathcal{D}_\mu(\hat{\Xi}, \alpha)} p(\xi). \tag{13}$$

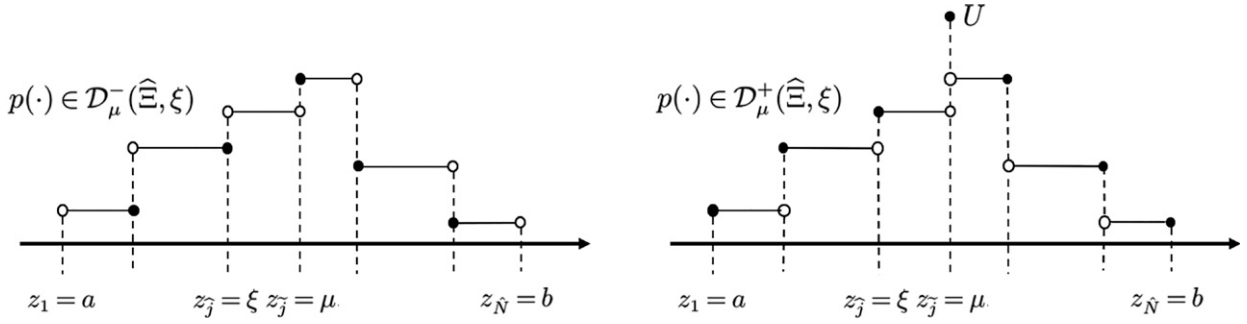
Here, the superscript "SR" refers to "shape restricted." By its definition, the value of $l_\alpha^{SR}(\xi)$ (respectively, $u_\alpha^{SR}(\xi)$) equals the smallest (largest) value at ξ among all density functions that have the piecewise-constant form in (11) ((12)) and satisfy

$$c^-(\alpha) \le \int_{\hat{\xi}_{(k_{i-1})}}^{\hat{\xi}_{(k_i)}} p(\xi) d\xi \le c^+(\alpha) \text{ for } i = 2, 3, \ldots, M.$$

Finally, the ambiguity set based on the shape-restricted confidence band in (13) is defined as

$$\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha) := \left\{ p \in \mathcal{L} \,\middle|\, l_\alpha^{SR}(\xi) \le p(\xi) \le u_\alpha^{SR}(\xi), \ \forall \xi \in \Xi, \ \int_\Xi p(\xi)\, d\xi = 1 \right\}. \tag{14}$$

Figure 2. Examples of the Functions in $\mathcal{D}_\mu^-(\hat{\Xi}, \xi)$ and $\mathcal{D}_\mu^+(\hat{\Xi}, \xi)$ Defined in (11) and (12)



Notes. A solid break point means that the end point is included in the piece, and a hollow break point means not included. The height of the jth piece is β_j. We also mark the mode $z_{\tilde{j}} = \mu$ in the figures as well as possible locations of $z_{\hat{j}} = \xi$.

Remark 1. *In the original work by Hengartner and Stark (1995), the constant U in (11) and (12) is $+\infty$. In this paper, we require U to be finite so that $u_\alpha^{SR}(\xi) \le U < +\infty$ for any ξ, which is needed to establish Theorem 2. Requiring a finite U does not change any statistical property of $\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)$ we need (i.e., Theorem 1) to obtain the results in this paper.*

The following theorem is from theorem 3.1 and equation (12) in Hengartner and Stark (1995).

Theorem 1 (Hengartner and Stark 1995). *Suppose Assumption 1 holds. Then,*

$$\mathbb{P}\left\{p^\star \in \mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)\right\} \ge 1 - \alpha.$$

We then describe the numerical procedure for computing $l_\alpha^{SR}(\xi)$ and $u_\alpha^{SR}(\xi)$ for a given ξ. Recall that z_j denotes the jth smallest element in $\{a, b, \mu, \xi, \hat{\xi}_{(k_1)}, \ldots, \hat{\xi}_{(k_M)}\}$. Let $\hat{\jmath}$ and $\tilde{\jmath}$ be the indexes in $\{1, 2, \ldots, \hat{N}\}$ such that $z_{\hat{\jmath}} = \xi$ and $z_{\tilde{\jmath}} = \mu$. By (11) and (12), if a density function p belongs to either $\mathcal{D}_\mu^-(\hat{\Xi}, \xi)$ or $\mathcal{D}_\mu^+(\hat{\Xi}, \xi)$, the value $p(\xi)$ only depends on either $\beta_{\hat{\jmath}}$ or $\beta_{\hat{\jmath}-1}$. Hence, to solve (13), one only needs to maximize or minimize $\beta_{\hat{\jmath}}$ or $\beta_{\hat{\jmath}-1}$ subject to the appropriate constraints on $\boldsymbol{\beta}$ that restrict p in $\mathcal{D}_\mu^-(\hat{\Xi}, \xi) \cap \mathcal{D}_\mu(\hat{\Xi}, \alpha)$ or $\mathcal{D}_\mu^+(\hat{\Xi}, \xi) \cap \mathcal{D}_\mu(\hat{\Xi}, \alpha)$. Such constraints can be characterized as the following polyhedron:

$$\mathcal{H}(\hat{\Xi}_N, \alpha) = \Bigg\{\boldsymbol{\beta} \in \mathbb{R}_+^{\hat{N}-1} \mid \beta_1 \le \beta_2 \le \cdots \le \beta_{\tilde{\jmath}-1}, \quad \beta_{\tilde{\jmath}} \ge \beta_{\tilde{\jmath}+1} \ge \cdots \ge \beta_{\hat{N}-1} c^-(\alpha) \le \sum_{j:\hat{\xi}_{(k_{i-1})} \le z_j < \hat{\xi}_{(k_i)}} \beta_j (z_{j+1} - z_j) \le c^+(\alpha),$$
$$\forall i = 2, \ldots, M \sum_{j=1}^{\hat{N}-1} \beta_j (z_{j+1} - z_j) = 1, 0 \le \beta_j \le U, \quad \forall \ 1 \le j \le \hat{N} - 1\Bigg\}. \tag{15}$$

Here, the first line of the constraints in (15) ensures μ is the mode of p as p must be in $\mathcal{D}_\mu(\hat{\Xi}, \alpha) \subset \mathcal{L}_\mu$. The second line of the constraints in (15) is obtained by instantiating the condition

$$c^-(\alpha) \le \int_{\hat{\xi}_{(k_{i-1})}}^{\hat{\xi}_{(k_i)}} p(\xi) d\xi \le c^+(\alpha)$$

in (9) with the piecewise-constant function p in $\mathcal{D}_\mu^-(\hat{\Xi}, \xi)$ or $\mathcal{D}_\mu^+(\hat{\Xi}, \xi)$. The third line requires that p must be a density function, and the last line is from Assumption 1(A3). With this constraint set, $l_\alpha^{SR}(\xi)$ and $u_\alpha^{SR}(\xi)$ in (13) can be computed, respectively, by solving the following linear programs

$$l_\alpha^{SR}(\xi) = \begin{cases} \min\limits_{\boldsymbol{\beta} \in \mathcal{H}(\hat{\Xi}_N, \alpha)} \beta_{\hat{\jmath}-1}, & \text{if } \xi \le \mu, \\ \min\limits_{\boldsymbol{\beta} \in \mathcal{H}(\hat{\Xi}_N, \alpha)} \beta_{\hat{\jmath}}, & \text{if } \xi > \mu, \end{cases} \quad \text{and} \quad u_\alpha^{SR}(\xi) = \begin{cases} \max\limits_{\boldsymbol{\beta} \in \mathcal{H}(\hat{\Xi}_N, \alpha)} \beta_{\hat{\jmath}}, & \text{if } \xi \le \mu, \\ \max\limits_{\boldsymbol{\beta} \in \mathcal{H}(\hat{\Xi}_N, \alpha)} \beta_{\hat{\jmath}-1}, & \text{if } \xi > \mu. \end{cases} \tag{16}$$

We summarize this procedure for constructing a confidence band for a unimodal density function in Algorithm 1. According to Hengartner and Stark (1995), Algorithm 1 can also be extended for constructing a confidence band when the mode of $p^\star$ is not known exactly but only known to be in an interval $[\mu^+, \mu^-] \subset [a, b]$. The convergence rate of the constructed confidence band is given in the following lemma, which is a paraphrase of equation (104) in Hengartner and Stark (1995).[4]

Algorithm 1 (Shape-Restricted Confidence Band $(l_\alpha^{SR}(\xi), u_\alpha^{SR}(\xi))$ at $\xi \in \Xi$)

Input: Data $\hat{\Xi}_N = \{\hat{\xi}^1, \ldots, \hat{\xi}^N\}$ sampled from $p^\star$, an interval $\Xi = [a, b]$, the mode μ of $p^\star$, a constant $U \ge p^\star(\xi)$ for any $\xi \in \Xi$, a significance level $\alpha \in (0, 1)$, a group size K with $0 < K < N$ and a targeted point $\xi \in \Xi$.

1: Let $M' := \lfloor N/K \rfloor$ and $M := \lceil N/K \rceil$.

2: Let $\hat{\xi}_{(1)}, \ldots, \hat{\xi}_{(N)}$ be the order statistics of $\hat{\Xi}_N$ with $\hat{\xi}_{(1)} < \cdots < \hat{\xi}_{(N)}$.

3: Let $k_i := \begin{cases} (i-1)K + 1 & \text{if } i = 1, 2, \ldots, M' \\ N & \text{if } i = M \ne M' \end{cases}$.

4: Let $X_1, X_2, \ldots, X_{N+1}$ be i.i.d. exponential random variables with mean one and define

$$\tilde{\Delta}_i := \frac{\sum_{j=k_{(i-1)}+1}^{k_i} X_j}{\sum_{j=1}^{N+1} X_j}, \text{ for } i = 2, 3, \ldots, M.$$

5: Generate a sufficient number of samples of $\tilde{\Delta}_i$ s and use them to estimate constants $c^-(\alpha)$ and $c^+(\alpha)$ that satisfy $\mathbb{P}\{c^-(\alpha) \le \tilde{\Delta}_i \le c^+(\alpha), i = 2, 3, \ldots, M\} \ge 1 - \alpha$.

6: Let z_j be the jth smallest value in the set $\{a, \mu, \xi, \hat{\xi}_{(k_1)}, \ldots, \hat{\xi}_{(k_M)}, b\}$ for $j = 1, \ldots, \hat{N}$, where $\hat{N}$ represents the number of distinct elements in that set.

7: Let $\hat{j}$ and $\tilde{j}$ be the indexes in $\{1,2,\ldots,\hat{N}\}$ such that $z_{\hat{j}} = \xi$ and $z_{\tilde{j}} = \mu$. Define the polyhedron $\mathcal{H}(\hat{\Xi}_N, \alpha)$ in (15).
8: Compute $(l_\alpha^{SR}(\xi), u_\alpha^{SR}(\xi))$ by solving the corresponding linear programs in (16).
Output: $(l_\alpha^{SR}(\xi), u_\alpha^{SR}(\xi))$.

Lemma 1 (Hengartner and Stark 1995).
Suppose Assumption 1 holds and $p^\star$ is (C,ρ)-Hölder continuous for constants $C > 0$ and $\rho > 0$, that is, $| p^\star(\xi') - p^\star(\xi) | \le C | \xi' - \xi |^\rho$ for any ξ and ξ' in Ξ. If $K = \lceil B\left(N^{2\rho}\log N\right)^{1/(1+2\rho)} \rceil$ in Algorithm 1 for a constant $B > 0$, we have

$$\lim_{N\to\infty} \mathbb{P}\left\{\left|u_\alpha^{SR}(\xi) - l_\alpha^{SR}(\xi)\right| \le 4p^\star(\xi)\left(\sqrt{\frac{3+2\rho}{1+2\rho}}B^{-1/2} + C(p^\star(\xi))^{-(\rho+1)}B^\rho\right)\left(\frac{\log N}{N}\right)^{\rho/(1+2\rho)}\right\} = 1. \tag{17}$$

This is for any $\xi \in \Xi$ except μ.

The width of the confidence band at ξ, that is, $| u_\alpha^{SR}(\xi) - l_\alpha^{SR}(\xi) |$, converges to zero at a rate $O\left((\log N)/N)^{\rho/(1+2\rho)}\right)$ according to Lemma 1. According to the lower bound in Khas'minskii (1976), this convergence rate is optimal (up to a constant factor) for a confidence band of a unimodal density. It is also worthwhile to note that the rate of convergence from Kolmogorov–Smirnov distance is slower as compared with this approach. As shown in Hartigan and Hartigan (1985), the confidence band for the cumulative density function formed by the Kolmogorov–Smirnov distance is only $O\left((1/N)^{1/4}\right)$, which is slower than the rate of $O\left((\log N/N)^{1/3}\right)$ in Lemma 1 when $\rho = 1$.

Lemma 1 shows the pointwise convergence of the confidence band to the true distribution $p^\star$. In the next theorem, we show that, under additional assumptions to Lemma 1, we can characterize the convergence of $v^\star_{\mathcal{D}^{SR}(\hat{\Xi}_N,\alpha)}$ defined in (7) to $v^\star$ in (2).

Theorem 2. *Suppose the assumptions in Lemma 1 hold and $\max_{x\in\mathcal{X},\xi\in\Xi}|f(x,\xi)| < +\infty$. For any $\epsilon > 0$, we have*

$$\lim_{N\to+\infty} \mathbb{P}\left\{\max_{x\in\mathcal{X}}\left|\int_\Xi f(x,\xi)p^\star(\xi)d\xi - \sup_{p\in\mathcal{D}^{SR}(\hat{\Xi}_N,\alpha)}\int_\Xi f(x,\xi)p(\xi)d\xi\right| \le \epsilon\right\} \ge 1-\alpha$$

and $\lim_{N\to+\infty}\mathbb{P}\left\{\left|v^\star_{\mathcal{D}^{SR}(\hat{\Xi}_N,\alpha)} - v^\star\right| \le \epsilon\right\} \ge 1-\alpha$.

Proof. See Appendix A.

We present some examples of the confidence bands constructed by Algorithm 1 with various values of α and N in Figures 3 and 4 in which the true distribution $p^\star$ is chosen to be a scaled beta distribution and a truncated exponential distribution, respectively. The spikes of the upper bands near mode μ in all figures are explained in Appendix E.

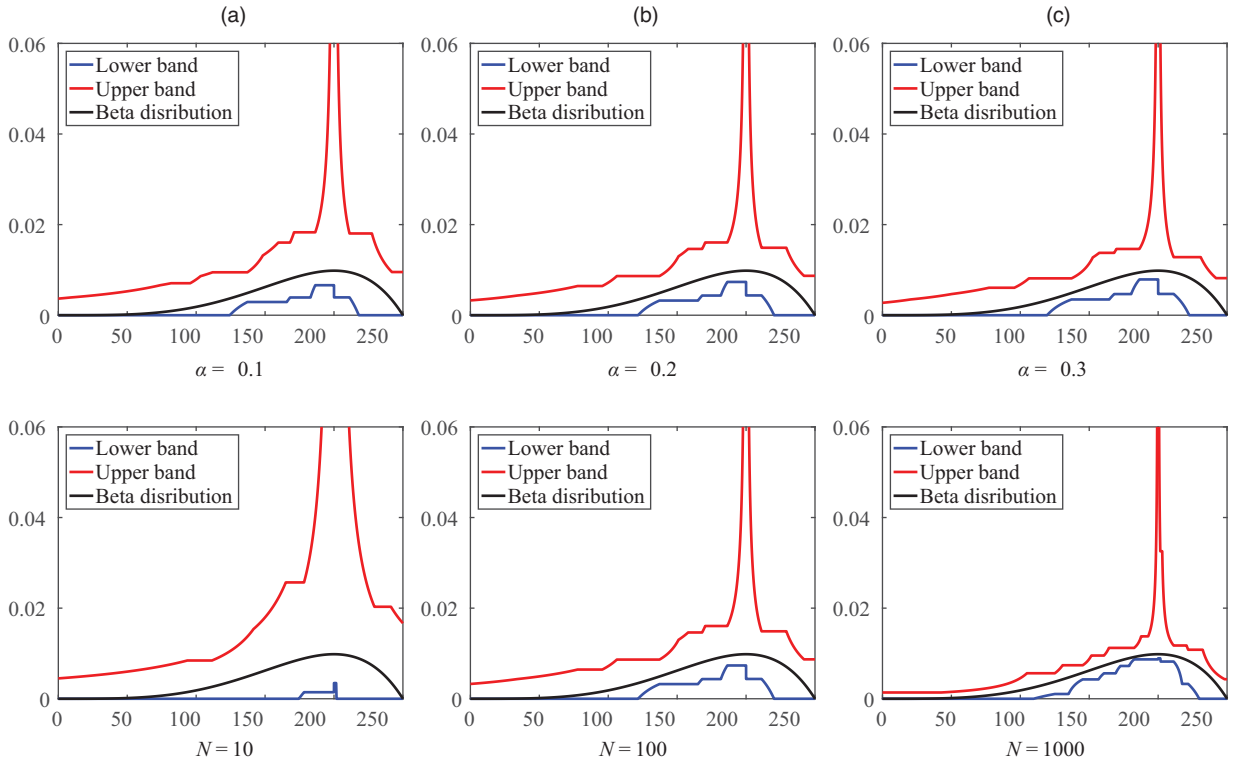
3.2. Kernel Density Estimation Confidence Bands

The shape-restricted confidence bands described in the previous section can only be applied to a univariate density function. In this section, we describe a method to construct a confidence band for a multivariate density function based on the classical *kernel density estimation* (KDE) (Rosenblatt 1956, Parzen 1962). This method requires a *kernel function*, which is a mapping $\mathcal{K}: \mathbb{R}^m \to [0, +\infty)$ satisfying $\int_{\mathbb{R}^m}\mathcal{K}(\xi)d\xi = 1$. The commonly used kernel functions include uniform kernel $\mathcal{K}(\xi) = \frac{1}{2^m}\mathbb{I}_{\|\xi\|_\infty\le 1}(\xi)$ 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 $\hat{\Xi}_N := \{\hat{\xi}_1,\ldots,\hat{\xi}_N\} \subseteq \mathbb{R}^m$ is the set of N i.i.d. samples drawn from $p^\star$. The KDE of $p^\star$ based on $\hat{\Xi}_N$, $\mathcal{K}$ and h is

$$\hat{p}_h(\xi) := \frac{1}{N}\sum_{i=1}^N \frac{1}{h^m}\mathcal{K}\left(\frac{\xi - \hat{\xi}_i}{h}\right). \tag{18}$$

The convergence of $\hat{p}_h(\xi)$ to the true density $p^\star(\xi)$ has been studied for a long time (see Tsybakov 2008 for example) with most of the existing works focusing the asymptotic convergence property. Recently, the finite-sample nonasymptotic convergence property of KDE is characterized by Rinaldo and Wasserman (2010) and Jiang (2017). The confidence band we construct based on KDE utilizes the nonasymptotic convergence property shown by Jiang (2017).

Figure 3. (Color online) The Confidence Bands Generated by Algorithm 1 in Which $p^\star = 250 \cdot \text{Beta}(5,2)$



Notes. The bands in the first row are generated with $n = 100$ but different α's. The bands in the second row are generated with $\alpha = 0.2$ but different Ns. (a) $\alpha = 0.1$. (b) $\alpha = 0.2$. (c) $\alpha = 0.3$. (d) $n = 10$. (e) $n = 100$. (f) $n = 1{,}000$.

We need the following assumptions in this section.

Assumption 2 (For KDE-Based Confidence Bands).
The following statements hold:
A1. *There exists a nonincreasing function* $\kappa : [0, +\infty) \to [0, +\infty)$ *such that* $\mathcal{K}(\xi) = \kappa(\|\xi\|_2)$.
A2. *There exist constants* $r > 0$, $C_r > 0$, *and* $\tau > 0$ *such that* $\kappa(t) \le C_r \cdot \exp(-t^r)$ *for any* $t > \tau$.
A3. *There exists a constant* U *such that* $p(\xi) \le U$ *for any* $\xi \in \Xi$.

Assumption 2 (A1 and A2) are from Jiang (2017), and they hold if κ is one of the popular kernel densities (up to scaling) in $\mathbb{R}$, including the two mentioned as well as exponential, tricube, triangular, and Epanechnikov kernels. Under these assumptions, the following finite-sample convergence result is established by Jiang (2017).

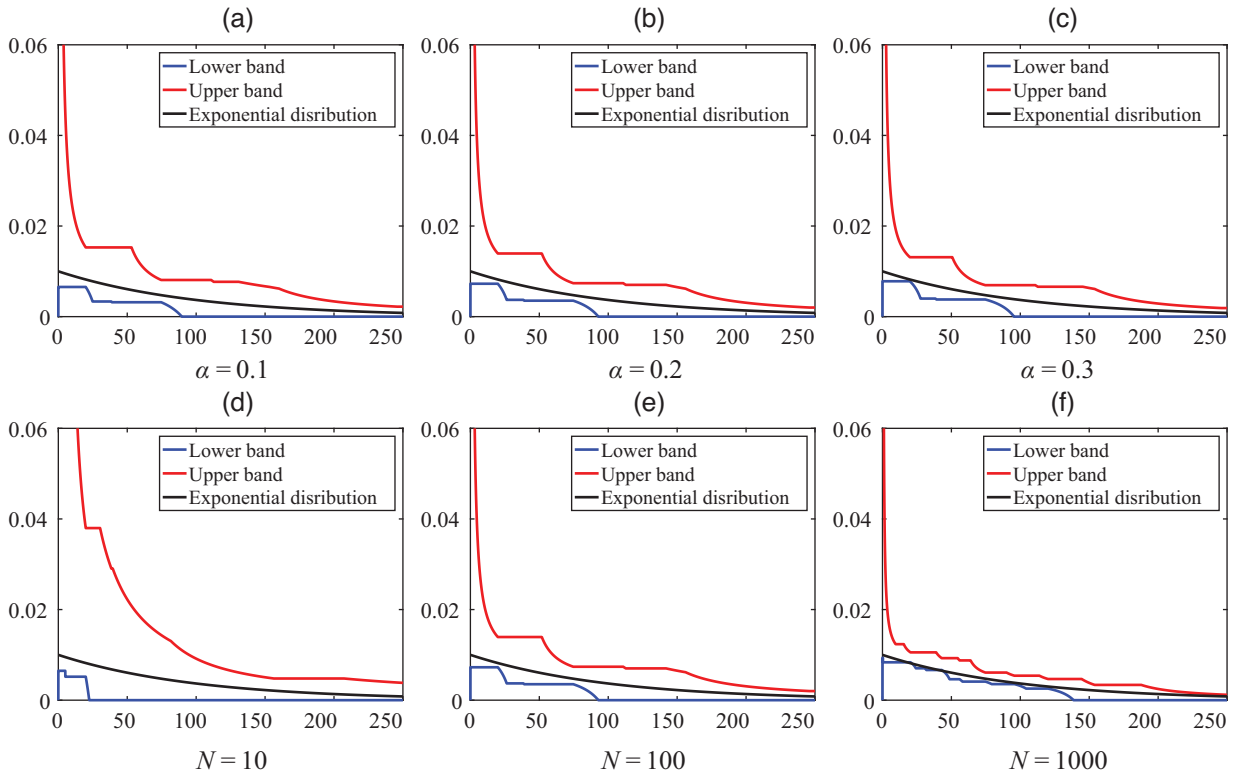
Lemma 2 (Jiang 2017, Theorem 2).
Suppose Assumption 2 holds and $p^\star$ *is* (C, ρ)*-Hölder continuous for constants* $C > 0$ *and* $\rho \in (0, 1]$, *that is,* $|\, p^\star(\xi') - p^\star(\xi)\,| \le C\|\xi' - \xi\|_2^\rho$ *for any* ξ *and* ξ'. *Let* $\alpha \in (0,1)$ *and* V_m *be the volume of the unit ball in* $\mathbb{R}^m$. *If* $h > \left(\log(N/\alpha)/N\right)^{1/m}$, *we have*

$$\mathbb{P}\left\{\sup_{\xi\in\Xi} |\hat{p}_h(\xi) - p^\star(\xi)| \le C_1 h^\rho + C_2\sqrt{\frac{\log(N/\alpha)}{Nh^m}}\right\} \ge 1 - \alpha.$$

Here[5] $\hat{p}_h(\xi)$ *is defined in* (18),

$$C_1 = V_m C \int_0^\infty \kappa(t) t^{m+\rho} dt, \text{ and } C_2 = 8m\sqrt{V_m U}\left(\int_0^\infty \kappa(t) t^{m/2} dt + 1\right) + 64m^2\kappa(0).$$

Figure 4. (Color online) The Confidence Bands Generated by Algorithm 1 in Which $p^\star$ Is a Truncated Exponential Distribution Exp(1/100) with the Support on [0, 250]



Notes. The bands in the first row are generated with $n = 100$ but different α's. The bands in the second row are generated with $\alpha = 0.2$ but different Ns. (a) $\alpha = 0.1$. (b) $\alpha = 0.2$. (c) $\alpha = 0.3$. (d) $n = 10$. (e) $n = 100$. (f) $n = 1{,}000$.

In particular, if $h = (\log(N/\alpha)/N)^{1/(2\rho+m)}$, *we have*

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

Based on the convergence property in Lemma 2, with $h > (\log(N/\alpha)/N)^{1/m}$, we can construct the KDE-based confidence band for $p^\star$ with a significance level of α as follows:

$$l_\alpha^{KDE}(\xi) = \max\{0, \hat{p}_h(\xi) - \delta\},\ u_\alpha^{KDE}(\xi) = \hat{p}_h(\xi) + \delta, \tag{19}$$

where

$$\delta = C_1 h^\rho + C_2\sqrt{\frac{\log(N/\alpha)}{Nh^m}}. \tag{20}$$

The corresponding uncertainty set is

$$\mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha) := \left\{p \in \mathcal{L} \,\middle|\, l_\alpha^{KDE}(\xi) \le p(\xi) \le u_\alpha^{KDE}(\xi),\ \forall\, \xi \in \Xi,\ \int_\Xi p(\xi)\, d\xi = 1\right\}. \tag{21}$$

The following property is a direct consequence of Lemma 2.

Theorem 3. *Suppose Assumption 2 holds. Then,* $\mathbb{P}\left\{p^\star \in \mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)\right\} \ge 1-\alpha$.

We summarize this procedure for constructing a KDE-based 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 $\hat{\Xi}_N = \{\hat{\xi}^1, \ldots, \hat{\xi}^N\}$ sampled from $p^\star$, a constant $U \geq p^\star(\xi)$ for any $\xi \in \Xi$, a significance level $\alpha \in (0,1)$, a kernel function $\mathcal{K}(\xi) = \kappa(\|\xi\|_2)$ with κ satisfying Assumption 2(A1 and A2), a bandwidth $h > 0$, 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 $\hat{p}_h(\xi)$ and δ be defined as in (18) and (20), respectively.

3: Compute $(l_\alpha^{KDE}(\xi), u_\alpha^{KDE}(\xi))$ as in (19).

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

The following theorem shows that, under additional assumptions to Assumption 2, $v^\star_{\mathcal{D}^{KDE}(\hat{\Xi}_N,\alpha)}$ defined in (7) converges to $v^\star$ in (2).

Theorem 4. *Suppose the assumptions of Lemma 2 hold,* $\max_{x\in\mathcal{X},\xi\in\Xi}|f(x,\xi)| < +\infty$, *and* Ξ *is bounded. Let* $h = (\log(N/\alpha)/N)^{1/(2\rho+m)}$ *in Algorithm 2. For any* $\epsilon > 0$, *we have*

$$\lim_{N\to+\infty} \mathbb{P}\left\{\sup_{x\in\mathcal{X}} \left| \int_\Xi f(x,\xi) p^\star(\xi) d\xi - \sup_{p\in\mathcal{D}^{KDE}(\hat{\Xi}_N,\alpha)} \int_\Xi f(x,\xi) p(\xi) d\xi \right| \leq \epsilon \right\} \geq 1-\alpha$$

and $\lim_{N\to+\infty}\mathbb{P}\left\{ \left| v^\star_{\mathcal{D}^{KDE}(\hat{\Xi}_N,\alpha)} - v^\star \right| \leq \epsilon \right\} \geq 1-\alpha$.

Proof. See Appendix B.

3.3. Example: Reduced Conservativeness by the Ambiguity Set Consisting of Only Absolutely Continuous Distributions

We consider a cost function $f(\xi)$ on $[0,1]$, which is zero on $[0,1-\delta)$ for $\delta \in (0,1)$ and linearly increases to $\frac{1}{\delta}$ on $[1-\delta,1]$. In particular, we define

$$f(\xi) = \begin{cases} 0, & \text{if } \xi \in [0,1-\delta) \\ \dfrac{1}{\delta^2}(\xi - (1-\delta)), & \text{if } \xi \in [1-\delta,1]. \end{cases}$$

Let $\xi^\star$ follow a uniform distribution on $[0,1]$, that is, $p^\star(\xi) = 1$ on $[0,1]$. For simplicity, we do not introduce a decision variable x in f but only consider the problem of estimating the worst-case expected cost based on different ambiguity sets. We choose δ to be a small number so that the cost is very high when ξ is close to one but quickly decreases to zero when ξ is deviated away from one. In this example, for any $\delta \in (0,1)$, the expected cost under the true distribution is

$$v^\star := \int_0^1 f(\xi) d\xi = \frac{1}{2}.$$

Given a sample $\hat{\Xi}_N = \{\hat{\xi}_1, \ldots, \hat{\xi}_N\}$ from $p^\star$, according to Mohajerin Esfahani and Kuhn (2018), the worst-case expected cost over the ambiguity set constructed by the Wasserstein metric with a radius of ϵ can be calculated by solving

$$v_W^\star := \begin{cases} \sup\limits_{Q_i, i=1,\ldots,N} & \dfrac{1}{N}\sum\limits_{i=1}^N \int_0^1 f(\xi) Q_i(d\xi) \\ \text{s.t.} & Q_i \text{ is a probability measure on } [0,1], \\ & \dfrac{1}{N}\sum\limits_{i=1}^N \int_0^1 |\xi - \hat{\xi}_i| Q_i(d\xi) \leq \epsilon. \end{cases} = \begin{cases} \sup\limits_{\xi_i, i=1,\ldots,N} & \dfrac{1}{N}\sum\limits_{i=1}^N f(\xi_i) \\ \text{s.t.} & \dfrac{1}{N}\sum\limits_{i=1}^N |\xi^i - \hat{\xi}_i| \leq \epsilon. \end{cases} \tag{22}$$

According to equation (8) in Mohajerin Esfahani and Kuhn (2018), to ensure a significance level α, one should choose the radius to be $\epsilon = \Omega(1/\sqrt{N})$ so that we can assume $\frac{1}{N} \leq \epsilon$. As a result, the solution $(\xi_1, \ldots, \xi_{N-1}, \xi_N) = (\hat{\xi}_1, \ldots, \hat{\xi}_{N-1}, 1)$ is feasible to (22) so that the optimal objective value of (22) is lower bounded as

$$v_W^\star \geq \frac{1}{N}\sum_{i=1}^{N-1} f(\hat{\xi}^i) + \frac{1}{N} f(1) \geq \frac{1}{N\delta},$$

which means $v_W^\star$ increases to positive infinity as δ approaches zero (but N is fixed). In other words, the worst-case expected cost over the ambiguity set constructed by the Wasserstein metric can be significantly larger than the true expected cost in this example.

Now let's consider the worst-case expected cost over the ambiguity set constructed by the shape-restricted confidence band in (14). We choose the mode of $p^\star$ to be $\mu = 0$. In fact, we can make the density function of the uniform distribution slightly tilt up at zero while keeping $v_W^\star$ and $v^\star$ almost unchanged. We first claim that $u_\alpha^{SR}(\xi)$ defined in (16) is no more than $\frac{1}{1-\delta}$ when $\xi \in [1-\delta, 1]$. To show this, we first recall that $\hat{N} \in \mathbb{Z}_+$ is the number of distinct elements in the set $\{0, 1, \xi, \hat{\xi}_{(k_1)}, \ldots, \hat{\xi}_{(k_M)}\}$ and z_j is the jth smallest element (so that $z_1 = 0$). Let $\hat{j}$ be the indexes in $\{1, 2, \ldots, \hat{N}\}$ such that $z_{\hat{j}} = \xi$. Suppose $u_\alpha^{SR}(\xi) > \frac{1}{1-\delta}$ at some $\xi \in [1-\delta, 1]$. Because $\xi > \mu = 0$, by (15) and (16), there must exist $\boldsymbol{\beta} \in \mathcal{H}(\hat{\Xi}_N, \alpha)$ such that $\beta_{\hat{j}-1} = u_\alpha^{SR}(\xi) > \frac{1}{1-\delta}$ and

$$1 = \sum_{j=1}^{\hat{N}-1} \beta_j (z_{j+1} - z_j) \geq \sum_{j=1}^{\hat{j}-1} \beta_j (z_{j+1} - z_j) > \frac{1}{1-\delta}(z_{\hat{j}} - z_1) = \frac{\xi}{1-\delta} \geq 1,$$

where the first equality is from a constraint in (15), the first inequality is because $\beta_j \geq 0$, the second inequality is because $\beta_{\hat{j}-1} > \frac{1}{1-\delta}$ and the fact that β_j is nonincreasing in j, and the last equality is because $z_{\hat{j}} = \xi$ and $z_1 = 0$. This contradiction means we must have $u_\alpha^{SR}(\xi) \leq \frac{1}{1-\delta}$ when $\xi \in [1-\delta, 1]$. As a result, we have

$$v^\star_{\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)} = \sup_{p \in \mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)} \int_0^1 f(\xi) p(\xi) d\xi \leq \int_{1-\delta}^1 f(\xi) u_\alpha^{SR}(\xi) d\xi \leq \int_{1-\delta}^1 \frac{f(\xi)}{1-\delta} d\xi = \frac{1}{2(1-\delta)}.$$

In summary, as δ decreases to zero, we have $v^\star = \frac{1}{2}$, $v^\star_{\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)} \leq \frac{1}{2(1-\delta)} \to \frac{1}{2}$ and $v_W^\star \to +\infty$, which indicates that, when the true distribution is absolutely continuous, an ambiguity set consisting of only absolutely continuous distributions can be preferred to an ambiguity set that includes discrete distributions because the former may provide a less conservative estimation of the worst-case cost.

Our example can be easily extended by introducing a decision variable $x \in \Xi = [0, 1]$ in the cost function $f(x, \xi) = x \cdot f(\xi) + (1-x) \cdot c$, where c is a constant significantly higher than $\frac{1}{2}$ but not more than $\frac{1}{N\delta}$. The optimal decisions made using either the true distribution or the confidence-band based ambiguity set are $x = 1$, but the decision solved using a Wasserstein-based ambiguity set is $x = 0$, which leads to a significantly higher expected cost under the true distribution.

3.4. Other Confidence Bands

Besides (16) and (19), there exist other techniques to construct a confidence band for $p^\star$ that can also be used to construct an ambiguity set in the form of (6). One of the simplest approaches is based on the histogram density estimator. For simplicity, we assume $p^\star$ is univariate ($m = 1$), and $\Xi = [a, b]$ with finite a and b. Suppose Ξ is evenly divided into K intervals at points $a_k = a + (b-a)\frac{k}{K}$ for $k = 0, 1, \ldots, K$. The histogram estimator of $p^\star$ based on $\hat{\Xi}_N$ is defined as

$$\hat{p}_K(\xi) := \sum_{k=0}^{K-1} \frac{K \mathbb{I}_{[a_k, a_{k+1})}(\xi)}{b-a} \cdot \frac{N_k}{N},$$

where N_k is the number of elements in $\hat{\Xi}_N \cap [a_k, a_{k+1})$. We assume $p^\star$ is C-Lipschitz continuous for a constant $C > 0$, that is, $| p^\star(\xi') - p^\star(\xi) | \leq C | \xi' - \xi |$ for any ξ and ξ' in Ξ. According to section 3.2.1 in Scott (2015), we have

$$|\mathbb{E}[\hat{p}_K(\xi)] - p^\star(\xi)| \leq \frac{C(b-a)}{K}, \quad \forall \xi \in \Xi. \tag{23}$$

According to theorem 2.2 in Chen (2019) and its proof, we have, for $\alpha \in (0, 1)$,

$$\mathbb{P}\left\{ \sup_{\xi \in \Xi} |\hat{p}_K(\xi) - \mathbb{E}[\hat{p}_K(\xi)]| \leq \frac{K}{b-a} \sqrt{\frac{\log(2K/\alpha)}{2N}} \right\} \geq 1 - \alpha. \tag{24}$$

Using this result and (23), we can construct a confidence band for $p^\star$ with a significance level α as

$$l_\alpha^H(\xi) = \max\{0, \hat{p}_K(\xi) - \delta\},\ u_\alpha^H(\xi) = \hat{p}_K(\xi) + \delta, \tag{25}$$

where

$$\delta = \frac{C(b-a)}{K} + \frac{K}{b-a}\sqrt{\frac{\log(2K/\alpha)}{2N}}.$$

The corresponding uncertainty set is

$$\mathcal{D}^H(\hat{\Xi}_N, \alpha) := \left\{p \in \mathcal{L} \mid l_\alpha^H(\xi) \le p(\xi) \le u_\alpha^H(\xi),\ \ \forall\, \xi \in \Xi,\ \int_\Xi p(\xi)\, d\xi = 1\right\}, \tag{26}$$

which satisfies

$$\mathbb{P}\left\{p^\star \in \mathcal{D}^H(\hat{\Xi}_N, \alpha)\right\} \ge 1 - \alpha.$$

By choosing K such that $K \to +\infty$ and $K = o(\sqrt{N})$ as $N \to +\infty$, one can show the convergence of the DRO objective value in a theorem similar to Theorem 4.

Another confidence band can be constructed using the frequency polygon, which is the linearly smoothed version of $\hat{p}_K(\xi)$ described previously (see chapter 4 in Scott 2015). In particular, we define

$$\hat{p}_K^{FP}(\xi) := \frac{K(\xi - a_i)}{b-a}\hat{p}_K(a_{i+1}) + \frac{K(a_{i+1} - \xi)}{b-a}\hat{p}_K(a_i) \text{ if } \xi \in [a_i, a_{i+1}).$$

Assuming $p^\star$ is C-Lipschitz continuous, the distribution of $\hat{p}_K^{FP}(\xi)$ is determined by $\hat{p}_K(a_i)$ for $i = 0, \ldots, K-1$ so that we can easily derive the properties of $\hat{p}_K^{FP}$ similar to (23) and (24), which allows us to construct a confidence band and the corresponding ambiguity set similar to (25) and (26).

4. A Numerical Method for DRO

In this section, we present a numerical scheme that solves the DRO problem in (7). Although we only described a few specific ambiguity sets in Section 3, the optimization method we propose here can be potentially applied to DRO with any ambiguity set in the form of (6) as long as the following assumption holds after $\hat{\Xi}_N$ is drawn.

Assumption 3 (On Ambiguity Set $\mathcal{D}(\hat{\Xi}_N, \alpha)$). *The following statements hold:*

A1. $\int_\Xi u_\alpha(\xi) d\xi < \infty, \quad \forall\ x \in \mathcal{X}.$

A2. $\mathcal{D}(\hat{\Xi}_N, \alpha) \neq \emptyset.$

Note that this assumption is made for $\mathcal{D}(\hat{\Xi}_N, \alpha)$ after $\hat{\Xi}_N$ is drawn because the numerical algorithm in this section is proposed to solve (7), which is only defined after a particular $\hat{\Xi}_N$ is drawn. We show that this assumption holds with a high probability for $\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)$ and $\mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)$.

Suppose Assumption 1 holds. The ambiguity set $\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)$ in (14) satisfies Assumption 3 with a probability of $1-\alpha$. This is because we have $u_\alpha^{SR}(\xi) \le U$ for any $\xi \in \Xi$ according to the definition of $u_\alpha^{SR}(\xi)$ in (16) and the constraint $0 \le \beta_j \le U$ in (15). Moreover, by Theorem 1, $\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)$ at least contains $p^\star$ and, thus, is nonempty with a probability of $1-\alpha$.

Suppose Assumption 2 holds and Ξ is bounded. The ambiguity set $\mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)$ in (21) can also satisfy Assumption 3 with a probability of $1-\alpha$. In fact, by Theorem 2, with a bandwidth $h > (\log(N/\alpha)/N)^{1/m}$, $\mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)$ at least contains $p^\star$ with a probability of $1-\alpha$, which further implies $u_\alpha^{KDE}(\xi) = \hat{p}_h(\xi) + \delta \le p^\star(\xi) + 2\delta \le U + 2\delta$ according to (19) and (20).

We define $v_P(x)$ as the optimal value of the inner maximization problem of (7), that is,

$$v_P(x) := \sup_{p \in \mathcal{D}(\hat{\Xi}_N, \alpha)} \int_\Xi f(x, \xi) p(\xi) d\xi. \tag{27}$$

When Assumption 3 holds and $\max_{\xi\in\Xi}|f(x,\xi)| < +\infty$ for every $x \in \mathcal{X}$, we have

$$|v_P(x)| \le \sup_{p\in\mathcal{D}(\hat{\Xi}_N,\alpha)} \int_\Xi |f(x,\xi)|p(\xi)d\xi \le \int_\Xi |f(x,\xi)|u_\alpha(\xi)d\xi < \infty. \tag{28}$$

Note that (27) is a *continuous linear program* that is formulated with a functional decision variable (i.e., p) and continuously many constraints. In general, (27) 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 (27) and present its convergence rate. For simplicity, we write (u_α, l_α) as (u, l) by suppressing α in the rest of this section.

The dual problem of (27) is given as follows:

$$\begin{array}{l} v_D(x) := \inf\limits_{\lambda,\alpha,\beta} \lambda - \int l(\xi)\alpha(\xi)d\xi + \int u(\xi)\beta(\xi)d\xi \\ \text{s.t.} \quad \lambda - \alpha(\xi) + \beta(\xi) \ge \mathrm{f}(\mathrm{x},\xi) \quad \forall\, \xi \in \Xi, \\ \alpha(\xi) \ge 0,\ \beta(\xi) \ge 0 \quad \forall\, \xi \in \Xi. \end{array} \tag{29}$$

Weak duality implies $v_P(x) \le v_D(x)$. According to Shapiro (2001), strong duality holds between (27) and (29). We state the strong duality in the following lemma whose proof is given Appendix C. Because this result is not new, in the proof, we only discuss what result in Shapiro (2001) is used to establish the strong duality and why the assumptions of that result hold in our case.

Lemma 3. *Suppose Assumption 3 holds and* $\max_{\xi\in\Xi}|f(x,\xi)| < +\infty$ *for any* $x \in \mathcal{X}$. *We have* $v_P(x) = v_D(x)$ *for any* $x \in \mathcal{X}$.

The objective function of (29) can be rewritten as

$$\lambda + \int l(\xi)(\beta(\xi) - \alpha(\xi))d\xi + \int (u(\xi) - l(\xi))\beta(\xi)d\xi.$$

Let $(z)_- := \max\{-z, 0\}$ and $(z)_+ := \max\{z, 0\}$. The constraints of (29) require $\beta(\xi) \ge (f(x,\xi) - \lambda + \alpha(\xi))_+$. Because $u(\xi) \ge l(\xi) \ge 0$ for any $\xi \in \Xi$, by optimizing $\beta(\xi)$ for any fixed $\alpha(\xi)$, we can always require $\beta(\xi) = (f(x,\xi) - \lambda + \alpha(\xi))_+$ for any $\xi \in \Xi$ in the optimal solution. Eliminating $\beta(\xi)$ in the objective function gives

$$\lambda + \int [l(\xi)[(f(x,\xi) - \lambda + \alpha(\xi))_+ - \alpha(\xi)]d\xi + (u(\xi) - l(\xi))(f(x,\xi) - \lambda + \alpha(\xi))_+]d\xi.$$

Because $u(\xi) \ge l(\xi) \ge 0$, the preceding integrand is nonincreasing in $\alpha(\xi)$ when $\alpha(\xi) \le \lambda - f(x,\xi)$ and nondecreasing in $\alpha(\xi)$ when $\alpha(\xi) \ge \lambda - f(x,\xi)$. Considering the constraint $\alpha(\xi) \ge 0$, the objective function can be minimized at $\alpha(\xi) = (\lambda - f(x,\xi))_+ = (f(x,\xi) - \lambda)_-$ for any fixed λ. After further eliminating $\alpha(\xi)$, the optimal value of (29) can be stated equivalently as

$$v_D(x) = \inf_\lambda \lambda - \int l(\xi)(f(x,\xi) - \lambda)_- d\xi + \int u(\xi)(f(x,\xi) - \lambda)_+ d\xi, \tag{30}$$

and thus, we have

$$v^\star_{\mathcal{D}(\hat{\Xi}_N,\alpha)} := \inf_{x\in\mathcal{X},\,\lambda} \left\{ F(x,\lambda) := \lambda - \int l(\xi)(f(x,\xi) - \lambda)_- d\xi + \int u(\xi)(f(x,\xi) - \lambda)_+ d\xi \right\}. \tag{31}$$

In general, the two integrals appearing in (31) do not have analytical forms, raising a challenge for finding the optimal solution. Hence, we consider an SGD method for solving (31) by constructing a stochastic subgradient of $F(x,\lambda)$ in (31). To do so, we need the following result.

Lemma 4. *Suppose (i) Assumption 3 holds, (ii)* $f(x,\xi)$ *is lower semicontinuous and convex in* x *for any* $\xi \in \Xi$ *and bounded over* Ξ *for any* $x \in X$, *and (iii)* $\mathcal{X}$ *has a nonempty interior. There exists a Lebesgue integrable mapping* $f'(x,\xi) : \mathcal{X} \times \Xi \to \mathbb{R}^d$ *such that* $f'(x,\xi) \in \partial_x f(x,\xi)$ *and*

$$\int l(\xi)f'(x,\xi)\mathbb{I}_{f(x,\xi)<\lambda}(\xi)d\xi + \int u(\xi)f'(x,\xi)\mathbb{I}_{f(x,\xi)\ge\lambda}(\xi)d\xi \in \partial_x F(x,\lambda), \tag{32}$$

$$1 - \int l(\xi)\mathbb{I}_{f(x,\xi)<\lambda}(\xi)d\xi - \int u(\xi)\mathbb{I}_{f(x,\xi)\ge\lambda}(\xi)d\xi \in \partial_\lambda F(x,\lambda), \tag{33}$$

where $\mathbb{I}_E(\cdot)$ *is the* $0-1$ *indicator function of the event* E.

Proof. Because $f(x,\xi)$ is measurable in x and ξ, lower semicontinuous, and convex in x and also because $0 \le l(\xi) \le u(\xi)$, the integrand $-l(\xi)(f(x,\xi)-\lambda)_- + u(\xi)(f(x,\xi)-\lambda)_+$ in (31) is random lower semicontinuous (see definition 7.35 in Shapiro et al. 2014) and convex in x for any $\xi \in \Xi$ according to theorem 7.36 in Shapiro et al. (2014). By Assumption 3, we have $F(x,\lambda) < +\infty$ for any $x \in \mathcal{X}$ and λ. Because $\mathcal{X}$ has a nonempty interior, all the assumptions of theorem 7.47 in Shapiro et al. (2014) hold, which implies the conclusions of this lemma. □

According to Lemma 4, we can construct the stochastic subgradients of $F(x,\lambda)$ by stochastically approximating the integrals in (32) and (33). We have assumed Ξ is bounded so that there exists a box $I \subset \mathbb{R}^m$ containing Ξ. We denote the volume of I by $|I|$. Suppose ξ is sampled from a uniform distribution on I. We can show that the left-hand sides of (32) and (33) are the expectations of

$$|I|\mathbb{I}_{f(x,\xi)<\lambda}(\xi)l(\xi)f'(x,\xi) + |I|\mathbb{I}_{f(x,\xi)\ge\lambda}(\xi)u(\xi)f'(x,\xi) \tag{34}$$

and

$$1 - |I|\mathbb{I}_{f(x,\xi)<\lambda}(\xi)l(\xi) + |I|\mathbb{I}_{f(x,\xi)\ge\lambda}(\xi)u(\xi), \tag{35}$$

respectively. Hence, we can use (34) and (35) with ξ uniformly sampled from I as the unbiased stochastic subgradients of F with respect to x and λ, respectively. To further reduce the sampling noise, we use a minibatch technique by generating B i.i.d. samples from the uniform distribution on I and constructing such a stochastic subgradient using each sample and then taking the average over all samples. Based on this idea, we proposed the SGD method for (31) in Algorithm 3. The convergence of Algorithm 3 is well known (see, e.g., Nemirovski et al. 2009), so we only present a short proof in Appendix D by directly using some results from Nemirovski et al. (2009) and explaining why the assumptions of those results holds in our case.

Algorithm 3 (SGD for (31))

Input: Initial solution $(x_0,\lambda_0) \in \mathcal{X}\times\mathbb{R}$, batch size $B \ge 1$, step length $\eta_k > 0$, a box $I \subset \mathbb{R}^d$ containing Ξ, the volume of I denoted by $|I|$, and the total number of iteration K.

1: **for** $k = 0,1,\dots,K-1$ **do**

2: Compute $(\bar{x}_k,\bar{\lambda}_k) = \dfrac{\sum_{i=0}^k \eta_i(x_i,\lambda_i)}{\sum_{i=0}^k \eta_i}$.

3: Sample $\{\xi_1,\xi_2,\dots,\xi_B\}$ from a uniform distribution over I.

4: Construct the stochastic gradients

$$g_x^k = \frac{|I|}{B}\sum_{i:f(x_k,\xi_i)<\lambda_k} l(\xi_i)f'(x_k,\xi_i) + \frac{|I|}{B}\sum_{i:f(x_k,\xi_i)\ge\lambda_k} u(\xi_i)f'(x_k,\xi_i),$$

$$g_\lambda^k = 1 - \frac{|I|}{B}\sum_{i:f(x_k,\xi_i)<\lambda_k} l(\xi_i) - \frac{|I|}{B}\sum_{i:f(x_k,\xi_i)\ge\lambda_k} u(\xi_i).$$

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

6: **end for**

Output: $(\bar{x}_{K-1},\bar{\lambda}_{K-1})$

Theorem 5 (Nemirovski et al. 2009).

Suppose (i) the assumptions of Lemma 4 hold and (ii) there exists a constant M such that $\mathbb{E}\|(g_x^k,g_\lambda^k)\|_2^2 \le M^2$ for all k in Algorithm 3 and a constant D such that $\frac{1}{2}\|(x_0,\lambda_0)-(x_*,\lambda_*)\|_2^2 \le D^2$. Algorithm 3 with step length $\eta_k = \frac{\sqrt{2}D}{M\sqrt{K}}$ for all k ensures

$$\mathbb{E}\left[v_D(\bar{x}_k) - v^\star_{\mathcal{D}(\hat{\Xi}_N,\alpha)}\right] \le \mathbb{E}\left[F(\bar{x}_k,\bar{\lambda}_k) - v^\star_{\mathcal{D}(\hat{\Xi}_N,\alpha)}\right] \le \frac{\sqrt{2}DM}{\sqrt{K}}. \tag{36}$$

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 Bertsimas et al. (2014), which applies hypothesis tests to construct ambiguity sets, and for the portfolio selection example, we compare our approach with that in Mohajerin Esfahani and Kuhn (2018), which applies the Wasserstein metric to construct ambiguity sets. We implement all methods in MATLAB (R2014a) version 8.3.0.532 on a Windows

computer with an Intel Core i3 2.93 GHz and 4 GB of RAM. The linear programs involved in the implementations are solved by CPLEX 12.4 and the modeling language YALMIP (Lofberg 2004).

5.1. Single-Item Newsvendor

We consider a classic single-item newsvendor problem in which we assume the demand $\xi^\star$ of an item follows a continuous distribution with a bounded support set $[a,b] \subseteq \mathbb{R}$ with $0 \le a < b$ and a bounded density function. An order of x units ($x \ge 0$) 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. Assuming a set of historical demand data are available, we construct an ambiguity set $\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)$ in (14) and solve the DRO in (7) with $\mathcal{D}(\hat{\Xi}_N, \alpha) = \mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)$. We compare the optimal order obtained with the one found by the DRO model in Bertsimas et al. (2014) in which the ambiguity set is built using the 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 $[0,250]$.

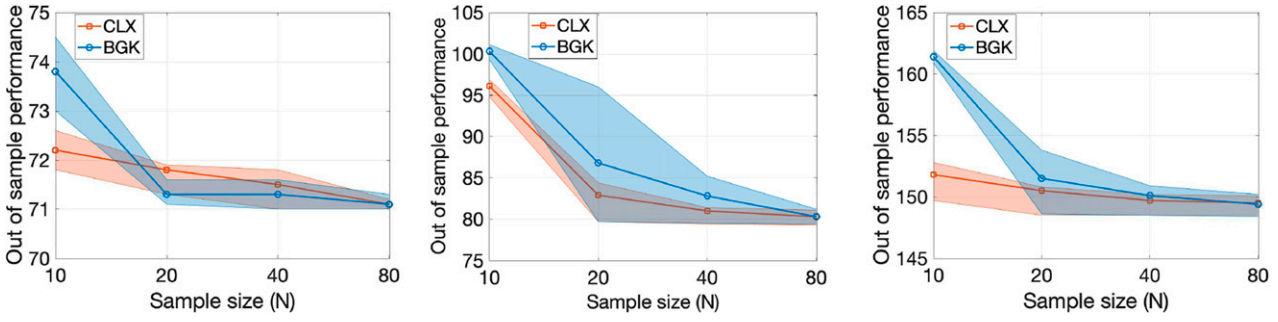
For each distribution, we consider four different sample sizes, that is, $N \in \{10,20,40,80\}$. For each size, we randomly generate a data set $\hat{\Xi}_N$ by i.i.d. sampling from the demand distribution. Using $\hat{\Xi}_N$, we apply our approach and the method by Bertsimas et al. (2014) to construct 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. data set $\{\xi_i'\}_{i=1}^{N_{\text{large}}}$ with $N_{\text{large}} = 100{,}000$ from the true distribution and calculate the sample average approximation of the expected cost, that is, $\frac{1}{N_{\text{large}}}\sum_{i=1}^{N_{\text{large}}} f(\hat{x},\xi_i')$ 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 Lemma 1, it involves quantities that are hard to estimate (e.g., B and ρ). When we construct $\mathcal{D}^{SR}(\hat{\Xi}_N, \alpha)$, we set

$$K = \mathcal{K}(N,c) := \min\left\{\left\lceil c\left(N^2 \log N\right)^{1/3}\right\rceil, N-1\right\}$$

and select c from $\{0.5, 0.75, 1, 1.25, 1.5\}$ based on the holdout validation method. The 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 $\hat{\Xi}_{train}$ and $\hat{\Xi}_{test}$ with $|\hat{\Xi}_{train}| = 0.7N$ and $|\hat{\Xi}_{test}| = 0.3N$. Given a combination of c and α, we first construct

Figure 5. (Color online) The Out-of-Sample Performance of CLX and BGK on a Newsvendor Problem



Notes. The boundaries of the shaded areas 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).

$\mathcal{D}^{SR}(\hat{\Xi}_{train}, \alpha)$ using $K = \mathcal{K}(0.7N, c)$, and then we solve

$$x^*_{c,\alpha} \in \arg\min_{x\in\mathcal{X}} \sup_{p\in\mathcal{D}^{SR}(\hat{\Xi}_{train},\,\alpha)} \int_{\Xi} f(x,\xi)p(\xi)d\xi$$

using Algorithm 3. Then, we select the combination of c and α that give the smallest $\frac{1}{|\hat{\Xi}_{test}|}\sum_{\hat{\xi}\in\hat{\Xi}_{test}} f(x^*_{c,\alpha}, \hat{\xi})$. For each of the 100 independent trials, we repeat this process to select c and α. For a fair comparison, we also apply the same validation scheme to choose the significance level α used in the method by Bertsimas et al. (2014).

We denote our approach by CLX and the method in Bertsimas et al. (2014) by BGK. Figure 5 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 the 80th percentile of the out-of-sample performances in the 100 trials. Figure 5 indicates that CLX has 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. We note that the average computation times of CLX and BGK over all instances are about, respectively, 80 and 2 seconds for all tested sample sizes.

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 $w = (w_1, w_2, \ldots, w_n)$ 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 $\xi^\star_i$. The return from each unit of budget is, thus, $w^\top \xi^\star$. 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 Rockafellar and Uryasev (2000). Suppose the joint distribution of the return $\xi^\star = (\xi^\star_1, \xi^\star_2, \ldots, \xi^\star_n)$ has a density function $p^\star(\xi)$. The CVaR at level $\epsilon \in (0,1)$ of the return of a portfolio with respect to a probability distribution $p^\star$ is defined as

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

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

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

Here, $\gamma > 0$ 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 distribution $p^\star$ of $\xi^\star$ is unknown, but a collection of historical data returns is collected, the investor can construct a data-driven ambiguity set of $p^\star$ and solve the DRO problem corresponding to the SP. We construct the ambiguity set $\mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)$ in (21) and solve the DRO in (7) with $\mathcal{D}(\hat{\Xi}_N, \alpha) = \mathcal{D}^{KDE}(\hat{\Xi}_N, \alpha)$ to construct an portfolio. Then, we compare our solution with the one obtained by the DRO model in Mohajerin Esfahani and Kuhn (2018) in which the ambiguity set is constructed using the Wasserstein metric.

Following the numerical experiments by Mohajerin Esfahani and Kuhn (2018), we consider $n = 10$ assets with decomposable returns $\xi^\star_i = \phi + \zeta_i$ for $i = 1, 2, \ldots, 10$, 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 the ith asset. By the construction, assets with higher indexes promise higher mean returns at higher risks. We set $\epsilon = 20\%$ and $\gamma = 10$ in our all experiments. We consider six different sample sizes, that is, $N \in \{30, 60, 120, 240, 480, 960\}$. For each size, we randomly generate a data set $\hat{\Xi}_N$ by i.i.d. sampling returns from the aforementioned distribution of $\xi^\star$. Using $\hat{\Xi}_N$, we apply our approach and the method by Mohajerin Esfahani and Kuhn (2018) to construct the ambiguity

sets and then solve the DRO to obtain a portfolio $\hat{w}$ from each approach. To evaluate the out-of-sample performance of $\hat{w}$, we sample another i.i.d. data set $\{\xi_i'\}_{i=1}^{N_{\text{large}}}$ with $N_{\text{large}} = 100,000$ from the true distribution and calculate the sample average approximation of

$$\mathbb{E}_{\xi^\star \sim p^\star}[-\hat{w}^\top \xi^\star] + \gamma \text{CVaR}_{\xi^\star \sim p^\star, \epsilon}(-\hat{w}^\top \xi^\star)$$

using $\{\xi_i'\}_{i=1}^{N_{\text{large}}}$ 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 that 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 (19) with the parameters δ and h selected by the holdout validation method rather than their theoretical values in (20) and Lemma 2. In particular, we set $h = c(\log(N)/N)^{1/(2+m)}$ (see Lemma 2) 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 $\hat{\Xi}_{train}$ and $\hat{\Xi}_{test}$ with $|\hat{\Xi}_{train}| = 0.7N$ and $|\hat{\Xi}_{test}| = 0.3N$. Given a combination of c and δ, we first construct $\mathcal{D}^{KDE}(\hat{\Xi}_{train})$ using and δ and $h = c(\log(N)/N)^{1/(2+m)}$, and then solve

$$x_{c,\delta}^* \in \arg\min_{x \in \mathcal{X}} \sup_{p \in \mathcal{D}^{KDE}(\hat{\Xi}_{train})} \int_\Xi f(x, \xi) p(\xi) d\xi$$

using Algorithm 3. Let $w_{c,\delta}^*$ be the portfolio component in $x_{c,\delta}^*$. Then, we select the combination of c and δ that makes the smallest sample approximation of

$$\mathbb{E}_{\xi^\star \sim p^\star}[-(w_{c,\delta}^*)^\top \xi^\star] + \gamma \text{CVaR}_{\xi^\star \sim p^\star, \epsilon}(-(w_{c,\delta}^*)^\top \xi^\star)$$

on $\hat{\Xi}_{train}$. For each of the 100 independent trials, we repeat this process to select c and δ.

For a fair comparison, we apply the same validation scheme to choose the radius (the only parameter) of the Wasserstein ball in Mohajerin Esfahani and Kuhn (2018). More specifically, we randomly partition $\hat{\Xi}_N$ into $\hat{\Xi}_{train}$ and $\hat{\Xi}_{test}$ with $|\hat{\Xi}_{train}| = 0.7N$ and $|\hat{\Xi}_{test}| = 0.3N$ and select the value for radius r from the set $\mathcal{R} := \{r = b \cdot 10^c | b \in \{0, 1, \ldots, 9\}, c \in \{-3, -2, -1\}\}$. For each candidate of the radius, we construct the Wasserstein ball

$$\mathcal{D}^W(\hat{\Xi}_{train}, r) := \{\mathcal{P} | \mathcal{P} \text{ is a distribution on } \mathbb{R}^m,\ d^W(\hat{\mathcal{P}}, \mathcal{P}) \le r\},$$

where $\hat{\mathcal{P}}$ is the empirical distribution defined over $\hat{\Xi}_{train}$ and $d^W(\hat{\mathcal{P}}, \mathcal{P})$ is defined as

$$d^W(\hat{\mathcal{P}}, \mathcal{P}) := \inf \left\{ \int_{\mathbb{R}^m \times \mathbb{R}^m} \|\xi - \xi'\|_1 \Pi(d\xi, d\xi') \,\middle|\, \begin{array}{l} \Pi \text{ is a joint distribution on } \mathbb{R}^m \times \mathbb{R}^m \\ \text{with marginals } \mathcal{P} \text{ and } \hat{\mathcal{P}} \text{ respectively} \end{array} \right\}.$$

We then solve

$$x_r^* \in \arg\min_{x \in \mathcal{X}} \sup_{\mathcal{P} \in \mathcal{D}^W(\hat{\Xi}_{train}, r)} \int_\Xi f(x, \xi) \mathcal{P}(d\xi)$$

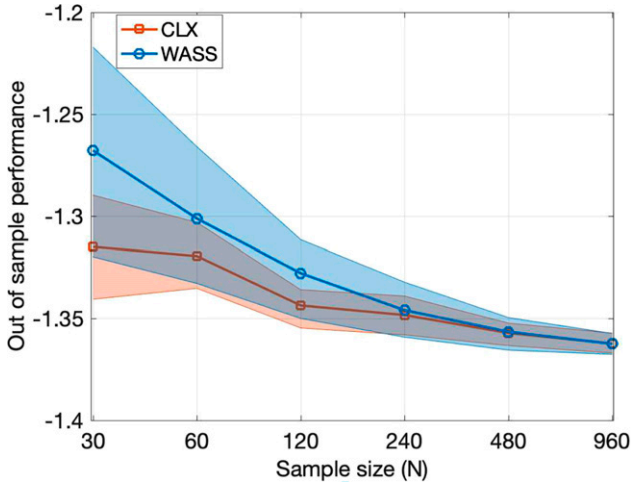
using their proposed method. Let w_r^* be the portfolio component in x_r^*. Then, we select radius r that makes the smallest sample approximation of

$$\mathbb{E}_{\xi^\star \sim p^\star}[-(w_r^*)^\top \xi^\star] + \gamma \text{CVaR}_{\xi^\star \sim p^\star, \epsilon}(-(w_r^*)^\top \xi^\star)$$

on $\hat{\Xi}_{train}$. We employ this procedure to choose the best radius for each of the 100 independent trials.

We denote our approach by CLX and the method in Mohajerin Esfahani and Kuhn (2018) by WASS and plot the numerical results in Figure 6. In particular, for each of the 100 trials in each sample-size scenario, we evaluate the solutions from both CLX and WASS by their out-of-sample performances. We then plot the 20th percentile, the mean, and the 80th percentile of the out-of-sample performances of both approaches over the six sample-size scenarios. Figure 6 indicates that both approaches converge to the true expectation at a comparable speed with the sample size increases. We also report the computation times for both CLX and WASS methods. The average computation times of CLX are 20, 33, 59, 115, 249, and 437 seconds for sample sizes of

Figure 6. (Color online) The Out-of-Sample Performance of CLX and WASS on a Portfolio Selection Problem



Note. The boundary of the shaded area represents the 20th and 80th percentiles, and the solid line represents the mean from 100 independent trials.

30, 60, 120, 240, 480, and 960, respectively, and the average computation times of WASS are 0.002, 0.002, 0.004, 0.008, 0.031, and 0.112 seconds for sample sizes of 30, 60, 120, 240, 480, and 960, respectively.

6. Conclusions

In this paper, we consider a DRO problem and propose data-driven approaches to construct ambiguity sets that only consist of absolutely continuous distributions. Such ambiguity sets are constructed using confidence bands of the density functions. Two different existing techniques in statistics literature are borrowed to create the confidence bands. We show that the optimal objective values of the DRO problems formulated using these two types of confidence bands converge to the optimal objective value of the corresponding SP based on the true distribution. The DRO problem we formulate is a continuous linear program. We then propose a stochastic gradient decent method to solve it. Numerical experiments in a newsvendor problem and a portfolio selection problem verify the effectiveness of our approach.

Acknowledgments

The authors thank Aditya Guntuboyina for referring us to the papers about confidence band constructions. The authors also thank the two anonymous reviewers for their suggestions and comments, which were very helpful in improving this manuscript.

Appendix A. Proof of Theorem 2

For $\delta > 0$, we define a subset of Ξ as

$$\Xi_\delta := \{\xi \in \Xi | p^\star(\xi) > \delta\}.$$

We first bound the differences between u_α, l_α and $p^\star$ at a given point $\xi \in \Xi_\delta$. We assume $\xi > \mu$ first, and the proof for $\xi < \mu$ is similar. We don't need to bound the differences at $\xi = \mu$ because the point μ has a zero measure so that it does not affect the Lebesgue integrals appearing in this theorem. For simplicity of notation, we use c^+ and c^- to represent $c^-(\alpha)$ and $c^-(\alpha)$ in this proof.

There exists an index i such that $\hat{\xi}_{(k_i)} < \xi < \hat{\xi}_{(k_{i+1})}$. Because $\xi > \mu$ and $p^\star$ is (C, ρ)-Hölder continuous, there exists N_δ independent of ξ such that, for $N \geq N_\delta$, we further have $p^\star(\hat{\xi}_{(k_{i+1})}) \geq \frac{\delta}{2}$ (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 $p^\star$ is monotonically decreasing over $[\hat{\xi}_{(k_i)}, \hat{\xi}_{(k_{i+1})}]$. By the definitions of $l_\alpha^{SR}(\xi)$ and $u_\alpha^{SR}(\xi)$ in (16) and the second line of the constraints in (15), we must have

$$u_\alpha^{SR}(\xi) \leq \frac{c^+}{\hat{\xi}_{(k_i)} - \hat{\xi}_{(k_{i-1})}} \text{ and } l_\alpha^{SR}(\xi) \geq \frac{c^-}{\hat{\xi}_{(k_{i+1})} - \hat{\xi}_{(k_i)}},$$

which implies

$$u_\alpha^{SR}(\xi) - l_\alpha^{SR}(\xi) := D_\xi \le \frac{c^+}{\hat{\xi}_{(k_i)} - \hat{\xi}_{(k_{i-1})}} - \frac{c^-}{\hat{\xi}_{(k_{i+1})} - \hat{\xi}_{(k_i)}}. \tag{A.1}$$

By the monotonicity of $p^\star$ on $[\mu, b]$, we can show that

$$(\hat{\xi}_{(k_i)} - \hat{\xi}_{(k_{i-1})})p^\star(\hat{\xi}_{(k_i)}) \le \Delta_i = F^\star(\hat{\xi}_{(k_i)}) - F^\star(\hat{\xi}_{(k_{(i-1)})}) \le (\hat{\xi}_{(k_i)} - \hat{\xi}_{(k_{i-1})})p^\star(\hat{\xi}_{(k_{i-1})}), \tag{A.2}$$

$$(\hat{\xi}_{(k_{i+1})} - \hat{\xi}_{(k_i)})p^\star(\hat{\xi}_{(k_{i+1})}) \le \Delta_{i+1} = F^\star(\hat{\xi}_{(k_{i+1})}) - F^\star(\hat{\xi}_{(k_i)}) \le (\hat{\xi}_{(k_{i+1})} - \hat{\xi}_{(k_i)})p^\star(\hat{\xi}_{(k_i)}). \tag{A.3}$$

Applying (A.2) and (A.3) to (A.1), we obtain

$$\begin{aligned}
D_\xi &\le \left| \frac{c^+ p^\star(\hat{\xi}_{(k_{i-1})})}{\Delta_i} - \frac{c^- p^\star(\hat{\xi}_{(k_{i+1})})}{\Delta_{i+1}} \right| \\
&\le p^\star(\hat{\xi}_{(k_{i+1})}) \left| \frac{c^+}{\Delta_i} - \frac{c^-}{\Delta_{i+1}} \right| + \frac{c^+}{\Delta_i} \left| p^\star(\hat{\xi}_{(k_{i+1})}) - p^\star(\hat{\xi}_{(k_{i-1})}) \right| \\
&\le U \left| \frac{c^+}{\Delta_i} - \frac{c^-}{\Delta_{i+1}} \right| + \frac{Cc^+}{\Delta_i} \left| \hat{\xi}_{(k_{i+1})} - \hat{\xi}_{(k_{i-1})} \right|^\rho \\
&\le U \left| \frac{c^+}{\Delta_i} - \frac{c^-}{\Delta_{i+1}} \right| + \frac{Cc^+}{\Delta_i} \left| \frac{\Delta_i + \Delta_{i+1}}{p^\star(\hat{\xi}_{(k_{i+1})})} \right|^\rho \\
&\le U \left| \frac{c^+}{\Delta_i} - \frac{c^-}{\Delta_{i+1}} \right| + \frac{Cc^+}{\Delta_i} \left| \frac{\Delta_i + \Delta_{i+1}}{\delta/2} \right|^\rho,
\end{aligned} \tag{A.4}$$

where the second inequality is by the triangle inequality, the third by Assumption 1(A3) and the (C, ρ)-Hölder continuity of $p^\star$, the fourth by (A.2) and (A.3) as well as the fact that $p^\star(\hat{\xi}_{(k_{i+1})}) \le p^\star(\hat{\xi}_{(k_i)})$, and the last by the fact that $p^\star(\hat{\xi}_{(k_{i+1})}) \ge \frac{\delta}{2}$.

The proof shows that (A.4) holds when $\xi \in \Xi_\delta$ and $\xi > \mu$. Using a similar argument, we can also show that (A.4) holds when $\xi \in \Xi_\delta$ and $\xi < \mu$. As a result, (A.4) holds for any $\xi \in \Xi_\delta$ except $\xi = \mu$. However, we don't need to upper bound D_ξ when $\xi = \mu$ because the point μ has a zero measure so that it does not affect the Lebesgue integrals that we use in the rest of the proof.

According to equations (98) and (101) with ($\tau = 2$) in Hengartner and Stark (1995), we have

$$\lim_{N\to+\infty} \mathbb{P}\left(\left| \frac{c^+}{\Delta_i} - \frac{c^-}{\Delta_{i+1}} \right| \le 4\sqrt{\frac{\log(N/K)}{K}} \right) \ge 1 - 2\theta \tag{A.5}$$

and

$$\lim_{N\to+\infty} \mathbb{P}\left(\frac{c^+}{\Delta_i}(\Delta_i + \Delta_{i+1})^\rho \le 2\left(\frac{K}{N}\right)^\rho \right) \ge 1 - 2\theta. \tag{A.6}$$

for any $\theta \in (0,1)$ and for all $i = 2, 3, \ldots, M$. We then apply (A.5) and (A.6) to (A.4) to achieve

$$\lim_{N\to+\infty} \mathbb{P}\left(D_\xi \le 4U\sqrt{\frac{\log(N/K)}{K}} + \frac{2^{\rho+1}C}{\delta^\rho}\left(\frac{K}{N}\right)^\rho \right) \ge 1 - 4\theta, \quad \forall \xi \in \Xi_\delta \backslash \{\mu\}. \tag{A.7}$$

Because (A.7) holds for any $\theta \in (0,1)$, using a union bound, we can show that, for any finite set $\Xi' \subset \Xi_\delta$,

$$\lim_{N\to+\infty} \mathbb{P}\left(\max_{\xi\in\Xi'} D_\xi \le 4U\sqrt{\frac{\log(N/K)}{K}} + \frac{2^{\rho+1}C}{\delta^\rho}\left(\frac{K}{N}\right)^\rho \right) \ge 1 - 4\theta \tag{A.8}$$

for any $\theta \in (0,1)$, which further implies

$$\lim_{N\to+\infty} \mathbb{P}\left(\int_{\Xi_\delta} |u_\alpha^{SR}(\xi) - l_\alpha^{SR}(\xi)| d\xi \le (b-a)\left[4U\sqrt{\frac{\log(N/K)}{K}} + \frac{2^{\rho+1}C}{\delta^\rho}\left(\frac{K}{N}\right)^\rho \right] \right) \ge 1 - 4\theta \tag{A.9}$$

for any $\theta \in (0,1)$. Let $\mathcal{M}(\cdot)$ represent the Lebesgue measure. We then set δ small enough such that

$$\max_{x\in\mathcal{X},\,\xi\in\Xi}|f(x,\xi)|\cdot U\cdot\mathcal{M}(\Xi\backslash\Xi_\delta)\le\frac{\epsilon}{4}. \tag{A.10}$$

Because $p^\star\in\mathcal{D}^{SR}(\hat{\Xi}_N,\alpha)$ with a probability of $1-\alpha$, we have

$$\begin{aligned}
&\max_{x\in\mathcal{X}}\left|\int_\Xi f(x,\xi)p^\star(\xi)d\xi-\sup_{p\in\mathcal{D}^{SR}(\hat{\Xi}_N,\alpha)}\int_\Xi f(x,\xi)p(\xi)d\xi\right|\\
&\le\max_{x\in\mathcal{X},\,\xi\in\Xi}|f(x,\xi)|\cdot\int_\Xi|u_\alpha^{SR}(\xi)-l_\alpha^{SR}(\xi)|d\xi\\
&=\max_{x\in\mathcal{X},\,\xi\in\Xi}|f(x,\xi)|\cdot\int_{\Xi_\delta}|u_\alpha^{SR}(\xi)-l_\alpha^{SR}(\xi)|d\xi+\max_{x\in\mathcal{X},\,\xi\in\Xi}|f(x,\xi)|\int_{\Xi\backslash\Xi_\delta}|u_\alpha^{SR}(\xi)-l_\alpha^{SR}(\xi)|d\xi\\
&\le\max_{x\in\mathcal{X},\,\xi\in\Xi}|f(x,\xi)|\cdot\int_{\Xi_\delta}|u_\alpha^{SR}(\xi)-l_\alpha^{SR}(\xi)|d\xi+2\max_{x\in\mathcal{X},\,\xi\in\Xi}|f(x,\xi)|\cdot U\cdot\mathcal{M}(\Xi\backslash\Xi_\delta)
\end{aligned}$$

with a probability of $1-\alpha$. Here, the first inequality is because $p^\star\in\mathcal{D}^{SR}(\hat{\Xi}_N,\alpha)$ and the second is because $0\le l_\alpha^{SR}(\xi)\le u_\alpha^{SR}(\xi)\le U$. Using a union bound, the preceding inequality, (A.9), and (A.10) together imply that

$$\lim_{N\to+\infty}\mathbb{P}\left(\begin{array}{l}\max\limits_{x\in\mathcal{X}}\left|\int_\Xi f(x,\xi)p^\star(\xi)d\xi-\sup\limits_{p\in\mathcal{D}^{SR}(\hat{\Xi}_N,\alpha)}\int_\Xi f(x,\xi)p(\xi)d\xi\right|\\ \le\ \max\limits_{x\in\mathcal{X},\,\xi\in\Xi}|f(x,\xi)|(b-a)\left[4U\sqrt{\dfrac{\log(N/K)}{K}}+\dfrac{2^{\rho+1}C}{\delta^\rho}\left(\dfrac{K}{N}\right)^\rho\right]+\dfrac{\epsilon}{2}\end{array}\right)\ge1-4\theta-\alpha,$$

for any $\theta\in(0,1)$. Because

$$\lim_{N\to+\infty}\left[4U\sqrt{\frac{\log(N/K)}{K}}+\frac{2^{\rho+1}C}{\delta^\rho}\left(\frac{K}{N}\right)^\rho\right]=0,$$

we have, for any ϵ,

$$\lim_{N\to+\infty}\mathbb{P}\left(\max_{x\in\mathcal{X}}\left|\int_\Xi f(x,\xi)p^\star(\xi)d\xi-\sup_{p\in\mathcal{D}^{SR}(\hat{\Xi}_N,\alpha)}\int_\Xi f(x,\xi)p(\xi)d\xi\right|\le\epsilon\right)\ge1-\alpha,$$

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

Appendix B. Proof of Theorem 4

Let $\mathcal{M}(\cdot)$ represent the Lebesgue measure on $\mathbb{R}^m$. Suppose $p^\star\in\mathcal{D}^{KDE}(\hat{\Xi}_N,\alpha)$. We have

$$\begin{aligned}
&\max_{x\in\mathcal{X}}\left|\int_\Xi f(x,\xi)p^\star(\xi)d\xi-\sup_{p\in\mathcal{D}^{KDE}(\hat{\Xi}_N,\alpha)}\int_\Xi f(x,\xi)p(\xi)d\xi\right|\\
&\le\max_{x\in\mathcal{X}}\int_\Xi|f(x,\xi)||u_\alpha^{KDE}(\xi)-l_\alpha^{KDE}(\xi)|d\xi\\
&\le\max_{x\in\mathcal{X}}\int_\Xi|f(x,\xi)|2\delta d\xi\le2\delta\max_{x\in\mathcal{X},\,\xi\in\Xi}|f(x,\xi)|\mathcal{M}(\Xi),
\end{aligned}$$

where δ is defined in (20). By the definition of δ and the fact that $h=(\log(N/\alpha)/N)^{1/(2\rho+m)}$, we have

$$\lim_{N\to+\infty}2\delta\max_{x\in\mathcal{X},\,\xi\in\Xi}|f(x,\xi)|\mathcal{M}(\Xi)=0.$$

As a result of Theorem 3, we have

$$\lim_{N\to+\infty}\mathbb{P}\left(\max_{x\in\mathcal{X}}\left|\int_\Xi f(x,\xi)p^\star(\xi)d\xi-\sup_{p\in\mathcal{D}^{KDE}(\hat{\Xi}_N,\alpha)}\int_\Xi f(x,\xi)p(\xi)d\xi\right|\le\epsilon\right)\ge\lim_{N\to+\infty}\mathbb{P}\left(p^\star\in\mathcal{D}^{KDE}(\hat{\Xi}_N,\alpha)\right)\ge1-\alpha.$$

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

Appendix C. Proof of Lemma 3

We first reformulate (27) as the linear conic program considered in Shapiro (2001). We consider the Banach space of Lebesgue integrable functions on Ξ, denoted by $X=\mathcal{L}_1(\Xi)$, and its dual space of Lebesgue integrable essentially bounded functions on Ξ, denoted by $X'=\mathcal{L}_\infty(\Xi)$. In addition, we define $\mathcal{L}_1^+(\Xi)$ as the cone of the almost surely nonnegative functions in $\mathcal{L}_1(\Xi)$. Note that u_α and l_α are in $\mathcal{L}_1^+(\Xi)$ because of Assumption 3(A1) and $f(x,\xi)\in\mathcal{L}_\infty(\Xi)$ by the assumption that

$\max_{\xi\in\Xi}|f(x,\xi)|<+\infty$. Let the bilinear form $\langle\cdot,\cdot\rangle: X'\times X\to\mathbb{R}$ be defined as $\langle h,p\rangle=\int_{\Xi}h(\xi)p(\xi)d\xi$. Then, we define the following parameterized linear conic program

$$\begin{aligned} V(y) &:= \inf_{p\in X}\langle f(x,\cdot),p\rangle \\ \text{s.t}\quad & \mathrm{A(p)+y}\in\mathrm{K}, \end{aligned} \tag{C.1}$$

where $A: X\to Y:=\mathbb{R}\times\mathcal{L}_1(\Xi)\times\mathcal{L}_1(\Xi)$ is the linear mapping

$$A(p):=\left(\int_{\Xi}p(\xi)d\xi,-p,p\right),$$

$y=(\delta,\mu,\nu)\in Y$, and $K=\{0\}\times\mathcal{L}_1^+(\Xi)\times\mathcal{L}_1^+(\Xi)$ is a cone in Y. Note that (27) is exactly (C.1) with $b=(-1,u_\alpha,-l_\alpha)\in Y$. Also note that the linear mappings $\langle f(x,\cdot),p\rangle$ and $A(p)$ are continuous, and the cone K is closed. We equip the space Y with the norm $\|y\|=\sqrt{\delta^2+\|\mu\|_1^2+\|\nu\|_1^2}$ for $y=(\delta,\mu,\nu)\in Y$. Let

$$D:=\|f(x,\cdot)\|_\infty=\max_{\xi\in\Xi}|f(x,\xi)|.$$

Consider any $y=(\delta,\mu,\nu)\in Y$. We claim that the set

$$S_y:=\{p\in X:\langle f(x,\cdot),p\rangle\le D, A(p)+y\in K\}$$

is nonempty when $y=b$ and that S_y is contained in the compact set $\{p\in X:\|p\|_1\le\|u_\alpha\|_1+\epsilon\}$ if y is in the ϵ-neighborhood of b, namely

$$\sqrt{(\delta+1)^2+\|\mu-u_\alpha\|_1^2+\|\nu+l_\alpha\|_1^2}\le\epsilon..$$

In fact, the set $\mathcal{D}(\hat{\Xi}_N,\alpha)$ is exactly the set of functions p satisfying $A(p)+b\in K$, and it is nonempty by Assumption 3(A2). Moreover, each $p\in\mathcal{D}(\hat{\Xi}_N,\alpha)$ satisfies $\langle f(x,\cdot),p\rangle\le D$ by the Cauchy–Schwartz inequality. Hence, S_y is nonempty when $y=b$. If y is in the ϵ-neighborhood of b, by the constraint $-\nu\le p\le\mu$ satisfied by any $p\in S_y$, we have

$$\|p\|_1\le\|\max(|\nu|,|\mu|)\|_1\le\epsilon+\|u_\alpha\|_1.$$

As a result, all conditions in proposition 2.4 in Shapiro (2001) hold such that the set of optimal solutions of (27) is nonempty and compact and $V(y)$ is lower semicontinuous at $y=b$. By proposition 2.3 in Shapiro (2001), the optimal objective value of the dual problem (29) equals $\mathrm{lsc}V(b)$, that is, the value of the lower semicontinuous hull of the function V at b. Because $V(y)$ is lower semicontinuous at $y=b$, we have $\mathrm{lsc}V(b)=V(b)$, where $V(b)$ is the optimal objective value of (27) according to (C.1). Hence, strong duality holds between (27) and (29).

Appendix D. Proof of Theorem 5

According to Lemma 4, there exists a Lebesgue integrable mapping $f'(x,\xi):\mathcal{X}\times\Xi\to\mathbb{R}^d$ such that (32) and (33) hold. Therefore, for any $x\in\mathcal{X}$ and a batch size B, if we generate a sample $\{\xi_1,\xi_2,\ldots,\xi_B\}$ from a uniform distribution over I and construct

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

we must have

$$\begin{aligned} \mathbb{E}(g_x) &= \int l(\xi)f'(x,\xi)\mathbb{I}_{f(x,\xi)<\lambda}(\xi)d\xi+\int u(\xi)f'(x,\xi)\mathbb{I}_{f(x,\xi)\ge\lambda}(\xi)d\xi\ \in\ \partial_x F(x,\lambda) \\ \mathbb{E}(g_\lambda) &= 1-\int l(\xi)\mathbb{I}_{f(x,\xi)<\lambda}(\xi)d\xi-\int u(\xi)\mathbb{I}_{f(x,\xi)\ge\lambda}(\xi)d\xi\ \in\ \partial_\lambda F(x,\lambda), \end{aligned}$$

where the expectation is taken over the sample $\{\xi_1,\xi_2,\ldots,\xi_B\}$. Hence, assumptions A1 and A2 in Nemirovski et al. (2009) are satisfied. By the assumption of Theorem 5, there exist a constant M such that $\mathbb{E}\|(g_x^k,g_\lambda^k)\|_2^2\le M^2$ for all k, which ensures condition (2.40) in holds Nemirovski et al. (2009). Because assumptions A1, A2, and (2.40) in Nemirovski et al. (2009) hold, the convergence property (2.48) in Nemirovski et al. (2009) holds with $\theta=1$ and $\alpha=1$. Applying (2.48) in Nemirovski et al. (2009) to the sequence $(\bar{x}_k,\bar{\lambda}_k)$ from Algorithm 3 yields the second inequality of (36). Note that the first inequality in (36) is because of (30), which indicates $v_D(x)\le F(x,\lambda)$ for any λ.

Appendix E. Behavior of $u_\alpha^{SR}(\xi)$ Near μ in Figures 3 and 4

In this section, we provide an explanation of the spikes in the upper bands u_α^{SR} in Figures 3 and 4 near the location of mode μ. We only need to focus on the case in which $\xi < \mu$ because the case in which $\xi > \mu$ can be explained similarly by symmetry.

Recall that $\hat{N} \in \mathbb{Z}_+$ is the number of distinct elements in the set $\{a, b, \mu, \xi, \hat{\xi}_{(k_1)}, \ldots, \hat{\xi}_{(k_M)}\}$, and z_j is the jth smallest element in this set. Recall also that $\hat{\jmath}$ and $\tilde{\jmath}$ are the indexes such that $z_{\hat{\jmath}} = \xi$ and $z_{\tilde{\jmath}} = \mu$. According to the definition of $u_\alpha^{SR}(\xi)$ in (15) and (16), we have $u_\alpha^{SR}(\xi) = \max_{\boldsymbol{\beta} \in \mathcal{H}(\hat{\Xi}_N, \alpha)} \beta_{\hat{\jmath}}$, which is a linear program. Suppose ξ takes a value, say ξ', which is less than μ but close enough to μ so that $\tilde{\jmath} - 1 = \hat{\jmath}$. The constraints involving variable $\beta_{\hat{\jmath}}$ in this linear program are as follows:

$$c^-(\alpha) \le \beta_{\hat{\jmath}-1}(\xi' - z_{\hat{\jmath}-1}) + \beta_{\hat{\jmath}}(\mu - \xi') + \beta_{\tilde{\jmath}}(z_{\tilde{\jmath}+1} - \mu) \le c^+(\alpha), \tag{E.1}$$

$$\sum_{j=1}^{\hat{N}-1} \beta_j (z_{j+1} - z_j) = 1, \quad \beta_{j-1} \le \beta_j, \quad 0 \le \beta_j \le U. \tag{E.2}$$

Suppose there exists a feasible solution $\boldsymbol{\beta}' = (\beta_1', \ldots, \beta_{\hat{N}-1}')$ for this linear program that satisfies $\beta_{\hat{\jmath}-1}' < \beta_{\hat{\jmath}}'$. We define

$$c(\xi') := \beta_{\hat{\jmath}-1}'(\xi' - z_{\hat{\jmath}-1}) + \beta_{\hat{\jmath}}'(\mu - \xi').$$

Then, for any $\xi \in (\xi', \mu)$, we define a solution $\boldsymbol{\beta}^\xi := (\beta_1^\xi, \ldots, \beta_{\hat{N}-1}^\xi)$, where $\beta_j^\xi = \beta_j'$ for $j \ne \hat{\jmath}$ and

$$\beta_{\hat{\jmath}}^\xi = \frac{c(\xi') - \beta_{\hat{\jmath}-1}^\xi(\xi - z_{\hat{\jmath}-1})}{\mu - \xi} = \frac{\beta_{\hat{\jmath}-1}'(\xi' - \xi) + \beta_{\hat{\jmath}}'(\mu - \xi')}{\mu - \xi} = \frac{\beta_{\hat{\jmath}-1}'(\mu - \xi) + (\beta_{\hat{\jmath}}' - \beta_{\hat{\jmath}-1}')(\mu - \xi')}{\mu - \xi}.$$

With this construction, we have $\beta_{\hat{\jmath}-1}^\xi(\xi - z_{\hat{\jmath}-1}) + \beta_{\hat{\jmath}}^\xi(\mu - \xi) = c(\xi')$ for any $\xi \in (\xi', \mu)$ so that all constraints in (E.1) and (E.2) remain satisfied except the constraint $\beta_{\hat{\jmath}} \le U$. Note that, as a conservative global upper bound of $p^*(\xi)$, U is a relatively large number. By definition, $u_\alpha^{SR}(\xi) \ge \beta_{\hat{\jmath}}^\xi$ and $\beta_{\hat{\jmath}}^\xi$ increases to infinity as ξ increases to μ. Hence, the upper band $u_\alpha^{SR}(\xi)$ increases to U as ξ approaches μ from the left. This explains the spike at μ in Figures 3 and 4. Although U is large, it is still a finite number so the height of the spikes in both figures are U instead of infinity. However, we do not include the peaks of the spikes in the figure in order to display the curves in a readable scale.

Endnotes

[1] In this paper, the notations $\xi^\star \sim P$ and $\xi^\star \sim p$ mean random variable $\xi^\star$ follows distribution P and has density function p, respectively.

[2] In this paper, an integral in the form of $\int (\cdot) d\xi$ represents the Lebesgue integral.

[3] Note that $p^*(\xi)$ can still be zero on Ξ.vf.

[4] Equation (104) in Hengartner and Stark (1995) was stated slightly differently from (17) by replacing "lim" by "lim inf" and "= 1" by "> 0" in (17). That equation was obtained by choosing the parameter τ^2 in their proof to be $2 + 2\rho$. However, the same proof leads to Lemma 1 once we choose $\tau^2 = 3 + 2\rho$ in their proof.

[5] Note that $\int_0^\infty \kappa(t) t^{m+\rho} dt < +\infty$ because of Assumption 2(A2).

References

Ben-Tal A, den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Sci.* 59(2):341–357.

Bertsimas D, Popescu I (2005) Optimal inequalities in probability theory: A convex optimization approach. *SIAM J. Optim.* 15(3):780–804.

Bertsimas D, Gupta V, Kallus N (2014) Robust sample average approximation. *Math. Programming* 171(1):217–282.

Bertsimas D, Gupta V, Kallus N (2018) Data-driven robust optimization. *Math. Programming* 167(2):235–292.

Birge JR, Louveaux F (2011) *Introduction to Stochastic Programming* (Springer Science & Business Media, New York).

Calafiore GC, El Ghaoui L (2006) On distributionally robust chance-constrained linear programs. *J. Optim. Theory Appl.* 130(1):1–22.

Chen YC (2019) Lecture 2: Density estimation. STAT 535: Statistical Machine Learning. Accessed January 22, 2019, http://faculty.washington.edu/yenchic/19A_stat535/Lec2_density.pdf.

Chen Z, Sim M, Xu H (2019) Distributionally robust optimization with infinitely constrained ambiguity sets. *Oper. Res.* 67(5):1328–1344.

DasGupta A (2019) Finite sample theory of order statistics and extremes. Lecture note. Accessed January 22, 2019, https://www.stat.purdue.edu/dasgupta/orderstats.pdf.

de Klerk E, Kuhn D, Postek K (2019) Distributionally robust optimization with polynomial densities: Theory, models and algorithms. *Math. Programming* 181(2):265–296.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

Duchi J, Glynn P, Namkoong H (2021) Statistics of robust optimization: A generalized empirical likelihood approach. *Math. Oper. Res.* Forthcoming.
Dupačová J (1987) The minimax approach to stochastic programming and an illustrative application. *Stochastics* 20(1):73–88.
El Ghaoui L, Oks M, Oustry F (2003) Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Oper. Res.* 51(4):543–556.
Erdoğan E, Iyengar G (2006) Ambiguous chance constrained problems and robust optimization. *Math. Programming* 107(1–2):37–61.
Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with Wasserstein distance. Preprint, submitted April 8, https://arxiv.org/abs/1604.02199.
Gao R, Chen X, Kleywegt AJ (2017) Wasserstein distributional robustness and regularization in statistical learning. Preprint, submitted December 17, https://arxiv.org/abs/1712.06050.
Hanasusanto GA, Kuhn D, Wallace SW, Zymler S (2015) Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Math. Programming* 152(1):1–32.
Hartigan JA, Hartigan PM (1985) The dip test of unimodality. *Ann. Statist.* 13(1):70–84.
Hengartner NW, Stark PB (1995) Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* 23(2):525–550.
Hu Z, Hong LJ (2013) *Kullback-Leibler Divergence Constrained Distributionally Robust Optimization* (Optimization Online).
Jiang H (2017) Uniform convergence rates for kernel density estimation. Precup D, Teh YW, eds. *Internat. Conf. Machine Learn.* (JMLR.org, Cambridge, MA), 1694–1703.
Jiang R, Guan Y (2015) Data-driven chance constrained stochastic program. *Math. Programming* 158(1):291–327.
Khas'minskii RZ (1976) A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory Probab. Appl.* 23(4): 794–798.
Klabjan D, Simchi-Levi D, Song M (2013) Robust stochastic lot-sizing by means of histograms. *Production Oper. Management* 22(3):691–710.
Lam H (2019) Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Oper. Res.* 67(4): 1090–1105.
Lam H, Mottet C (2017) Tail analysis without parametric models: A worst-case perspective. *Oper. Res.* 65(6):1696–1711.
Li B, Jiang R, Mathieu JL (2018) Ambiguous risk constraints with moment and unimodality information. *Math. Programming* 173(1):151–192.
Lofberg J (2004) YALMIP: A toolbox for modeling and optimization in MATLAB. *IEEE Internat. Sympos. Comput. Aided Control Systems Design* (IEEE Press, Piscataway, NJ), 284–289.
Mak WK, Morton DP, Wood RK (1999) Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Oper. Res. Lett.* 24(1–2):47–56.
Mevissen M, Ragnoli E, Yu JY (2013) Data-driven distributionally robust polynomial optimization. Burges C, ed. *Adv. Neural Inform. Processing Systems* 26 (Neural Information Processing Systems Foundation, Inc., La Jolla, CA), 37–45.
Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming* 171(1–2):115–166.
Natarajan K, Teo CP (2017) On reduced semidefinite programs for second order moment bounds with applications. *Math. Programming* 161(1): 487–518.
Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* 19(4):1574–1609.
Parzen E (1962) On estimation of a probability density function and mode. *Ann. Math. Statist.* 33(3):1065–1076.
Pflug G, Wozabal D (2007) Ambiguity in portfolio selection. *Quant. Finance* 7(4):435–442.
Rinaldo A, Wasserman L (2010) Generalized density clustering. *Ann. Statist.* 38(5):2678–2722.
Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. *J. Risk* 2:21–42.
Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* 27(3):832–837.
Scarf H (1958) A min-max solution of an inventory problem. Arrow K, Karlin S, Scarf H, eds. *Studies in the Mathematical Theory of Inventory and Production* (Stanford University Press, Stanford, CA), 201–209.
Scott DW (2015) *Multivariate Density Estimation: Theory, Practice, and Visualization* (John Wiley & Sons, Hoboken, NJ).
Shapiro A (2001) On duality theory of conic linear problems. Goberna M, López M, eds. *Semi-infinite Programming* (Springer, Boston), 135–165.
Shapiro A, Dentcheva D, Ruszczynski A (2014) *Lectures on Stochastic Programming: Modeling and Theory*, vol. 16, 2nd ed. (SIAM, Philadelphia).
Tsybakov AB (2008) *Introduction to Nonparametric Estimation* (Springer Science & Business Media, New York).
Vandenberghe L, Boyd S, Comanor K (2007) Generalized Chebyshev bounds via semidefinite programming. *SIAM Rev.* 49(1):52–64.
Wang Z, Glynn PW, Ye Y (2016) Likelihood robust optimization for data-driven problems. *Comput. Management Sci.* 13(2):241–261.
Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Oper. Res.* 62(6):1358–1376.
Zymler S, Kuhn D, Rustem B (2013a) Distributionally robust joint chance constraints with second-order moment information. *Math. Programming* 137(1–2):167–198.
Zymler S, Kuhn D, Rustem B (2013b) Worst-case value at risk for nonlinear portfolios. *Management Sci.* 59(1):172–188.