Bayesian Distributionally Robust Optimization

Alexander Shapiro, Enlu Zhou, Yifan Lin
School of Industrial and Systems Engineering
Georgia Institute of Technology

February 9, 2023

Abstract. We introduce a new framework, Bayesian Distributionally Robust Optimization (Bayesian-DRO), for data-driven stochastic optimization where the underlying distribution is unknown. Bayesian-DRO contrasts with most of the existing DRO approaches in the use of Bayesian estimation of the unknown distribution. To make computation of Bayesian updating tractable, Bayesian-DRO first assumes the underlying distribution takes a parametric form with unknown parameter and then computes the posterior distribution of the parameter. To address the model uncertainty brought by the assumed parametric distribution, Bayesian-DRO constructs an ambiguity set of distributions with the assumed parametric distribution as the reference distribution and then optimizes with respect to the worst case in the ambiguity set. We show the consistency of the Bayesian posterior distribution and subsequently the convergence of objective functions and optimal solutions of Bayesian-DRO. Our consistency result of the Bayesian posterior requires simpler assumptions than the classical literature on Bayesian consistency. We also consider several approaches for selecting the ambiguity set size in Bayesian-DRO and compare them numerically. Our numerical experiments demonstrate the out-of-sample performance of Bayesian-DRO in comparison with Kullback-Leibler-based (KL-) and Wasserstein-based empirical DRO as well as risk-neutral Bayesian Risk Optimization. Our numerical results shed light on how to choose the modeling framework (Bayesian-DRO, KL-DRO, Wasserstein-DRO) for specific problems, but the choice for general problems still remains an important and open question.

1 Introduction

Consider the following stochastic optimization problem

$$\min_{x\in\mathcal{X}} \mathbb{E}_Q[G(x,\xi)], \tag{1.1}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is a nonempty closed set, Q is a probability distribution of random vector ξ supported on $\Xi \subset \mathbb{R}^d$, and $G : \mathcal{X} \times \Xi \to \mathbb{R}$ is the cost function. The notation

$$\mathbb{E}_Q[Z] = \int_\Xi Z(\xi) dQ(\xi) \tag{1.2}$$

emphasizes that the expectation is taken with respect to the probability measure[1] (distribution) Q of random variable (measurable function) $Z : \Xi \to \mathbb{R}$. We use the same notation ξ viewed as random vector or as its realization, the particular meaning will be clear from the context.

In many applications, the underlying 'true' distribution of ξ is not known and should be derived (estimated) from the available data. A popular approach to deal with this distributional uncertainty is to construct an ambiguity set $\mathfrak{M}$ of probability distributions and to consider the following minimax (worst-case) counterpart of problem (1.1):

$$\min_{x \in \mathcal{X}} \sup_{Q \in \mathfrak{M}} \mathbb{E}_Q[G(x, \xi)]. \tag{1.3}$$

Such Distributionally Robust Optimization (DRO) approach to stochastic programming has a long history. In the setting of an inventory model, it was considered in the pioneering paper [23]. Various methods have been developed for construction of the ambiguity sets, such as methods based on moment constraints (e.g., [5]), ϕ-divergence (e.g. [2]), Wasserstein distance (e.g., [8]), and Bayesian guarantees [13].

A different approach is to fit a parametric family P_θ, $\theta \in \Theta$, of distributions to the (observed) data $(\xi_1, ..., \xi_N)$. We assume that the parameter set $\Theta \subset \mathbb{R}^k$ is closed, and that the parametric family is defined by density $f(\cdot|\theta)$. The value of the parameter vector θ is then estimated, say by the Maximum Likelihood method. This involves two approximations of the 'true' distribution. First, the parametric family is just a model, and as the famous quote is saying "every model is wrong, but some are useful". Second, the estimated value of the parameter vector may be not accurate especially when the available data are limited. The popular Bayesian approach is aimed at reducing variability of the parameter evaluation. That is, the parameter vector θ is assumed to be random whose probability distribution is supported on the set Θ and defined by a prior probability density $p(\theta)$. Then given the data (sample) $\boldsymbol{\xi}^{(N)} = (\xi_1, ..., \xi_N)$, the posterior distribution is determined by Bayes' rule

$$p(\theta|\boldsymbol{\xi}^{(N)}) = \frac{f(\boldsymbol{\xi}^{(N)}|\theta)p(\theta)}{\int_\Theta f(\boldsymbol{\xi}^{(N)}|\theta)p(\theta)d\theta}, \tag{1.4}$$

where $f(\boldsymbol{\xi}^{(N)}|\theta) = \prod_{i=1}^N f(\xi_i|\theta)$ is the conditional density of the sample by assuming ξ_i's are independent and identically distributed (i.i.d.).

Recently, [28] takes the Bayesian approach with the motivation to use the Bayesian posterior distribution (which encodes the likelihoods of all possibilities) to replace the ambiguity set (which treats every possibility inside the set with equal probability), and further take a risk functional with respect to the posterior distribution to allow more flexible risk attitude. This leads to the following Bayesian Risk Optimization (BRO) formulation

$$\min_{x \in \mathcal{X}} \rho_{\theta_N} \left(\mathbb{E}_{\xi|\theta}[G(x, \xi)] \right), \tag{1.5}$$

where ρ_{θ_N} is a risk functional (such as expectation, mean-variance, Value-at-Risk, Conditional Value-at-Risk) taken with respect to the posterior distribution $p(\theta|\boldsymbol{\xi}^{(N)})$, and $\mathbb{E}_{\xi|\theta}$ is the expectation taken with respect to the parametric distribution $f(\xi|\theta)$ conditional on θ. However, as mentioned above, the assumed parametric family introduces model uncertainty.

[1] Probability measure Q is defined on the sample (measurable) space $(\Xi, \mathcal{B})$, where $\mathcal{B}$ is the Borel sigma algebra of Ξ.

In this paper, we propose a new formulation termed Bayesian Distributionally Robust Optimization (Bayesian-DRO), which poses robustness against the model uncertainty (ambiguity) of the assumed parametric distributions while maintaining the advantage of Bayesian estimation when data are limited. It constructs an ambiguity set by taking the parametric distribution as the reference distribution and optimizes the worst-case of the Bayesian average of the true problem. More specifically, for every $\theta \in \Theta$ let $\mathfrak{M}^\theta$ be a set of probability measures on $(\Xi, \mathcal{B})$. We propose the following DRO formulation:

$$\min_{x\in\mathcal{X}} \mathbb{E}_{\theta_N}\left[\sup_{Q\in\mathfrak{M}^\theta} \mathbb{E}_{Q|\theta}[G(x,\xi)]\right], \tag{1.6}$$

where $\mathbb{E}_{Q|\theta}$ is the expectation with respect to distribution Q of ξ conditional on θ and

$$\mathbb{E}_{\theta_N}[Y] := \int_\Theta Y(\theta) p(\theta|\boldsymbol{\xi}^{(N)}) d\theta \tag{1.7}$$

denotes the expectation of random variable $Y : \Theta \to \mathbb{R}$ with respect to the posterior distribution $p(\theta|\boldsymbol{\xi}^{(N)})$. We refer to $\mathfrak{M}^\theta$ as the *ambiguity set*; a specific construction of the ambiguity sets will be discussed in the next section. Please note that the posterior distribution depends on choice of the prior density $p(\theta)$ and parametric model $f(\cdot|\theta)$. The choice of $p(\theta)$ and $f(\cdot|\theta)$ could both be subject to ambiguity. In this paper we mainly deal with ambiguity with respect to the parametric model. In Section 2.1.3 we give a brief discussion of modeling ambiguity of the posterior distribution, which of course also depends on ambiguity of the prior density.

We show the consistency of Bayesian posterior distributions. In particular, when the parametric model is mis-specified (i.e., when the true distribution lies outside the parametric family of distributions), the posterior distribution converges to the parametric distribution which has the minimum Kullback-Leibler (KL) divergence (within the parametric family) from the true distribution. Consistency of Bayesian posterior distribution under model mis-specification has been studied in the literature (e.g., [10, 14, 17]), but the assumptions required in our results are in general simpler and easier to verify than constructing a testing sequence as usually required in the existing literature. Built on this result, we show the objective functions and optimal solutions of Bayesian-DRO are strongly consistent.

When the ambiguity set is constructed using the KL divergence and its radius is small, we show that Bayesian-DRO is approximately equivalent to a weighted sum of the mean and standard deviation under the posterior distribution, where the weight depends on the size of the ambiguity set. This reveals that the robustness of Bayesian-DRO comes from the trade-off between the posterior mean and variability of the solution performance. Similar interpretation of robustness has also been observed in divergence-based empirical DRO (see [7, 11]), but the difference is that empirical DRO trades off the empirical mean and standard deviation (i.e., with respect to the empirical distribution) and in Bayesian-DRO these are with respect to the posterior distribution. To determine the ambiguity set size, we propose several theoretical and empirical methods and compare their performance numerically.

The rest of the paper is organized as follows. Section 2 formally introduces the Bayesian-DRO formulation, discusses the construction of the ambiguity set, and understands the robustness of Bayesian-DRO by sensitivity analysis. Section 3 analyzes convergence of Bayesian-DRO and

considers how to determine the size of the ambiguity set. Section 4 presents numerical results to illustrate the performance of Bayesian-DRO in comparison with empirical DRO as well as BRO-mean. Section 5 concludes the paper with a brief discussion of future work.

2 Bayesian distributionally robust optimization

The risk neutral formulation of the Bayesian counterpart of problem (1.1) can be written as

$$\min_{x\in\mathcal{X}} \left\{ g(x) := \mathbb{E}_{\theta_N}\left[\mathbb{E}_{\xi|\theta}[G(x,\xi)]\right]\right\}, \tag{2.1}$$

where the expectation $\mathbb{E}_{\xi|\theta}$ is taken with respect to the distribution of ξ conditional on θ, defined by density $f(\cdot|\theta)$, and the expectation $\mathbb{E}_{\theta_N}$ is taken with respect to the posterior distribution $p(\theta|\boldsymbol{\xi}^{(N)})$ defined in (1.4). Note that the nested expectation in (2.1) can be considered as the expectation with respect to the joint distribution of ξ and θ. An unbiased estimate of $g(x)$ can be obtained by generating a random realization of θ according to the posterior distribution $p(\theta|\boldsymbol{\xi}^{(N)})$ and then generating a random realization of $\xi \sim f(\cdot|\theta)$ conditional on generated θ. This allows to apply either the Sample Average Approximation (SAA) or Stochastic Approximation (SA) optimization methods for solving problem (2.1), provided that there is an efficient way to generate such random samples.

Now let us consider the uncertainty with respect to the choice of the parametric family of distributions of ξ, with a specified prior distribution of θ, which is often taken as an uninformative prior when there is no prior knowledge. We view (2.1) as the *nominal model* with observed (given) data $\boldsymbol{\xi}^{(N)}$, and the reference parametric family defined by the probability density function (pdf) $f(\cdot|\theta)$, $\theta \in \Theta$. We assume that the ambiguity set $\mathfrak{M}^\theta$ consists of probability measures defined by density functions, i.e., every distribution of the ambiguity set has respective pdf $q(\cdot|\theta)$, $\theta \in \Theta$. We also assume that the ambiguity set contains the nominal distribution. There are many ways how the ambiguity set can be constructed, and we will discuss a specific construction, well suited for our purposes, in Section 2.1 below.

By constructing an ambiguity $\mathfrak{M}^\theta$ for each fixed θ in (2.1), we define the Bayesian distributionally robust optimization problem (1.6), which is re-stated below for clarity:

$$\min_{x\in\mathcal{X}} \mathbb{E}_{\theta_N}\left[\sup_{Q\in\mathfrak{M}^\theta} \mathbb{E}_{Q|\theta}[G(x,\xi)]\right]. \tag{2.2}$$

For this problem, define the following distributionally robust functional

$$\mathfrak{R}(Z) := \mathbb{E}_{\theta_N}\left[\sup_{Q\in\mathfrak{M}^\theta} \mathbb{E}_{Q|\theta}[Z]\right]. \tag{2.3}$$

This functional is defined on an appropriate linear space of measurable functions (random variables) $Z : \Xi \to \mathbb{R}$. The functional $\mathfrak{R}$ can be viewed as a nested conditional functional. We can refer to [20] for a detailed discussion of such conditional functionals. For random variable $Z : \Xi \to \mathbb{R}$, the respective expectation in (2.3) is

$$\mathbb{E}_{Q|\theta}[Z] = \int_\Xi Z(\xi) q(\xi|\theta) d\xi, \tag{2.4}$$

where $q(\cdot|\theta)$ is the pdf of $Q \in \mathfrak{M}^\theta$. The maximum (supremum) in the right hand side of (2.3) is taken over all pdfs $q_\theta(\xi) = q(\xi|\theta)$ from the ambiguity set $\mathfrak{M}^\theta$.

The distributionally robust counterpart of problem (2.1) is obtained by employing the above distributionally robust functional. That is, problem (2.2) can be written as

$$\min_{x\in\mathcal{X}} \mathfrak{R}(G_x), \tag{2.5}$$

where $G_x(\xi) := G(x,\xi)$. Of course, it should be verified that the above distributionally robust functionals are well defined for every $Z = G_x$, $x \in \mathcal{X}$. We will discuss this in the next section.

Remark 2.1. In problem (2.2), if we take $\mathfrak{M}^\theta$ constant for all $\theta \in \Theta$, i.e., $\mathfrak{M}^\theta = \mathfrak{M}, \forall \theta \in \Theta$, then this problem becomes a DRO problem

$$\min_{x\in\mathcal{X}} \sup_{Q\in\mathfrak{M}} \mathbb{E}_Q[G(x,\xi)]. \tag{2.6}$$

Hence, $\{\mathfrak{M}^\theta, \theta \in \Theta\}$ in (2.2) can be viewed as a finer characterization of the ambiguity set based on the likelihood of each θ, whereas the usual DRO takes a "blanket" ambiguity set for every θ. Moreover, the outer expectation in (2.2) aggregates all $\theta \in \Theta$ by their posterior density rather than fixating on the worst case in the ambiguity set.

When $\mathfrak{M}^\theta$ is a singleton consisting of only $f(\cdot|\theta)$, then (2.2) reduces to (2.1) or BRO-mean (i.e., (1.5) with expectation being the risk functional). This implies that as opposed to BRO-mean, Bayesian-DRO imposes additional robustness with respect to the possibly misspecified likelihood.

2.1 Construction of the ambiguity set

Consider now construction of the ambiguity set for the parametric family. The functional

$$\varrho_{|\theta}(\cdot) := \sup_{Q\in\mathfrak{M}^\theta} \mathbb{E}_{Q|\theta}[\,\cdot\,] \tag{2.7}$$

can be viewed as a coherent risk measure conditional on $\theta \in \Theta$. We have that $\mathbb{E}_{Q|\theta}[Z]$ is a function of $\theta \in \Theta$ defined by the corresponding integral (see (2.4)) which is assumed to be well defined. It could happen that by taking the maximum (supremum) of such functions over possibly uncountable family of distributions, the resulting value $\varrho_{|\theta}(Z)$, considered as a function of $\theta \in \Theta$, is not measurable. In that case the corresponding integral, defining $\mathfrak{R}(Z)$, does not exist. We will deal with this issue in the specific construction below.

There are many ways how the ambiguity sets can be constructed. The following approach, of the so-called ϕ-divergence ([4],[18]), is general and flexible. Let $\phi : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ be a convex lower semi-continuous function such that $\phi(1) = 0$ and $\phi(x) = +\infty$ for $x < 0$. For $\epsilon \geq 0$ and $f_\theta(\xi) := f(\xi|\theta)$ define the corresponding set of pdfs $q_\theta(\xi) = q(\xi|\theta)$, representing the ambiguity set, as

$$\mathfrak{M}^\theta_\epsilon := \left\{ q_\theta : \int_\Xi \phi\big(q_\theta(\xi)/f_\theta(\xi)\big) f_\theta(\xi) d\xi \leq \epsilon \right\}. \tag{2.8}$$

That is, the ambiguity set consists of pdfs having ϕ-divergence $\leq \epsilon$ from the reference parametric pdf $f(\xi|\theta)$. Note that $\mathfrak{M}_\epsilon^\theta$ contains the reference measure (distribution) defined by the pdf $f(\xi|\theta)$. Note also that the probability measure defined by the pdf q_θ in (2.8) is assumed to be absolutely continuous with respect to the reference measure f_θ for every $\theta \in \Theta$.

Consider the conjugate $\phi^*(y) = \sup_{x\geq 0}\{yx - \phi(x)\}$ of ϕ. Note that the conjugate of $\lambda\phi(\cdot)$ is $(\lambda\phi)^*(y) = \lambda\phi^*(y/\lambda)$ for $\lambda > 0$. It can be shown by duality arguments (cf., [1],[2],[25]), that for a random variable $Z : \Xi \to \mathbb{R}$,

$$\varrho_{|\theta}(Z) = \inf_{\lambda\geq 0,\mu} \left\{\lambda\epsilon + \mu + \mathbb{E}_{\xi|\theta}\big[(\lambda\phi)^*(Z-\mu)\big]\right\}. \tag{2.9}$$

Hence, the functional (2.3) can be written as

$$\mathfrak{R}(Z) = \mathbb{E}_\theta\Big[\underbrace{\inf_{\lambda>0,\mu} \mathbb{E}_{\xi|\theta}\big[\lambda\epsilon + \mu + \lambda\phi^*\big((Z-\mu)/\lambda\big)\big]}_{\varrho_{|\theta}(Z)}\Big]. \tag{2.10}$$

The measurability of the infimum $\varrho_{|\theta}(Z)$ in the right hand side of (2.9), considered as a function of θ, can be verified under mild regularity conditions. For example, we have the following result.

Proposition 2.1. *Suppose that for almost every (with respect to the Lebesgue measure) ξ the density function $f(\xi|\theta)$ is lower semicontinuous in $\theta \in \Theta$. Then $\varrho_{|\theta}(Z)$ is measurable in θ.*

Proof. Since the conjugate function $\phi^*(\cdot)$ is lower semicontinuous and $f(\xi|\cdot)$ is lower semicontinuous, we have that for almost every ξ the function $\lambda\epsilon + \mu + \lambda\phi^*\big((Z(\xi)-\mu)/\lambda\big)f(\xi|\theta)$ is lower semicontinuous in (λ, μ, θ). It follows by Fatou's lemma that its integral

$$\mathbb{E}_{\xi|\theta}\big[\lambda\epsilon + \mu + \lambda\phi^*\big((Z-\mu)/\lambda\big)\big] = \int_\Xi [\lambda\epsilon + \mu + \lambda\phi^*\big((Z(\xi)-\mu)/\lambda\big)]f(\xi|\theta)d\xi$$

is lower semicontinuous in (λ, μ, θ) and hence is measurable. Therefore the above integral is a normal integrand [22, Corollary 14.41], and hence its infimum over $(\lambda, \mu) \in \mathbb{R}_+ \times \mathbb{R}$ is measurable [22, Theorem 14.37]. □

2.1.1 Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence from a pdf $q(\cdot)$ to a pdf $f(\cdot)$, on Ξ, is

$$D_{KL}(q\|f) := \int_\Xi q(\xi)\ln\big(q(\xi)/f(\xi)\big)d\xi = \int_\Xi (q(\xi)/f(\xi))\ln\big(q(\xi)/f(\xi)\big)f(\xi)d\xi. \tag{2.11}$$

The KL-divergence is a particular instance of the ϕ-divergence with

$$\phi(x) := x\ln x - x + 1,\ x \geq 0.$$

The corresponding ambiguity set $\mathfrak{M}_\epsilon^\theta$ is formed by pdfs q_θ such that $D_{KL}(q_\theta\|f_\theta) \leq \epsilon$. We will show that the KL-divergence approach is in accordance with the consistency of the Bayesian

posterior distribution in Section 3.1, and therefore is a natural approach to construction of the corresponding ambiguity set.

We make the following assumption in the remainder of the paper: for $x \in \mathcal{X}$ and $Z := G_x$ it follows that

$$\mathbb{E}_{\xi|\theta}\left[e^{tZ}\right] < +\infty \quad \text{for any } t \in \mathbb{R} \text{ and } \theta \in \Theta. \tag{2.12}$$

For the KL-divergence, given $\lambda > 0$ the minimizer over μ in (2.9) is given by $\mu = \lambda \ln \mathbb{E}_{\xi|\theta}\left[e^{Z/\lambda}\right]$, and hence the minimum becomes

$$\varrho_{|\theta}(Z) = \inf_{\lambda>0} \left\{\lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}\left[e^{Z/\lambda}\right]\right\}. \tag{2.13}$$

Consequently, the DRO problem (2.3) can be written as

$$\min_{x\in\mathcal{X}} \mathbb{E}_{\theta_N}\left[\inf_{\lambda>0} \left\{\lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}[e^{G_x/\lambda}]\right\}\right]. \tag{2.14}$$

The above optimization problem (2.14) can be viewed as a two-stage stochastic program with the second stage given by the optimization problem with respect to $\lambda > 0$. It can be solved, for example, by the Sample Average Approximation (SAA) method; we will discuss this further in Section 4.

2.1.2 Robustness via sensitivity analysis

We now consider the sensitivity of the Bayesian-DRO objective value with respect to ϵ, size of the ambiguity set. Note that for $\epsilon = 0$ the minimum (infimum) in (2.13) is attained as $\lambda \to +\infty$, and equals $\mathbb{E}_{\xi|\theta}[Z]$. For $\epsilon > 0$ the optimization problem (2.13) has unique optimal solution $\bar{\lambda}$, with $\bar{\lambda}$ tending to $+\infty$ as $\epsilon \downarrow 0$.

Consider minimization problem in the right hand side of (2.13) for $\theta \in \Theta$ and small $\epsilon > 0$. By condition (2.12), the log-moment generation function $\Lambda(t) := \ln \mathbb{E}_{\xi|\theta}\left[e^{tZ}\right]$ is finite valued and infinitely differentiable with its first and second derivatives at $t = 0$ being the respective mean and variance. Consequently, by using the second order Taylor expansion of the log-moment generating function, we can write

$$\lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}\left[e^{Z/\lambda}\right] = \lambda\epsilon + \mu + \tfrac{1}{2}\sigma^2/\lambda + O(\lambda^{-2}), \tag{2.15}$$

where[2] $\mu := \mathbb{E}_{\xi|\theta}[Z]$, $\sigma^2 := \mathrm{Var}_{\xi|\theta}(Z)$ by minimizing the right hand side of (2.15) we obtain approximation $\bar{\lambda} \approx \frac{\sigma}{\sqrt{2\epsilon}}$ of the optimal solution of (2.13), and consequently for small $\epsilon > 0$ the approximation

$$\min_{\lambda>0} \left\{\lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}\left[e^{Z/\lambda}\right]\right\} \approx \mu + \sigma\sqrt{2\epsilon}. \tag{2.16}$$

Plugging the approximation (2.16) into the Bayesian-DRO problem (2.14) reveals that when the ambiguity set is small, Bayesian-DRO is approximately equal to a weighted sum of the posterior mean and posterior standard deviation of the performance function, with weight depending on the ambiguity set size ϵ. A similar interpretation of mean-variance trade-off has also been observed for divergence-based empirical DRO (see [7, 11]), but its mean and standard deviation

[2]Of course, μ and σ depend on θ, we suppress this in the notation.

are with respect to the empirical distribution. Moreover, [12] shows that the empirical DRO can be interpreted as a trade-off between the mean and worst-case sensitivity (we refer the reader to [12] for the definition of worst-case sensitivity); whether such an interpretation can be extended to Bayesian-DRO will be left as a future work.

2.1.3 Variants of Bayesian-DRO Formulations

In this section we briefly discuss some other possible DRO formulations in the Bayesian setting and their pros and cons. We first consider the alternative formulation

$$\min_{x\in\mathcal{X},\lambda>0} \mathbb{E}_{\theta_N}\left[\lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}[\exp(G_x/\lambda)]\right]. \tag{2.17}$$

The nested Bayesian-DRO problem (2.14) can be viewed as a relaxation of problem (2.17). In (2.17) the parameter λ is chosen before observing a realization of θ, while in (2.14) the parameter λ is a function of θ. We have that the optimal value of the Bayesian-DRO problem (2.14) is less than or equal to the optimal value of problem (2.17). It could be noted that the relaxation (2.17) is computationally easier to solve than (2.14), since it avoids nested Monte Carlo simulation that is needed in solving the nested formulation (2.14).

Now let's consider another variant of formulation. As it was mentioned in section 1, the posterior distribution depends on the choice of the prior density and parametric family. In the above derivations we considered the ambiguity with respect to the reference parametric pdf $f(\cdot|\theta)$, and consequently the corresponding Bayesian-DRO problem (2.2). It is possible to apply the KL-divergence ambiguity approach to the posterior distribution rather than the parametric family. That is for $\epsilon > 0$ let $\mathcal{M}_\epsilon$ be the family of pdfs $\mathfrak{p}(\theta)$, $\theta \in \Theta$, such that $D_{KL}\left(\mathfrak{p}\|p(\cdot|\boldsymbol{\xi}^{(N)})\right) \leq \epsilon$. Let

$$\mathcal{R}(Y) := \sup_{\mathfrak{p}\in\mathcal{M}_\epsilon}\left\{\mathbb{E}_{\mathfrak{p}}[Y] = \int_\Theta Y(\theta)\mathfrak{p}(\theta)d\theta\right\} \tag{2.18}$$

be the corresponding distributionally robust functional defined on a space of random variables $Y : \Theta \to \mathbb{R}$. Similar to (2.13) we have the following representation of that functional

$$\mathcal{R}(Y) = \inf_{\lambda>0}\left\{\lambda\epsilon + \lambda \ln \mathbb{E}_{\theta_N}\left[e^{Y/\lambda}\right]\right\}. \tag{2.19}$$

The corresponding DRO problem is obtained by replacing the expectation $\mathbb{E}_{\theta_N}$ in (2.1) with $\mathcal{R}$, that is minimization of $\mathcal{R}\big(\mathbb{E}_{\xi|\theta}[G_x]\big)$ over $x \in \mathcal{X}$. By (2.19) we can write this optimization problem as

$$\min_{x\in\mathcal{X},\lambda>0} \lambda\epsilon + \lambda \ln \mathbb{E}_{\theta_N}\left[\exp\left(\mathbb{E}_{\xi|\theta}[G_x/\lambda]\right)\right]. \tag{2.20}$$

Now by interchanging the expectation $\mathbb{E}_{\theta_N}$ and the supremum in the definition (2.3) of the distributionally robust functional $\mathfrak{R}$, we can consider the functional

$$\Re(Z) := \sup_{Q\in\mathfrak{M}^\theta} \mathbb{E}_{\theta_N}\left[\mathbb{E}_{Q|\theta}[Z]\right] \tag{2.21}$$

and the corresponding Bayesian-DRO problem. We have that $\Re(\cdot) \leq \mathfrak{R}(\cdot)$ and the inequality can be strict since the extreme measure Q in (2.3) could depend on θ. The maximization in (2.21) is over the pdfs of the ambiguity set. Because the expectation with respect to these pdfs is inside the expectation $\mathbb{E}_{\theta_N}$, it is not clear how to represent the corresponding optimization problem in the KL-divergence framework. It is also not clear what could be an interpretation of the functional $\Re$ and the corresponding optimization problem.

3 Analysis

Suppose that the data $\xi_1, \ldots, \xi_N$ are generated i.i.d. from the true (data-generating) distribution Q_*, i.e., $\xi_i \overset{\text{iid}}{\sim} Q_*$, and that Q_* has density (pdf) denoted q_*. Recall that $p(\theta)$ denotes the prior pdf, $f(\xi|\theta)$ denotes the reference parametric family, and $p(\theta|\boldsymbol{\xi}^{(N)})$ denotes the posterior pdf as defined in (1.4).

3.1 Consistency of Bayesian posterior distributions

In this section we discuss convergence of the posterior pdf $p(\theta|\boldsymbol{\xi}^{(N)})$ as N goes to infinity. The analysis of this section is a first step in establishing consistency of the Bayesian-DRO, discussed in the next section. We make the following assumptions which are relatively easy to verify and well suited for the considered framework.

Assumption 3.1. (i) *The set Θ is convex compact with nonempty interior.* (ii) $\ln p(\theta)$ *is bounded on Θ, i.e., there are constants $c_1 > c_2 > 0$ such that $c_1 \geq p(\theta) \geq c_2$ for all $\theta \in \Theta$.* (iii) $q_*(\xi) > 0$ *for $\xi \in \Xi$.* (iv) $f(\xi|\theta) > 0$, *and hence $p(\theta|\boldsymbol{\xi}^{(N)}) > 0$, for all $\xi \in \Xi$ and $\theta \in \Theta$.* (v) $f(\xi|\theta)$ *is continuous in $\theta \in \Theta$.* (vi) $\ln f(\xi|\theta)$, $\theta \in \Theta$, *is dominated by an integrable (with respect to Q_*) function.*

Assumptions 3.1(i)-(ii) provide sufficient conditions for uniform convergence of the posterior distribution. Without these assumptions, convergence of the posterior still holds but may not be uniform. The rest of Assumption 3.1 are regularity assumptions.

Consider function

$$\psi(\theta) := \mathbb{E}_{q_*}\big[\ln f(\xi|\theta)\big] = \int_\Xi \ln f(\xi|\theta) Q_*(d\xi) = \int_\Xi q_*(\xi) \ln f(\xi|\theta) d\xi. \tag{3.1}$$

Under Assumption 3.1, the function $\psi : \Theta \to \mathbb{R}$ is real valued. Moreover, we have that for $\theta \in \Theta$,

$$\lim_{\theta' \to \theta} \psi(\theta') = \lim_{\theta' \to \theta} \int_\Xi \ln f(\xi|\theta') Q_*(d\xi) = \int_\Xi \lim_{\theta' \to \theta} \ln f(\xi|\theta') Q_*(d\xi) = \psi(\theta),$$

where we use continuity of $f(\xi|\theta)$ in θ, and the interchange of the limit and integral follows by the Dominated Convergence Theorem since $\ln f(\cdot|\theta)$ is dominated by an integrable function. Thus $\psi(\theta)$ is continuous on Θ.

Consider the KL-divergence

$$D_{KL}\big(q_* \| f_\theta\big) = \int_\Xi q_*(\xi) \ln\left(\frac{q_*(\xi)}{f(\xi|\theta)}\right) d\xi = \mathbb{E}_{q_*}[\ln q_*(\xi)] - \underbrace{\mathbb{E}_{q_*}[\ln f(\xi|\theta)]}_{\psi(\theta)}.$$

Let

$$\Theta^* := \arg\min_{\theta \in \Theta} D_{KL}(q_* \| f_\theta) = \arg\max_{\theta \in \Theta} \underbrace{\mathbb{E}_{q_*}[\ln f(\xi|\theta)]}_{\psi(\theta)}.$$

Since the set Θ is compact and $\psi(\cdot)$ is continuous, it follows that the set Θ^* is nonempty. Note that if the model is correct, then $\Theta^* = \{\theta \in \Theta : q_* = f_\theta\}$.

For a point $\theta^* \in \Theta^*$ and $\epsilon > 0$, define the sets

$$V_\epsilon := \{\theta \in \Theta : \psi(\theta^*) - \psi(\theta) \geq \epsilon\},\ U_\epsilon := \Theta \setminus V_\epsilon = \{\theta \in \Theta : \psi(\theta^*) - \psi(\theta) < \epsilon\}. \tag{3.2}$$

Since $\psi(\theta^*) = \max_{\theta\in\Theta} \psi(\theta)$, the sets V_ϵ and U_ϵ remain the same for any $\theta^* \in \Theta^*$. Note that U_ϵ is a neighborhood of the set Θ^*. Since the set Θ is convex with nonempty interior, it follows that volume $\int_{U_\epsilon} d\theta$, of the set U_ϵ, is greater than zero for any $\epsilon > 0$.

The following theorem shows that the posterior pdf $p(\theta|\boldsymbol{\xi}^{(N)})$ converges almost surely to a distribution with probability mass concentrated on Θ^*. If Θ^* is the singleton $\{\theta^*\}$, then $p(\theta|\boldsymbol{\xi}^{(N)})$ converges almost surely to the Dirac delta function $\delta(\theta^*)$. The convergence is uniform in $\theta \in \Theta$ regardless of the choice of the prior pdf $p(\theta)$. In what follows by writing w.p.1 (almost surely) we mean that the considered property holds with probability one with respect to the probability measure Q_*^∞. Construction of the probability measure Q_*^∞ for the sequence $\{\xi_1, ...\}$ is verified by Kolmogorov's existence theorem. By saying that: "a property holds w.p.1 for N large enough", we mean that there is a subset of the considered probability space having measure zero such that for any element of the probability space outside this measure-zero set, there is N' (depending on that element) such that the property holds for that element for any $N \geq N'$.

Lemma 3.1. *Suppose that Assumption 3.1 holds. Then for $0 < \beta < \alpha < \epsilon$, it follows that w.p.1 for N large enough*

$$\sup_{\theta\in V_\epsilon} p(\theta|\boldsymbol{\xi}^{(N)}) \leq \kappa(\beta)^{-1} e^{-N(\alpha-\beta)}, \tag{3.3}$$

where V_ϵ and U_ϵ are defined in (3.2), and[3] *$\kappa(\beta) := \int_{U_\beta} d\theta$.*

Proof. Define

$$\phi_N(\theta) := N^{-1} \ln f(\boldsymbol{\xi}^{(N)}|\theta) = N^{-1} \sum_{i=1}^{N} \ln f(\xi_i|\theta).$$

By the Law of Large Number (LLN) we have for $\theta \in \Theta$ that

$$\lim_{N\to\infty} \phi_N(\theta) = \psi(\theta),\ \text{w.p.1}. \tag{3.4}$$

Hence we can write

$$N^{-1} \ln[f(\boldsymbol{\xi}^{(N)}|\theta)] = \psi(\theta) + \varepsilon_N(\theta), \tag{3.5}$$

where $\varepsilon_N(\theta)$ tends to 0 w.p.1 for any $\theta \in \Theta$. Now for $\theta^* \in \Theta^*$ and $\theta \in V_\epsilon$ we have

$$\ln p(\theta^*|\boldsymbol{\xi}^{(N)}) - \ln p(\theta|\boldsymbol{\xi}^{(N)}) = \ln f(\boldsymbol{\xi}^{(N)}|\theta^*) - \ln f(\boldsymbol{\xi}^{(N)}|\theta) + \ln p(\theta^*) - \ln p(\theta). \tag{3.6}$$

It follows by (3.5) that

$$N^{-1}[\ln p(\theta^*|\boldsymbol{\xi}^{(N)}) - \ln p(\theta|\boldsymbol{\xi}^{(N)})] = \psi(\theta^*) - \psi(\theta) + \varepsilon_N(\theta^*) - \varepsilon_N(\theta) + N^{-1}[\ln p(\theta^*) - \ln p(\theta)]. \tag{3.7}$$

Consider a point $\theta \in V_\epsilon$. Then

$$N^{-1}[\ln p(\theta^*|\boldsymbol{\xi}^{(N)}) - \ln p(\theta|\boldsymbol{\xi}^{(N)})] \geq \epsilon + \gamma_N(\theta), \tag{3.8}$$

[3]Recall that under Assumption 3.1, $\kappa(\beta) > 0$ for any $\beta > 0$.

where $\gamma_N(\theta)$ tends to zero w.p.1. It follows that for any $\alpha \in (0, \epsilon)$, w.p.1 for N large enough

$$\ln p(\theta^*|\boldsymbol{\xi}^{(N)}) - \ln p(\theta|\boldsymbol{\xi}^{(N)}) \geq N\alpha, \tag{3.9}$$

or equivalently

$$e^{-N\alpha} p(\theta^*|\boldsymbol{\xi}^{(N)}) \geq p(\theta|\boldsymbol{\xi}^{(N)}). \tag{3.10}$$

In the similar way by using (3.7), we obtain for $\theta \in U_\epsilon$ and $\beta \in (0, \epsilon)$ that w.p.1 for N large enough

$$\ln p(\theta^*|\boldsymbol{\xi}^{(N)}) - \ln p(\theta|\boldsymbol{\xi}^{(N)}) \leq N\beta,$$

or equivalently

$$e^{-N\beta} p(\theta^*|\boldsymbol{\xi}^{(N)}) \leq p(\theta|\boldsymbol{\xi}^{(N)}). \tag{3.11}$$

Now let us show that w.p.1 for N large enough

$$p(\theta^*|\boldsymbol{\xi}^{(N)}) \leq e^{N\beta}/\kappa(\beta). \tag{3.12}$$

Indeed since $p(\theta|\boldsymbol{\xi}^{(N)})$ is a density we have

$$1 = \int_\Theta p(\theta|\boldsymbol{\xi}^{(N)}) d\theta \geq \int_{U_\beta} p(\theta|\boldsymbol{\xi}^{(N)}) d\theta \geq e^{-N\beta} \kappa(\beta) p(\theta^*|\boldsymbol{\xi}^{(N)}),$$

where for the last inequality we used (3.11) with $\kappa(\beta) = \int_{U_\beta} d\theta$.

By Assumption 3.1 the set Θ is compact and $\ln f(\xi|\theta)$, $\theta \in \Theta$, is dominated by an integrable (with respect to Q_*) function. Then by the uniform LLN (e.g., [26, Theorem 7.48]) the limit (3.4) can be strengthened to the uniform limit

$$\lim_{N\to\infty} \sup_{\theta\in\Theta} |\phi_N(\theta) - \psi(\theta)| = 0, \text{ w.p.1}, \tag{3.13}$$

i.e., $\varepsilon_N(\theta) = N^{-1} \ln[f(\boldsymbol{\xi}^{(N)}|\theta)] - \psi(\theta)$ tends to 0 w.p.1 uniformly in $\theta \in \Theta$. Assumption 3.1 further supposes that $\ln p(\theta)$ is bounded on Θ, i.e., there are constants $c_1 > c_2 > 0$ such that $c_1 \geq p(\theta) \geq c_2$ for all $\theta \in \Theta$. Then

$$N^{-1}[\ln p(\theta^*|\boldsymbol{\xi}^{(N)}) - \ln p(\theta|\boldsymbol{\xi}^{(N)})] = \psi(\theta^*) - \psi(\theta) + \eta_N(\theta), \tag{3.14}$$

where

$$\eta_N(\theta) := \varepsilon_N(\theta^*) - \varepsilon_N(\theta) + N^{-1}[\ln p(\theta^*) - \ln p(\theta)]$$

tends to 0 w.p.1 uniformly in $\theta \in \Theta$. Thus for any $\alpha \in (0, \epsilon)$ we have that w.p.1 for N large enough

$$\ln p(\theta^*|\boldsymbol{\xi}^{(N)}) \geq N\alpha + \sup_{\theta\in V_\epsilon} \ln p(\theta|\boldsymbol{\xi}^{(N)}). \tag{3.15}$$

By (3.12) it follows that for $0 < \beta < \alpha < \epsilon$, w.p.1 for N large enough

$$\sup_{\theta\in V_\epsilon} p(\theta|\boldsymbol{\xi}^{(N)}) \leq e^{-N\alpha} p(\theta^*|\boldsymbol{\xi}^{(N)}) \leq e^{-N(\alpha-\beta)}/\kappa(\beta). \tag{3.16}$$

This completes the proof. □

Let θ_N be random vector with the posterior pdf $p(\theta|\boldsymbol{\xi}^{(N)})$. We have that probability of the event $\{\theta_N \in V_\epsilon\}$ is given by the integral $\int_{V_\epsilon} p(\theta|\boldsymbol{\xi}^{(N)})d\theta$. Consequently under Assumption 3.1, we have by (3.3) that for any $\epsilon > 0$, w.p.1 for N large enough,

$$\text{Prob}\{\theta_N \in V_\epsilon\} \leq \kappa(\beta)^{-1}\nu e^{-N(\alpha-\beta)}, \tag{3.17}$$

where ν is volume of the set Θ. It follows that probability of the event $\{\theta_N \in U_\epsilon\}$ converges w.p.1 to one as $N \to \infty$. Note that for an appropriate $\epsilon > 0$, the set $U_\epsilon = \Theta \setminus V_\epsilon$ can be an arbitrarily tight neighborhood of the set Θ^*. Therefore by (3.17) we have the following result.

Theorem 3.1. *Suppose that Assumption* 3.1 *holds. Then with w.p.1 the distance from $\hat{\theta}_N$ to the set Θ^* converges in probability to zero. In particular if $\Theta^* = \{\theta^*\}$ is the singleton, then for almost every sequence $\{\xi_1, ...\}$, we have that θ_N converges in probability to θ^*.*

Remark 3.1. Convergence of Bayesian posterior distributions has been studied for a long time, dating back to Doob's consistency [6]. We refer the reader to [10] for a nice overview of Bayesian consistency results. Our analysis here resembles the proof and result of Schwartz consistency [24], but we do not require the assumption of the existence of a testing sequence, which is a common assumption in many of Bayesian consistency results (e.g., [24, 10, 14, 29]) but usually hard to verify in practice. Instead we impose simpler and maybe stronger assumptions (see Assumption 3.1). These assumptions are easy to verify and sufficient for our problems.

3.2 Consistency of Bayesian optimization problems

As in the previous section by writing w.p.1 we mean this with respect to the probability measure Q_*^∞. Consider a function $H : \mathcal{X} \times \Theta \to \mathbb{R}$ and the corresponding optimization problem

$$\min_{x\in\mathcal{X}} \left\{ \mathbb{E}_{\theta_N}[H_x] = \int_\Theta H(x,\theta)p(\theta|\boldsymbol{\xi}^{(N)})d\theta \right\}. \tag{3.18}$$

In this section we discuss convergence of the optimal value and the set of optimal solutions of the above problem as $N \to \infty$. In the considered applications the function $H(x,\theta)$ is given by

$$H(x,\theta) := \mathbb{E}_{\xi|\theta}[G(x,\xi)] \;\text{ and }\; H(x,\theta) := \sup_{Q\in\mathfrak{M}^\theta} \mathbb{E}_{Q|\theta}[G(x,\xi)] \tag{3.19}$$

in the cases of the risk-neutral Bayesian problem (2.1) and the Bayesian-DRO problem (2.2), respectively. Note that in both cases, the function $H(x,\theta)$ is convex in x if $G(x,\xi)$ is convex in x.

Let us discuss convergence of random variables $H_x(\theta_N) = H(x,\theta_N)$, $\theta_N \sim p(\cdot|\boldsymbol{\xi}^{(N)})$.

Lemma 3.2. *Suppose that Assumption 3.1 holds and $\Theta^* = \{\theta^*\}$ is the singleton. Then for any upper semi-continuous*[4] *function $h : \Theta \to \mathbb{R}$ it follows that*

$$\lim_{N\to\infty} \int_\Theta h(\theta)p(\theta|\boldsymbol{\xi}^{(N)})d\theta = h(\theta^*), \text{ w.p.1.} \tag{3.20}$$

[4]Recall that function $h(\theta)$ is said to be upper semi-continuous if $h(\theta) \geq \limsup_{\theta'\to\theta} h(\theta')$ for $\theta \in \Theta$. Of course, any continuous function is upper semi-continuous.

Proof. Let $\epsilon > 0$ and consider $\gamma_\epsilon := \sup_{\theta \in U_\epsilon} h(\theta) - h(\theta^*)$. By the definition (3.2) we have that $V_\epsilon \cup U_\epsilon = \Theta$. Note that since $\theta^* \in U_\epsilon$, we have that $\gamma_\epsilon \geq 0$. Note also that since function $h(\theta)$ is upper semi-continuous, it attains its maximum over $\theta \in \Theta$, and thus the constant

$$\lambda := \sup_{\theta \in \Theta} \{h(\theta) - h(\theta^*)\}$$

is finite (and non-negative). Then we can write

$$\begin{aligned}
\left| \int_\Theta h(\theta) p(\theta|\boldsymbol{\xi}^{(N)}) d\theta - h(\theta^*) \right| &= \left| \int_\Theta h(\theta) p(\theta|\boldsymbol{\xi}^{(N)}) d\theta - h(\theta^*) \int_\Theta p(\theta|\boldsymbol{\xi}^{(N)}) d\theta \right| \\
&= \left| \int_{U_\epsilon} \big(h(\theta) - h(\theta^*)\big) p(\theta|\boldsymbol{\xi}^{(N)}) d\theta + \int_{V_e} \big(h(\theta) - h(\theta^*) p(\theta|\boldsymbol{\xi}^{(N)}) d\theta \right| \\
&\leq \gamma_\epsilon \int_{U_\epsilon} p(\theta|\boldsymbol{\xi}^{(N)}) d\theta + \lambda \int_{V_\epsilon} p(\theta|\boldsymbol{\xi}^{(N)}) d\theta \\
&\leq \gamma_\epsilon + \lambda \int_{V_\epsilon} p(\theta|\boldsymbol{\xi}^{(N)}) d\theta.
\end{aligned}$$

By (3.3) the term $\int_{V_\epsilon} p(\theta|\boldsymbol{\xi}^{(N)}) d\theta$ can be arbitrarily small w.p.1 for N large enough. Since $h(\cdot)$ is upper semi-continuous and U_ϵ shrinks to $\{\theta^*\}$ as $\epsilon \downarrow 0$, we have that $\limsup_{\epsilon \downarrow 0} \gamma_\epsilon \leq 0$. Because $\gamma_\epsilon \geq 0$, it follows that γ_ϵ tends to zero as $\epsilon \downarrow 0$. Consequently the assertion (3.20) follows. □

In both settings of (3.19) it can be verified under standard regularity conditions that $H_x(\cdot)$ is upper semi-continuous on Θ. Indeed, in the risk-neutral case we have

$$\lim_{\theta' \to \theta} H_x(\theta') = \lim_{\theta' \to \theta} \int G_x(\xi) f(\xi|\theta') d\xi = \int \lim_{\theta' \to \theta} G_x(\xi) f(\xi|\theta') d\xi = H_x(\theta), \tag{3.21}$$

i.e., $H_x(\cdot)$ is continuous, provided that $f(\xi|\theta)$ is continuous in $\theta \in \Theta$ and the limit and integral can be interchanged (this can be ensured by the respective dominance condition). In the DRO setting of KL-divergence approach, we have that

$$H_x(\theta) = \inf_{\lambda > 0} \left\{ \lambda \epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}[e^{G_x/\lambda}] \right\}. \tag{3.22}$$

The above function is finite valued by assumption (2.12). Since infimum of a family of continuous functions is upper semi-continuous, it follows that the above $H_x(\cdot)$ is upper semi-continuous provided that $\mathbb{E}_{\xi|\theta}[e^{G_x/\lambda}]$ is continuous in θ.

For $x \in \mathcal{X}$ suppose that $H_x(\cdot)$ is upper semi-continuous on Θ. Then under the assumptions of Lemma 3.2 we have by (3.20) that

$$\lim_{N \to \infty} \mathbb{E}_{\theta_N}[H_x] = H(x, \theta^*), \text{ w.p.1.} \tag{3.23}$$

The above can be viewed as a point-wise LLN for random variables $H_x(\theta_N)$. Under mild additional assumptions this point-wise LLN can be extended (we will discuss this below) to the respective uniform LLN:

$$\lim_{N \to \infty} \sup_{x \in \mathcal{X}} \left| \mathbb{E}_{\theta_N}[H_x] - H(x, \theta^*) \right| = 0, \text{ w.p.1.} \tag{3.24}$$

Now consider the limiting optimization problem

$$\min_{x\in\mathcal{X}} H(x,\theta^*). \tag{3.25}$$

Denote by ϑ_N and ϑ^* the optimal value of the respective problems (3.18) and (3.25), and the corresponding sets

$$\mathcal{S}_N := \operatorname*{argmin}_{x\in\mathcal{X}} \mathbb{E}_{\theta_N}[H_x] \ \text{ and } \ \mathcal{S}^* := \operatorname*{argmin}_{x\in\mathcal{X}} H(x,\theta^*)$$

of optimal solutions. Suppose that the optimal value ϑ^* of problem (3.25) is finite. Then the uniform LLN (3.24) implies that (e.g., [26, Proposition 5.2])

$$\lim_{N\to\infty} \vartheta_N = \vartheta^* \text{ w.p.1.} \tag{3.26}$$

Under mild additional conditions, it is possible to show that the uniform LLN implies that[5]

$$\lim_{N\to\infty} \mathbb{D}(\mathcal{S}_N,\mathcal{S}^*) = 0, \text{ w.p.1} \tag{3.27}$$

(see, e.g., [26, Theorems 5.3 and 5.4]). This means that if x_N is an optimal solution of problem (3.18), then the distance from x_N to $\mathcal{S}^*$ tends to zero w.p.1. In particular, if $\mathcal{S}^* = \{x^*\}$ is the singleton, then x_N converges to x^* w.p.1.

Let us discuss now the uniform LLN (3.24). It is relatively easy to derive the uniform LLN in the following convex case.

Assumption 3.2. *Suppose that the set $\mathcal{X}$ is compact and there is a convex neighborhood[6] $\mathcal{V}$ of $\mathcal{X}$ such that function $H(\cdot,\theta)$ is finite valued convex on $\mathcal{V}$ for every $\theta\in\Theta$.*

Convexity of $H(\cdot,\theta)$ implies convexity of the expectation function $\int_\Theta H(\cdot,\theta)p(\theta|\boldsymbol{\xi}^{(N)})d\theta$. It is known by convex analysis that an extended real valued convex function is continuous on the interior of its domain. Moreover, if $f_k:\mathbb{R}^n\to\overline{\mathbb{R}}$ is a sequence of convex functions and $f:\mathbb{R}^n\to\overline{\mathbb{R}}$ is a convex function such that its domain has a nonempty interior, and $f_k(x)$ converges to $f(x)$ for all x in a dense subset of $\mathbb{R}^n$, then $f_k(\cdot)$ converges uniformly to $f(\cdot)$ on every compact subset of $\mathbb{R}^n$ which does not contain a boundary point of the domain of f (e.g., [22, Theorem 7.17]). By using this result it is not difficult to derive the following uniform LLN (e.g., [26, Theorem 7.50]).

Proposition 3.1. *Suppose that Assumption 3.2 is fulfilled and the point-wise LLN* (3.23) *holds for every $x\in\mathcal{V}$. Then the uniform LLN* (3.24) *follows.*

Without the convexity assumption we need to impose additional conditions. The following is similar to a derivation of the uniform LLN in the standard case (e.g., [26, Theorem 7.48]).

Theorem 3.2. *Suppose that Assumption 3.1 holds, the set $\Theta^*=\{\theta^*\}$ is the singleton, the set $\mathcal{X}$ is compact, and the function $H(x,\theta)$ is continuous on $\mathcal{X}\times\Theta$. Then the uniform LLN* (3.24) *follows.*

[5] By $\mathbb{D}(A,B)$ we denote the deviation of set $A\subset\mathbb{R}^n$ from set $B\subset\mathbb{R}^n$, that is $\mathbb{D}(A,B) := \sup_{x\in A}\operatorname{dist}(x,B)$, with $\operatorname{dist}(x,B) = \sup_{y\in B}\|x-y\|$.

[6] By the "neighborhood" we mean that the set $\mathcal{V}$ is open and $\mathcal{X}\subset\mathcal{V}$.

Proof. For a point $\bar{x} \in \mathcal{X}$, a sequence ν_k of positive numbers converging to zero and $\mathcal{V}_k := \{x \in \mathcal{X} : \|x - \bar{x}\| \leq \nu_k\}$, consider

$$\Delta_k(\theta) := \sup_{x \in \mathcal{V}_k} |H(x, \theta) - H(\bar{x}, \theta)|, \ \theta \in \Theta.$$

Since $H(x, \theta)$ is continuous on $\mathcal{X} \times \Theta$ and $\mathcal{X}$ is compact, it follows that $\Delta_k(\cdot)$ is continuous on Θ. Then by Lemma 3.2 we have that

$$\lim_{N \to \infty} \mathbb{E}_{\theta_N}[\Delta_k] = \Delta_k(\theta^*), \text{ w.p.1.} \tag{3.28}$$

By continuity of $H(\cdot, \theta^*)$, we have that $\Delta_k(\theta^*)$ tends to zero as $k \to \infty$. We also have by Lemma 3.2 that

$$\lim_{N \to \infty} \mathbb{E}_{\theta_N}[H_{\bar{x}}] = H(\bar{x}, \theta^*), \text{ w.p.1.} \tag{3.29}$$

Furthermore for $x \in \mathcal{V}_k$,

$$\begin{array}{rcl} \left|\mathbb{E}_{\theta_N}[H_x] - \mathbb{E}_{\theta_N}[H_{\bar{x}}]\right| & \leq & \left|\mathbb{E}_{\theta_N}[H_x] - H(\bar{x}, \theta^*)\right| + \left|\mathbb{E}_{\theta_N}[H_{\bar{x}}] - H(\bar{x}, \theta^*)\right| \\ & \leq & \mathbb{E}_{\theta_N}[\Delta_k] + \left|\mathbb{E}_{\theta_N}[H_{\bar{x}}] - H(\bar{x}, \theta^*)\right|. \end{array}$$

It follows that for a given $\epsilon > 0$ there is a neighborhood $\mathcal{W}$ of $\bar{x}$ such that w.p.1 for N large enough

$$\sup_{x \in \mathcal{X} \cap \mathcal{W}} \left|\mathbb{E}_{\theta_N}[H_x] - \mathbb{E}_{\theta_N}[H_{\bar{x}}]\right| \leq \epsilon. \tag{3.30}$$

The proof can be completed now exactly in the same way as in the proof of Theorem 7.48 in [26] by using compactness of the set $\mathcal{X}$. □

The assumed continuity of $H(x, \theta)$ on $\mathcal{X} \times \Theta$ can be verified under mild regularity conditions. That is, assume that $G(x, \xi)$ is continuous in $x \in \mathcal{X}$, $f(\xi|\theta)$ is continuous in $\theta \in \Theta$ and $G_x(\xi) f_\theta(\xi)$, $(x, \theta) \in \mathcal{X} \times \Theta$, is dominated by an integrable function. Then in the risk neutral case the continuity of $H(x, \theta)$ can be verified similar to (3.21). In the DRO setting, with $H(x, \theta)$ given in (3.22), the continuity of $H(x, \theta)$ also follows since the objective function in the right hand side minimization of problem (3.22) is strictly convex in $\lambda > 0$, and thus the corresponding minimizer is unique. By convexity of the objective function, this minimizer is a continuous function of $(x, \theta) \in \mathcal{X} \times \Theta$. Therefore, for (x, θ) in a neighborhood of a considered point the minimization can be restricted to a bounded (compact) subset of $\mathbb{R}_+$, and hence the continuity at the considered point follows.

3.3 Determination of the ambiguity set size

We consider how to determine the ambiguity set size ϵ in the Bayesian-DRO problem (2.14). Recall that Q_* denotes the true distribution of ξ with q_* denoting its pdf, and $\mu := \mathbb{E}_{\xi|\theta}[Z]$, $\sigma^2 := \mathrm{Var}_{\xi|\theta}(Z)$ for $Z : \Xi \to \mathbb{R}$. The true objective function can be written as

$$\begin{array}{rcl} \mathbb{E}_{Q_*}[Z] & = & \mu + \mathbb{E}_{\xi|\theta}\left[Z(\xi)\dfrac{q_*(\xi) - f(\xi|\theta)}{f(\xi|\theta)}\right] \\ & = & \mu + \mathbb{E}_{\xi|\theta}\left[(Z(\xi) - \mu)\dfrac{q_*(\xi) - f(\xi|\theta)}{f(\xi|\theta)}\right], \end{array}$$

where the second equality uses the fact $\mathbb{E}_{\xi|\theta}\left[\frac{q_*(\xi)-f(\xi|\theta)}{f(\xi|\theta)}\right] = 0$. Applying Cauchy-Schwartz inequality to the right hand side of the equation above, we have

$$\mathbb{E}_{Q_*}[Z] \leq \mu + \sigma \mathbb{E}_{\xi|\theta}\left[\left(\frac{q_*(\xi) - f(\xi|\theta)}{f(\xi|\theta)}\right)^2\right]^{1/2},$$

where the last term can be simplified as

$$\mathbb{E}_{\xi|\theta}\left[\left(\frac{q_*(\xi) - f(\xi|\theta)}{f(\xi|\theta)}\right)^2\right] = \mathbb{E}_{Q_*}\left[\frac{q_*(\xi)}{f(\xi|\theta)}\right] - 1.$$

If we let $2\epsilon = \mathbb{E}_{Q_*}\left[\frac{q_*(\xi)}{f(\xi|\theta)}\right] - 1$, then by (2.16) we have

$$\mathbb{E}_{Q_*}[Z] \leq \mu + \sigma\sqrt{2\epsilon} \approx \min_{\lambda>0}\left\{\lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}\left[e^{Z/\lambda}\right]\right\}, \tag{3.31}$$

which implies the objective value of the Bayesian-DRO problem (2.14) is an upper bound on the true objective value. Note here ϵ depends on θ.

A plausible idea of choosing the ambiguity set size is to make sure the ambiguity set contains the true distribution. That is, we would set

$$\epsilon(\theta) = D_{KL}(q_* \| f_\theta).$$

When q_* is close to f_θ, we can write $D_{KL}(q_*\|f_\theta) \approx \mathbb{E}_{Q_*}\left[\frac{q_*(\xi)}{f(\xi|\theta)}\right] - 1$. However, (3.31) shows even choosing ϵ half of the size, i.e. $\epsilon = \left(\mathbb{E}_{Q_*}\left[\frac{q_*(\xi)}{f(\xi|\theta)}\right] - 1\right)/2$, the Bayesian-DRO objective is still an upper bound on the true objective, which indicates this choice of ambiguity set size might be too conservative. Moreover, since q_* is unknown and has to be replaced by a continuous approximation of its empirical distribution, the number of samples required to achieve a certain approximation accuracy grows exponentially in dimension, which makes this method impractical in high dimension.

Now we consider a different method, which is inspired by [3]. We choose the ambiguity set to be the minimum KL ball containing at least one distribution under which the corresponding problem has the same optimal solution as the true problem. More specifically, we define a set of distributions as

$$\mathcal{Q}(x^*) := \{Q : x^* \in \operatorname*{argmin}_x \mathbb{E}_Q[G(x,\xi)]\},$$

where x^* is an optimal solution to the true problem. When $G(x,\xi)$ is convex in x and x^* is an interior point of $\mathcal{X}$, we can simplify by the first-order optimality condition,

$$\mathcal{Q}(x^*) = \{Q : \mathbb{E}_Q[\nabla_x G(x^*,\xi)] = 0\}.$$

In general, we can represent the condition in $\mathcal{Q}(x^*)$ by KKT conditions. Clearly, $Q_* \in \mathcal{Q}(x^*)$, i.e., the true distribution falls in the set $\mathcal{Q}(x^*)$. Now we set the ambiguity set size by minimizing the KL divergence from $\mathcal{Q}(x^*)$ to f_θ:

$$\hat{\epsilon}(\theta) = \min_{q\in\mathcal{Q}(x^*)} D_{KL}(q||f_\theta). \tag{3.32}$$

We do not know the optimal solution x^*, so in implementation we can replace x^* by the empirical optimal solution $\hat{x}_N$, which is the optimal solution to the SAA problem $\min_{x\in\mathcal{X}} \mathbb{E}_{\hat{Q}_N}[G(x,\xi)]$, where $\hat{Q}_N$ is the empirical distribution of the data $\boldsymbol{\xi}^{(N)}$. Since $\hat{x}_N - x^* = O_p(N^{-1/2})$ under certain regularity conditions, in particular if the true optimal x^* is unique (see Section 5.1 of [26]), and under mild conditions $\epsilon(\theta, x) = \min_{q\in\mathcal{Q}(x)} D_{KL}(q||f_\theta)$ is a smooth function in x, one can expect that $\epsilon(\theta, \hat{x}_N)$ is a good approximation of $\hat{\epsilon}(\theta)$.

4 Numerical Experiments

In this section, we demonstrate the performance of Bayesian-DRO on problems of one-dimension and multi-dimension with randomness having continuous and finite support respectively. The Bayesian-DRO problem (2.14) is restated as follows:

$$\min_{x\in\mathcal{X}} \mathbb{E}_{\theta_N}\left[\inf_{\lambda>0}\left\{\lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}\left[e^{G_x/\lambda}\right]\right\}\right], \tag{4.1}$$

where N is the number of data points, G_x stands for the cost function $G(x,\xi)$. In implementation, we apply SAA (e.g., [26]) to solve problem (4.1). We generate 100 samples of θ from the posterior distribution $p\left(\theta \mid \boldsymbol{\xi}^{(N)}\right)$ and 100 samples of ξ from the reference distribution $f(\xi|\theta)$ conditioned on each sampled θ. We compare the following approaches.

(1) Bayesian-DRO, with pre-specified ambiguity set size ϵ, which varies in a certain range.

(2) Bayesian-DRO, with ambiguity set size $\epsilon_1(\theta) = D_{KL}\left(q_* \| f(\cdot;\theta)\right)$, where the unknown true distribution q_* is estimated by the empirical distribution of the data.

(3) Bayesian-DRO, with ambiguity set size $\epsilon_2(\theta) = \frac{\epsilon_1(\theta)}{2}$. It halves ϵ_1 to reduce the overestimation, as shown in Section 3.3.

(4) Bayesian-DRO, with ambiguity set size $\epsilon_3(\theta)$, that is, solving problem (3.32) with x^* replaced by the empirical optimal solution to the SAA problem $\min_{x\in\mathcal{X}} \mathbb{E}_{\hat{Q}_N}[G(x,\xi)]$, where $\hat{Q}_N$ is the empirical distribution.

(5) Bayesian average, that is, solving the Bayesian average problem (2.1), which is the risk-neutral Bayesian average and is equivalent to letting $\epsilon = 0$ in Bayesian-DRO.

(6) Empirical approach, that is, solving the SAA problem $\min_{x\in\mathcal{X}} \mathbb{E}_{\hat{Q}_N}[G(x,\xi)]$.

(7) When the distribution of ξ has a finite support $\{\xi_1, \ldots, \xi_m\}$, we compare with Empirical-DRO (KL) in [12]. Specifically, we solve the following optimization problem:

$$\min_{x\in\mathcal{X}} \max_{Q} \mathbb{E}_Q[G(x,\xi)], \quad \text{s.t.} \quad \sum_{i=1}^{m} q_i \log\left(\frac{q_i}{\hat{p}_i}\right) \le \epsilon, \sum_{i:\hat{p}_i>0} q_i = 1, q_i \ge 0,$$

where $Q = [q_1, \cdots, q_m]$, $\hat{p}_i$ is the probability mass on ξ_i in the empirical distribution.

(8) We also compare with the DRO - Wasserstein. That is, we solve the following optimization problem:

$$\min_{x\in\mathcal{X}} \max_{Q} \mathbb{E}_Q[G(x,\xi)], \quad \text{s.t.} \quad W_p(Q,\hat{Q}_N) \le \tilde{\epsilon}, \tag{4.2}$$

where $W_p(Q,\hat{Q}_N)$ is the Wasserstein distance of order p between Q and the empirical distribution $\hat{Q}_N$, and $\tilde{\epsilon}$ is the ambiguity set size. The dual of (4.2) is given by [8, 3, 9]:

$$\min_{x\in\mathcal{X},\lambda\ge 0} \lambda\tilde{\epsilon}^p + \frac{1}{N}\sum_{i=1}^{N} \sup_{\xi\in\Xi}[G(x,\xi) - \lambda d(\xi,\hat{\xi}_i)^p],$$

where Ξ is the space of ξ, $d(\xi,\hat{\xi}_i)$ is the metric (or distance function) between two points ξ and $\hat{\xi}_i$, and $\{\hat{\xi}_i\}_{i=1}^N$ are the data points. In our experiments, we consider Wasserstein distance of order $p = 1, 2$, and the metric is chosen to be Euclidean norm. It is shown in [8] that, under mild assumptions, the distributionally robust optimization problems over Wasserstein balls can be reformulated as finite convex programs.

When the randomness has finite support, we choose the prior distribution in Bayesian-DRO and Bayesian average to be an uninformative Dirichlet distribution on θ. Sampling from a Dirichlet posterior distribution given the data is the same as Bayesian bootstrapping [15]. Please note that in this case, we implicitly choose the parameterized family to contain all discrete distributions on the support, which is the correct model. Numerical results for finite-support examples are shown in the Online Appendix.

When the distribution of ξ is continuous, we compute the ambiguity set sizes in Bayesian-DRO with the following implementation details.

- In Bayesian-DRO with ambiguity set size $\epsilon_1(\theta)$ and $\epsilon_2(\theta)$, the KL divergence from the empirical distribution to the reference distribution is estimated using the estimation method in [19]. Specifically, we compute the empirical cumulative distribution function (cdf) given the data, construct linear interpolation of the empirical cdf, and then we use the finite difference method to compute the estimated KL divergence as:

$$\widehat{D}_{\mathrm{KL}}(Q\|f(\cdot;\theta)) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\delta P_c\left(\hat{\xi}_i\right)}{\Delta f\left(\hat{\xi}_i;\theta\right)}\right),$$

where $\left\{\hat{\xi}_i\right\}_{i=1}^N$ are the data points, P_c is the linear interpolation of the empirical cdf, $\delta P_c\left(\hat{\xi}_i\right) = P_c\left(\hat{\xi}_i\right) - P_c\left(\hat{\xi}_i - \Delta\right)$, $\Delta < \min_i\left\{\hat{\xi}_i - \hat{\xi}_{i-1}\right\}$.

- In Bayesian-DRO with ambiguity set size $\epsilon_3(\theta)$, to compute the minimum KL ball, we conduct Monte Carlo sampling from $f(\xi|\theta)$. Essentially, we employ SAA to solve the

problem

$$\min_q \frac{1}{L}\sum_{i=1}^{L}\log(\frac{q(\xi_i)}{f(\xi_i|\theta)})\frac{q(\xi_i)}{f(\xi_i|\theta)}$$
$$\text{s.t. } \frac{1}{L}\sum_{i=1}^{L}\frac{q(\xi_i)}{f(\xi_i|\theta)}=1, \quad \frac{1}{L}\sum_{i=1}^{L}\nabla_x G(x^*,\xi_i)\frac{q(\xi_i)}{f(\xi_i|\theta)}=0, \quad q(\xi_i)\geq 0,$$

where $\xi_1,\cdots,\xi_L$ are $L=100$ samples drawn from $f(\xi|\theta)$. We solve this optimization problem using Gurobi 9.1 with Python 3.7 API and scipy package in Python. Algorithmic description of this approach can be found in Algorithm 1 in the appendix.

We evaluate the performance of each algorithm following the procedure in [12], as follows. All algorithms are run for $K=200$ replications. In each replication $j=1,\cdots,K$, we collect N data points $\hat{\xi}_1,\cdots,\hat{\xi}_N$ drawn i.i.d. from the true distribution $\mathbb{P}_{\theta^c}$. Then we run each algorithm with the same data set and obtain its optimal solution, denoted by $x^{(j)}(\epsilon)$, where ϵ is the corresponding ambiguity set size. We then compute $\mu^{(j)}(\epsilon)=\mathbb{E}_{\mathbb{P}_{\theta^c}}[G(x^{(j)}(\epsilon),\xi)]$ and $v^{(j)}(\epsilon)=\mathrm{Var}_{\mathbb{P}_{\theta^c}}[G(x^{(j)}(\epsilon),\xi)]$, i.e., the (mean and variance) performance of the obtained solutions under the true system. The out-of-sample mean and variance are then approximated using these $K=200$ replications, with $\hat{\mu}_N(\epsilon)=\frac{1}{K}\sum_{j=1}^{K}\mu^{(j)}(\epsilon)$ and $\hat{v}_N(\epsilon)=\frac{1}{K}\sum_{j=1}^{K}v^{(j)}(\epsilon)+\frac{1}{K-1}\sum_{j=1}^{K}(\mu^{(j)}(\epsilon)-\hat{\mu}_N(\epsilon))^2$.

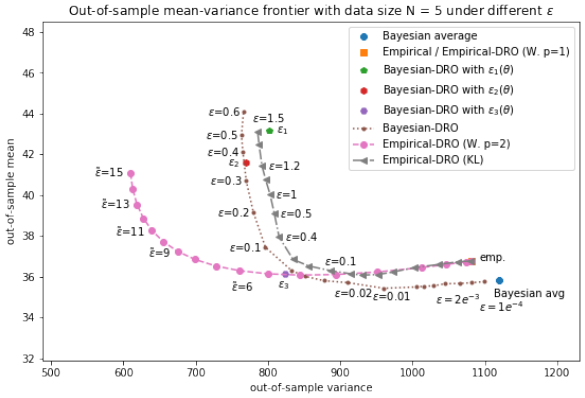
4.1 One-dimensional Newsvendor with Continuous Randomness

In this subsection, we run experiments on a one-dimensional newsvendor problem when the randomness ξ has a continuous distribution and the data all come from the true distribution (see [21] for a review on newsvendor models). We summarize notations used in the classical newsvendor problem as follows.

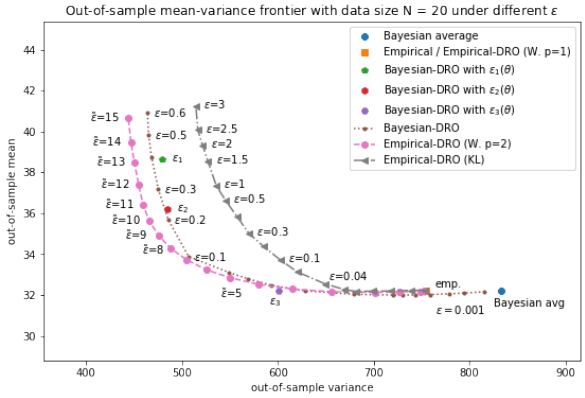
- x: order amount, assumed to be in $[0,M]$, M is the maximal order amount.
- ξ: random customer demand.
- b: backorder cost per unit.
- h: holding cost per unit.
- c: ordering cost per unit.

The cost function is given by $G(x,\xi)=h(x-\xi)^+ + b(\xi-x)^+ + cx$, where $(\cdot)^+=\max(\cdot,0)$. We assume the customer demand $\xi\in\Xi$, where $\Xi=(0,\infty)$. Parameters used in the newsvendor problem are summarized as follows: maximal ordering amount $M=50$, backorder cost $b=8$, holding cost $h=3$, ordering cost $c=0$.

In the first experiment, we test the performance of our proposed algorithms under model mis-specification. Specifically, the true distribution of the customer demand is normal distribution with mean 10 and variance 100 truncated above 0. In Bayesian-DRO, we choose the parametric family $f(\xi|\theta)$ to be the exponential distribution with rate parameter θ. To have closed-form posterior update, we use the conjugate prior of gamma distribution with parameter



(a) $N = 5$.



(b) $N = 20$.

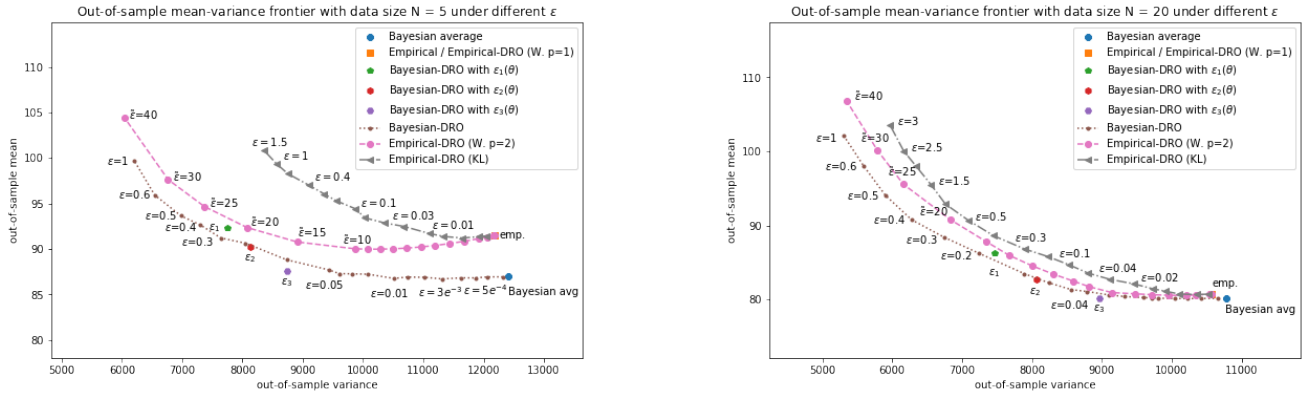
Figure 1: Newsvendor with continuous support: out-of-sample mean-variance frontiers of different algorithms under different ϵ values. Data size N is 5 and 20 respectively. Bayesian-DRO has model mis-specification.

N=5	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.58(0.04)	0.29(0.02)	0.07(0.01)	-	-	-
solution	26.13(0.80)	24.04(0.65)	18.15(0.38)	15.46(0.31)	16.44(0.38)	17.41
mean	43.14(0.33)	41.62(0.71)	36.16(0.53)	35.81(0.31)	36.77(0.25)	30.96
variance	802.24(2.39)	769.72(2.11)	823.97(2.56)	1119.30(2.73)	1082.21(2.93)	640.59
N=20	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.34(0.02)	0.17(0.01)	0.03(0.00)	-	-	-
solution	24.36(0.38)	22.13(0.30)	18.30(0.17)	16.16(0.15)	16.97(0.17)	17.41
mean	38.62(0.54)	36.19(0.60)	33.22(0.07)	32.22(0.06)	32.20(0.07)	30.96
variance	478.74(1.62)	485.15(1.49)	601.01(1.95)	832.66(2.38)	754.11(2.22)	640.59

Table 1: Newsvendor with continuous support: out-of-sample performance of variants of Bayesian-DRO with model mis-specification. Data size N is 5 and 20 respectively.

$(1, 1)$. Please note this choice of prior distribution is only for computational convenience. If the Bayesian updating does not admit closed-form posterior, we may use Monte Carlo simulation, such as Markov Chain Monte Carlo (MCMC) methods, to draw samples from the posterior; we only need sample average approximation of the expectations when solving the Bayesian-DRO problem. Figure 1 shows the out-of-sample mean-variance frontiers (with varying ϵ values) of different algorithms for data size $N = 5$ and 20. For the empirical approach (abbreviated as empirical), Bayesian average approach (abbreviated as Bayesian average), and Bayesian-DRO with calibrated ambiguity set size $\epsilon_1(\theta), \epsilon_2(\theta), \epsilon_3(\theta)$, their performance is denoted by one point (not a frontier) in the figure. Note that for Empirical-DRO with Wasserstein distance (abbreviated as W. in the figure) of order $p = 1$, it is equivalent to the empirical approach (see Remark 6.7 in [8] and Theorem 3.2 in [16]) and is independent of the ambiguity set size. Table 1 shows the out-of-sample performance of variants of Bayesian-DRO when data size is $N = 5$ and 20 respectively; solving the true problem (abbreviated as true) is included as a benchmark for all compared algorithms; standard errors of the average ϵ values, obtained solutions, and the out-of-sample

performances are shown within the parentheses in the table.



(a) $N = 5$. (b) $N = 20$.

Figure 2: Newsvendor with continuous randomness: out-of-sample mean-variance frontiers of different algorithms under different ϵ values. Data size N is 5 and 20 respectively. Bayesian-DRO chooses the correct model.

N=5	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.35(0.04)	0.17(0.02)	0.09(0.00)	-	-	-
solution	31.91(1.03)	28.73(0.96)	27.64(0.83)	23.30(0.73)	25.63(1.01)	25.99
mean	92.29(0.90)	90.25(0.85)	87.56(0.68)	86.99(0.72)	91.50(1.16)	77.66
variance	7743.27(44.87)	8134.65(44.00)	8743.18(37.55)	12409.66(37.35)	12184.37(42.12)	9760.90
N=20	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.18(0.01)	0.09(0.01)	0.03(0.00)	-	-	-
solution	33.11(0.87)	30.95(0.84)	28.32(0.46)	24.71(0.39)	25.31(0.44)	25.99
mean	86.21(0.88)	82.65(0.83)	80.18(0.29)	80.15(0.22)	80.70(0.25)	77.66
variance	7470.54(34.99)	8067.69(36.30)	8968.67(19.86)	10773.79(19.98)	10570.40(22.20)	9760.90

Table 2: Newsvendor with continuous randomness: out-of-sample performance of variants of Bayesian-DRO without model mis-specification. Data size N is 5 and 20 respectively.

In the second experiment, we test the performance of our proposed algorithms without model mis-specification. Specifically, the true distribution of the customer demand is exponential distribution with mean 20. We choose the parametric family $f(\xi|\theta)$ to be the correct model, i.e., the exponential distribution with rate parameter θ. Figure 2 shows the out-of-sample mean-variance frontiers (with varying ϵ values) of different algorithms for data size $N = 5$ and 20. Table 2 shows the out-of-sample performance of variants of Bayesian-DRO when data size is $N = 5$ and 20 respectively.

We have the following observations from the two experiments above.

(1) **Trade-off between out-of-sample mean and variance**: both Bayesian-DRO and Empirical-DRO show the trade-off. As the ambiguity set size ϵ grows larger, the out-of-sample mean deteriorates, which trades for more robustness in terms of smaller out-of-sample variance. Empirical approach is equivalent to Empirical-DRO with $\epsilon = 0$, and

Bayesian average is equivalent to Bayesian-DRO with $\epsilon = 0$. Therefore, empirical approach and Bayesian average produce solutions with larger out-of-sample variance and smaller out-of-sample mean compared to Empirical-DRO and Bayesian-DRO respectively.

(2) **Model mis-specification affects the performance of Bayesian-DRO**: when there is model mis-specification, Bayesian-DRO underperforms Empirical-DRO with Wasserstein distance of order $p = 2$, as can be seen from the worse mean-variance frontier in Figure 1. If we choose the correct model, Bayesian-DRO outperforms Empirical-DRO with Wasserstein distance of order $p = 2$, as can be seen from Figure 2. This is expected, since a poorly chosen model, which serves as the reference distribution of the ambiguity set in Bayesian-DRO, deteriorates the performance of Bayesian-DRO. However, the ambiguity set in Bayesian-DRO still provides robustness against model mis-specification, as it can be seen from Figure 1 that Bayesian-DRO (with ϵ_3) has about the same out-of-sample mean but much smaller variance than Bayesian average (which is equivalent to $\epsilon = 0$ in Baysian-DRO).

(3) **Bayesian-DRO outperforms Empirical-DRO with KL divergence**: in almost all the experiments, the mean-variance frontier of Bayesian-DRO dominates that of Empirical-DRO (KL). The reason is because the ambiguity sets of Bayesian-DRO contain distributions supported on the domain of the randomness if the prior distribution is chosen to cover the domain, whereas the Empirical-DRO with KL divergence only allows probability distributions in the ambiguity set that are absolutely continuous with respect to the empirical distribution (i.e., the observed data points) and leaves out distributions supported on the unobserved domain.

(4) **Parameter-dependent ambiguity set size outperforms pre-specified ones**: for Bayesian-DRO with parameter-dependent ambiguity set size $\epsilon_2(\theta), \epsilon_3(\theta)$, the out-of-sample performances are better compared to Bayesian-DRO with pre-specified ambiguity set size (i.e., fixed ϵ for all θ). It shows we can gain better performance for Bayesian-DRO by tuning an appropriate parameter-dependent ambiguity set size, although this incurs more computational cost.

(5) **Large data size reduces model uncertainty**: as expected, solutions of all the methods become more stabilized (smaller variance) as data size increases. In particular, solution of the empirical approach gets closer to the true optimal solution with more data.

4.2 Multi-dimensional Newsvendor with Continuous Randomness

In this subsection, we consider a three-dimensional newsvendor problem with multi-items, where the newsvendor sells three kinds of items (see [27] for a review on newsvendor models). Assume the customer demands for each kind of item are independent and follow normal distribution with mean 10, 12, 15 and standard deviation 20, 20, 20 respectively, truncated above 0. The objective function is given by: $G(x, \xi) = \sum_{i=1}^{3} h_i (x_i - \xi_i)^+ + b_i (\xi_i - x_i)^+$. We set $h_i = 3, b_i = 8$ for $i = 1, 2, 3$.

The parametric distribution we choose is the exponential distribution with rate parameter θ_i for each customer demand for item i. Figure 3 shows the out-of-sample mean-variance frontiers

N=10	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	1.05(0.03)	0.53(0.02)	0.17(0.01)	-	-	-
sol error	20.95(0.50)	18.84(0.43)	11.67(0.32)	10.33(0.28)	12.37(0.39)	0.00
mean	254.72(2.09)	238.56(1.83)	198.21(0.78)	184.13(0.76)	190.28(0.95)	171.28
variance	4585.39(35.09)	4759.10(28.18)	5845.42(16.55)	10030.76(20.25)	8613.13(20.35)	7066.05

Table 3: Multi-dimensional newsvendor with continuous randomness: out-of-sample performance of variants of Bayesian-DRO that has model mis-specification. Data size N is 10.

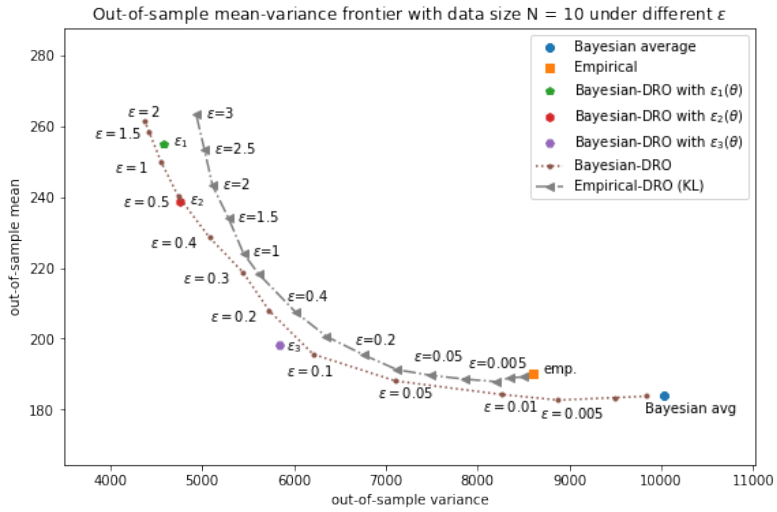


Figure 3: Multi-dimensional newsvendor with continuous randomness: out-of-sample mean-variance frontiers of different algorithms under different ϵ values. Data size N is 10. Bayesian-DRO has model mis-specification.

(with varying ϵ values) of different algorithms when data size $N = 10$. Table 3 shows the out-of-sample performance of variants of Bayesian-DRO when data size $N = 10$; in addition to out-of-sample performance, we also show the solution error, which is obtained by calculating each solution's Euclidean distance from the true optimal solution; standard errors of the average ϵ values, obtained solution error, and the out-of-sample performances are shown within the parentheses in the table. Similar to the one-dimensional newsvendor problem, **Bayesian-DRO outperforms Empirical-DRO (KL)** in the multi-dimensional newsvendor problem.

5 Conclusions and Future Work

We propose a new formulation, Bayesian Distributionally Robust Optimization (Bayesian-DRO), to address the ambiguity about the probability distribution in static stochastic optimization. Bayesian-DRO takes advantage of Bayesian estimation of parametric distributions and at the same time imposes robustness against the uncertainty introduced by the assumed parametric model. When the ambiguity set is constructed using Kullback-Leibler divergence and the size of the set is small, the robustness of Bayesian-DRO can be interpreted as a trade-off between the posterior mean and standard deviation of the cost function. We show the strong consistency of

Bayesian posterior distributions, and subsequently show the convergence of objectives and optimal solutions of Bayesian-DRO problems. Moreover, we consider several methods of determining the ambiguity set size in Bayesian-DRO. Our numerical results demonstrate that when data are limited, Bayesian-DRO has superior out-of-sample performance compared to KL-based empirical DRO, the Bayesian-average approach, and the empirical approach; Bayesian-DRO outperforms Wasserstein-based empirical DRO when the parametric family is correctly chosen (i.e., no model mis-specification) but underperforms when there is model mis-specification. More future research is needed to fully understand the connections between these frameworks (Bayesian-DRO, empirical-DRO, BRO) and how to choose a framework for specific data-driven stochastic optimization problems.

The nature of sequential Bayesian updating makes Bayesian approaches especially amenable to multi-stage (dynamic) settings where data come sequentially in time. One of the future works is to extend Bayesian-DRO to multi-stage stochastic optimization, including multi-stage stochastic programming, stochastic control, and Markov decision processes.

Acknowledgment

All authors are grateful for the support by Air Force Office of Scientific Research (AFOSR) under Grant FA9550-22-1-0244. The second and third authors are also grateful for the support by AFOSR under Grant FA9550-19-1-0283 and National Science Foundation (NSF) under Grant DMS2053489.

References

[1] Güzin Bayraksan and David K. Love. Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, pages 1–19. INFORMS, 2015.

[2] Aharon Ben-Tal and Marc Teboulle. Penalty functions and duality in stochastic programming via phi-divergence functionals. *Mathematics of Operations Research*, 12:224–240, 1987.

[3] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56:830–857, 2019.

[4] Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964.

[5] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[6] Joseph L. Doob. Application of the theory of martingales. *Actes du Colloque International Le Calcul des Probabilites et ses applications*, pages 23–27, 1948.

[7] John C. Duchi, Peter W. Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021.

[8] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

[9] Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[10] Subhashis Ghosal, Jayanta Kumar Ghosh, and Aad van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28:500–531, 2000.

[11] Jun-ya Gotoh, Michael J. Kim, and Andrew E. B. Lim. Robust empirical optimization is almost the same as mean-variance optimization. *Operations Research Letters*, 46(4):448–452, 2018.

[12] Jun-ya Gotoh, Michael J. Kim, and Andrew E. B. Lim. Calibration of distributionally robust empirical optimization models. *Operations Research*, 69(5):1349–1650,, 2021.

[13] Vishal Gupta. Near-optimal bayesian ambiguity sets for distributionally robust optimization. *Management Science*, 65:4242–4260, 2019.

[14] Bastiaan Kleijn and Aad van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, 34:837–877, 2006.

[15] Henry Lam and Enlu Zhou. Quantifying uncertainty in sample average approximation. In *Proceedings of the 2015 Winter Simulation Conference*, 2015.

[16] Sangyoon Lee, Hyunwoo Kim, and Ilkyeong Moon. A data-driven distributionally robust newsvendor model with a Wasserstein ambiguity set. *Journal of the Operational Research Society*, 72(8):1879–1897, 2021.

[17] Heng Lian. On rates of convergence for posterior distributions under misspecification. *Communications in Statistics—Theory and Methods*, 38:11:1893–1900, 2009.

[18] Tetsuzo Morimoto. Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, 1963.

[19] Fernando Pérez-Cruz. Kullback-leibler divergence estimation of continuous distributions. In *2008 IEEE international symposium on information theory*, pages 1666–1670. IEEE, 2008.

[20] Alois Pichler and Alexander Shapiro. Mathematical foundations of distributionally robust multistage optimization. *https://arxiv.org/abs/2101.02498*, 2021.

[21] Evan L Porteus. Stochastic inventory theory. *Handbooks in Operations Research and Management Science*, 2:605–652, 1990.

[22] R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[23] Herbert Scarf, KJ Arrow, and S Karlin. A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production*, 10:201–209, 1958.

[24] Lorraine Schwartz. On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4:10–26, 1965.

[25] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017.

[26] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[27] Nazli Turken, Yinliang Tan, Asoo J Vakharia, Lan Wang, Ruoxuan Wang, and Arda Yenipazarli. The multi-product newsvendor problem: Review, extensions, and directions for future research. *Handbook of Newsvendor Problems*, pages 3–39, 2012.

[28] Di Wu, Helin Zhu, and Enlu Zhou. A Bayesian risk approach to data-driven stochastic optimization: Formulations and asymptotics. *SIAM Journal on Optimization*, 28(2):1588–1612, 2018.

[29] Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *Annals of Statistics*, 48(4):2180 – 2207, 2020.

A Supplementary Numerical Experiments

A.1 Algorithm 1: Bayesian-DRO with ambiguity set size ϵ_3

Algorithm 1: Bayesian-DRO with ambiguity set size ϵ_3.

input : data points of size N, number of θ samples N_θ, number of ξ samples N_ξ, number of Monte Carlo samples to compute the ambiguity set size L

output: optimal solution $x(\epsilon_3)$

Solve for the SAA solution x_N^*;

for $i = 1 \leftarrow 1$ **to** N_θ **do**

Simulate θ_i from posterior distribution $p(\theta|\boldsymbol{\xi}^{(N)})$;

Simulate $\{\xi_j\}_{j=1}^L$ from reference distribution $f(\xi|\theta_i)$, solve the optimization problem

$$\epsilon_3(\theta_i) = \min_q \frac{1}{L}\sum_{j=1}^{L} \log(\frac{q(\xi_j)}{f(\xi_i|\theta_i)})\frac{q(\xi_i)}{f(\xi_j|\theta)}$$

$$\text{s.t. } \frac{1}{L}\sum_{j=1}^{L}\frac{q(\xi_j)}{f(\xi_j|\theta_i)} = 1, \quad \frac{1}{L}\sum_{j=1}^{L}\nabla_x G(x_N^*, \xi_j)\frac{q(\xi_j)}{f(\xi_j|\theta_i)} = 0, \quad q(\xi_j) \geq 0;$$

Simulate $\{\hat{\xi}_j\}_{j=1}^{N_\xi}$ from reference distribution $f(\xi|\theta_i)$ and store them as dataset $\mathcal{D}_i$;

end

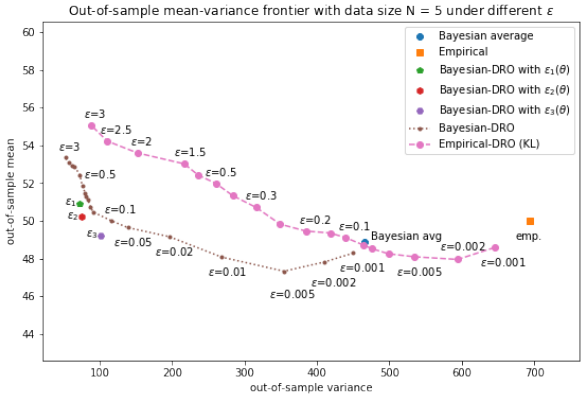
Solve the Bayesian-DRO problem and obtain the optimal solution $x(\epsilon_3)$

$$\min_{x\in\mathcal{X},\lambda_i>0}\left\{\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\left(\lambda_i\epsilon_3\left(\theta_i\right)+\lambda_i\log\left(\frac{1}{N_\xi}\sum_{\hat{\xi}\in\mathcal{D}_i}\exp\left(G(x,\hat{\xi})/\lambda_i\right)\right)\right)\right\}.$$

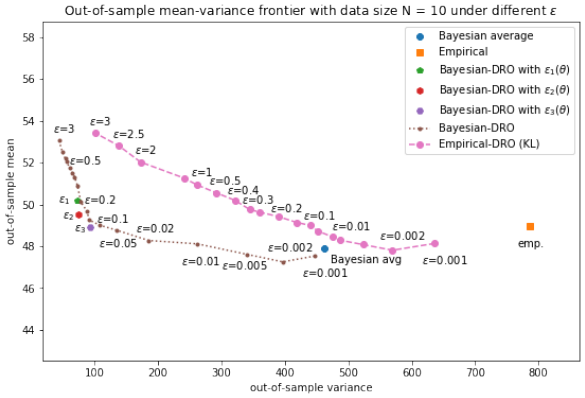
A.2 One-dimensional Newsvendor with Finite-support Randomness

In this subsection, we first run experiments on a one-dimensional newsvendor problem when the randomness ξ has a finite support and the data all come from the true distribution. Different from the continuous-support case, the random customer demand is assumed to take discrete values in $\{1, 2, \cdots, 14, 15\}$. The true probability mass $\theta^c \in \Delta_{15}$ is unknown to the decision maker, where Δ_{15} stands for a probability simplex. Parameters used in the newsvendor problem are summarized as follows. Maximal ordering amount $M = 20$, backorder cost $b = 10$, holding cost $h = 2$, ordering cost $c = 3$.

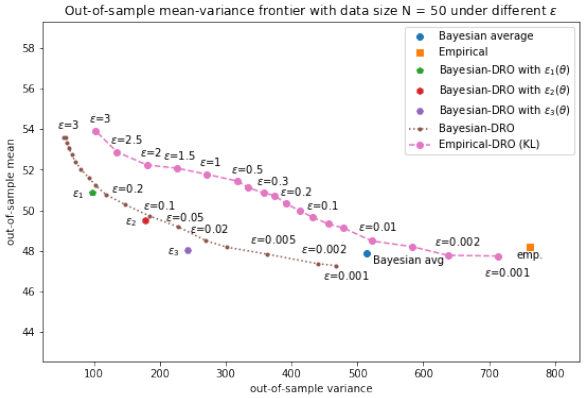
Figure 4 shows the out-of-sample mean-variance frontiers (with varying ϵ values) of different algorithms for data sizes $N = 5$, 10, 50 and 1000. Table 4 shows the out-of-sample performance of each algorithm when data size is $N = 5, 10, 50, 1000$ respectively. Similar to the continuous-support case, **Bayesian-DRO performs better than Empirical-DRO in most cases**, as the mean-variance frontier of Bayesian-DRO dominates that of empirical-DRO (KL). Note that for a small data size, Empirical-DRO (KL) will only put non-negative probability mass on the support point $\hat{\xi}$ that has been observed in the data. On the other hand, by imposing an appropriate prior



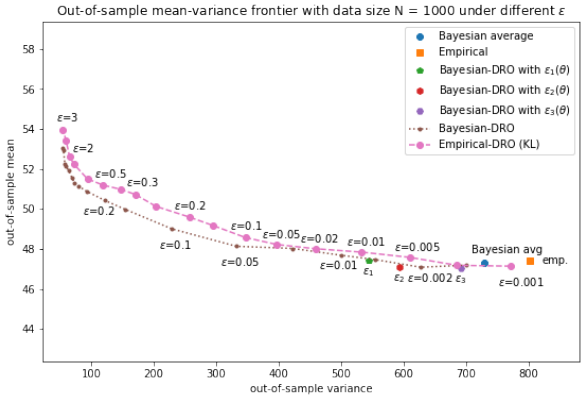
(a) $N = 5$.



(b) $N = 10$.



(c) $N = 50$.



(d) $N = 1000$.

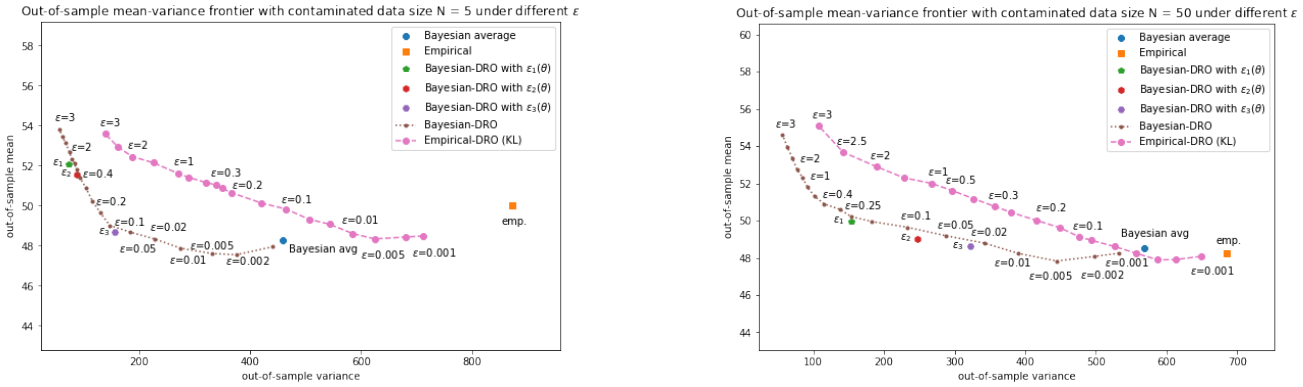
Figure 4: Newsvendor with finite-support randomness: out-of-sample mean-variance frontiers of different algorithms under different ϵ values. Data size varies from $5, 10, 50$ to 1000.

(in this problem we use a non-informative Dirichlet prior whose domain is a uniform distribution on the support of ξ), Bayesian-DRO can put non-negative probability mass on all the support points. Also note that the mean-variance frontiers of Bayesian-DRO and Empirical-DRO get closer as the data size N goes to infinity due to the reduced model uncertainty.

Next, we consider a contaminated data model, where 80% data are generated from the true distribution and 20% data are generated from an arbitrary distribution. In particular, the arbitrary distribution is randomly generated (specified by its probability mass) and is different in each replication. Figure 5 shows the out-of-sample mean-variance frontiers (with varying ϵ values) of different algorithms for data size $N = 5$ and 50. Table 5 shows the out-of-sample performance of all variants of Bayesian-DRO when data size is $N = 5$ and 50 respectively. Similar to the non-contaminated case, **Bayesian-DRO outperforms other benchmarks** even when data are contaminated. Note that the solution of the empirical approach does not get closer to the true optimal solution with more data, since part of the data are not from the true distribution and possibly become outliers.

N=5	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.99(0.02)	0.50(0.01)	0.14(0.01)	-	-	-
solution	12.77(0.17)	12.70(0.17)	11.29(0.14)	8.59(0.06)	8.55(0.24)	7.00
mean	50.92(0.43)	50.22(0.50)	49.24(0.47)	48.90(0.02)	50.01(0.28)	47.21
variance	72.27(2.15)	75.12(2.28)	101.24(2.38)	465.08(2.91)	693.63(3.63)	770.56
N=10	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.60(0.01)	0.30(0.01)	0.07(0.01)	-	-	-
solution	12.28(0.16)	12.20(0.14)	10.61(0.10)	8.57(0.09)	7.50(0.20)	7.00
mean	50.20(0.37)	49.54(0.28)	48.93(0.30)	47.87(0.05)	48.94(0.15)	47.21
variance	71.99(2.07)	74.97(2.09)	91.97(2.14)	461.88(2.55)	786.58(3.56)	770.56
N=50	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.14(0.00)	0.07(0.00)	0.02(0.00)	-	-	-
solution	11.24(0.04)	10.91(0.05)	9.26(0.06)	7.99(0.10)	7.36(0.13)	7.00
mean	50.44(0.05)	49.51(0.06)	48.22(0.04)	47.87(0.04)	48.21(0.05)	47.21
variance	118.65(1.99)	178.06(1.88)	241.68(1.89)	513.35(2.08)	761.45(2.86)	770.56
N=1000	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.006(0.00)	0.003(0.00)	0.001(0.00)	-	-	-
solution	8.08(0.01)	7.82(0.01)	7.25(0.03)	7.18(0.05)	6.93(0.06)	7.00
mean	47.42(0.03)	47.12(0.03)	47.03(0.02)	47.32(0.01)	47.39(0.02)	47.21
variance	544.42(0.92)	593.85(1.29)	691.00(1.15)	728.15(1.47)	801.66(1.50)	770.56

Table 4: Newsvendor with finite-support randomness: out-of-sample performance of variants of Bayesian-DRO. Data size N varies from 5, 10, 50 to 1000.



(a) $N = 5$. (b) $N = 50$.

Figure 5: Newsvendor with finite support: out-of-sample mean-variance frontiers of different algorithms under different ϵ values with contaminated data. Data size N is 5 and 50 respectively.

A.3 Multi-dimensional Portfolio Optimization with Finite-support Randomness

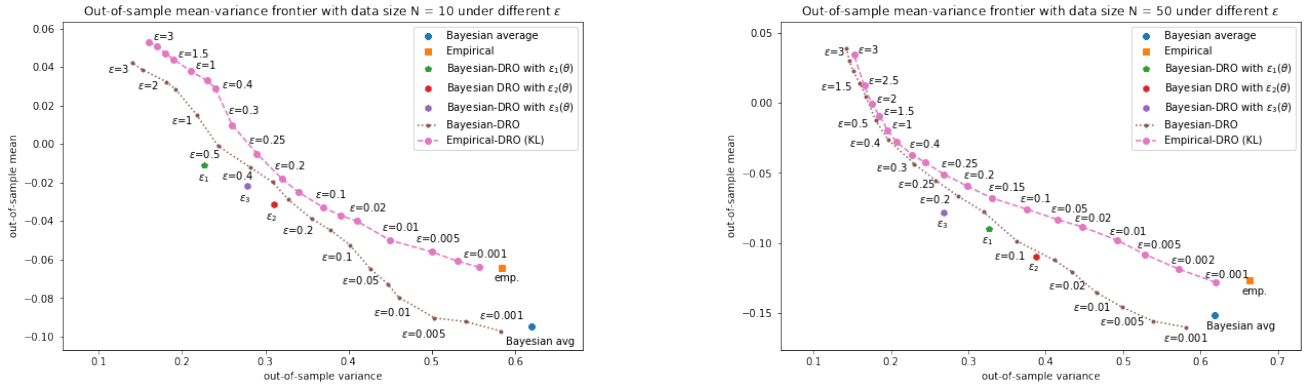
In this subsection, we consider a five-dimensional portfolio optimization problem when the randomness ξ has finite support and the data all come from the true distribution. We summarize notations used in the portfolio optimization problem as follows.

N=5	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	1.00(0.02)	0.50(0.01)	0.07(0.01)	-	-	-
solution	12.02(0.10)	11.72(0.13)	10.01(0.16)	8.59(0.08)	7.41(0.24)	7.00
mean	52.09(0.16)	51.58(0.27)	48.68(0.26)	48.29(0.05)	50.04(0.21)	47.21
variance	73.12(1.97)	87.09(1.98)	155.98(2.27)	459.39(2.47)	872.79(4.43)	770.56
N=50	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.13(0.00)	0.06(0.00)	0.02(0.00)	-	-	-
solution	11.32(0.05)	10.72(0.06)	9.18(0.05)	7.86(0.11)	7.46(0.12)	7.00
mean	49.97(0.06)	49.04(0.07)	48.63(0.11)	48.53(0.03)	48.26(0.04)	47.21
variance	153.08(1.41)	247.25(1.53)	321.62(1.63)	568.38(2.58)	684.59(2.69)	770.56

Table 5: Newsvendor with finite support: out-of-sample performance of variants of Bayesian-DRO algorithm with contaminated data. Data size N is 5 and 50 respectively.

- x: holding positions of assets. $x \in [0,1]^5$, $\sum_{i=1}^5 x_i = 1$.
- ξ: random returns of assets. ξ_i takes values in $\{-1, 0, 1\}$ for $i = 1, \cdots, 5$.

The cost function is given by $G(x, \xi) = -\xi^\top x$. Note that we do not allow shorting (i.e., $x_i > 0, i = 1, \cdots, 5$) and impose a budget constraint ($\sum_{i=1}^5 x_i = 1$). The true probability mass of dimension i, denoted by $\theta_i^c \in \Delta_3$, is unknown to the decision maker.



(a) $N = 10$. (b) $N = 50$.

Figure 6: Portfolio optimization with finite support: out-of-sample mean-variance frontiers of different algorithms under different ϵ values. Data size N is 10 and 50 respectively.

Figure 6 shows the out-of-sample mean-variance frontiers (with varying ϵ values) of different algorithms for data size $N = 5$ and 50. Table 6 shows the out-of-sample performance of variants of Bayesian-DRO when data size is $N = 10$ and 50 respectively. In addition to out-of-sample performance, we also show the solution error, which is obtained by calculating each solution's Euclidean distance from the true optimal solution, with sample standard deviation within the parentheses in the tables. Similar to the one-dimensional newsvendor problem, **Bayesian-DRO outperforms other benchmarks** in the multi-dimensional portfolio optimization problem.

N=10	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.49(0.01)	0.25(0.01)	0.47(0.02)	-	-	-
sol error	0.73(0.01)	0.70(0.02)	0.79(0.01)	0.76(0.05)	0.80(0.04)	0.00
mean	-0.01(0.00)	-0.03(0.00)	-0.02(0.00)	-0.09(0.01)	-0.06(0.01)	-0.17
variance	0.23(0.01)	0.31(0.01)	0.28(0.00)	0.62(0.01)	0.58(0.02)	0.59
N=50	ϵ_1	ϵ_2	ϵ_3	Bayesian avg	empirical	true
ϵ value	0.10(0.00)	0.05(0.00)	0.18(0.01)	-	-	-
sol error	0.53(0.02)	0.48(0.02)	0.61(0.01)	0.33(0.04)	0.40(0.04)	0.00
mean	-0.09(0.00)	-0.11(0.00)	-0.08(0.00)	-0.15(0.00)	-0.13(0.00)	-0.17
variance	0.33(0.01)	0.39(0.01)	0.27(0.00)	0.62(0.01)	0.66(0.01)	0.59

Table 6: Portfolio optimization with finite support: out-of-sample performance of variants of Bayesian-DRO. Data size N is 10 and 50 respectively.