Bayesian Distributionally Robust Optimization

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

December 14, 2021

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 strong exponential consistency of the Bayesian posterior distribution and subsequently the convergence of objective functions and optimal solutions of Bayesian-DRO. We also consider several approaches to selecting the ambiguity set size in Bayesian-DRO and compare them numerically. Our numerical results demonstrate the out-of-sample performance of Bayesian-DRO on the news vendor problem of different dimensions and data types.

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 objective 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.

[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 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 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 [18]. 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 [12].

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.

Recently, [22] 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.

In this paper, we propose a new formulation termed Bayesian Distributionally Robust Optimization (Bayesian-DRO), which poses robustness against the model uncertainty 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.

We show the strong (exponential) 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 divergence (within the parametric family) from the true distribution. Built on this result, we show the objective functions and optimal solutions of Bayesian-DRO are strongly consistent. When the ambiguity set is small, Bayesian-DRO is approximately equivalent to a weighted sum of the mean and standard deviation under the parametric model, 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 mean and variability of the solution performance. We further propose several theoretical and empirical methods to determine the ambiguity set size, and compare their out-of-sample performance numerically.

The rest of the paper is organized 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 on a news vendor problem. 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. 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 $p(\theta|\boldsymbol{\xi}^{(N)})$ defined in (1.4) 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. We also note that (2.1) is equivalent to the BRO formulation (1.5) with expectation being the risk functional.

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. 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.

Consider the following distributionally robust functional

$$\mathfrak{R}(Z) := \mathbb{E}_{\theta_N}\left[\sup_{Q \in \mathfrak{M}^\theta} \mathbb{E}_{Q|\theta}[Z]\right], \tag{2.2}$$

associated with problem (1.6). 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 [16] for a detail discussion of such conditional functionals. For random variable $Z : \Xi \to \mathbb{R}$ the respective expectation in (2.2) is

$$\mathbb{E}_{Q|\theta}[Z] = \int_\Xi Z(\xi) q(\xi|\theta) d\xi, \tag{2.3}$$

where $q(\cdot|\theta)$ is the pdf of $Q \in \mathfrak{M}^\theta$. The maximum (supremum) in the right hand side of (2.2) is taken over all pdfs $q_\theta(\xi) = q(\xi|\theta)$ from the ambiguity set $\mathfrak{M}^\theta$.

The distributionally robust counterparts of problem (2.1) is obtained by employing the above distributionally robust functional. That is, problem (1.6) can be written as

$$\min_{x \in \mathcal{X}} \mathfrak{R}(G_x), \tag{2.4}$$

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.

2.1 Construction of the ambiguity set

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

$$\varrho_{|\theta}(\cdot) := \sup_{Q \in \mathfrak{M}^\theta} \mathbb{E}_{Q|\theta}[\,\cdot\,], \tag{2.5}$$

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 (2.3)) 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],[14]), is general and flexible. Let $\phi : \mathbb{R} \to \mathbb{R}_+ \cup \{+\infty\}$ be a convex lower semicontinuous 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

$$\mathfrak{M}_\epsilon^\theta := \left\{ q_\theta : \int_\Xi \phi\big(q_\theta(\xi)/f_\theta(\xi)\big) f_\theta(\xi) d\xi \leq \epsilon \right\}. \tag{2.6}$$

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 (2.6) is assumed to be absolutely continuous with respect to the reference measure f_θ for every $\theta \in \Theta$.

It can be shown by duality arguments (cf., [1],[2],[20]), that for a random variable $Z : \Xi \to \mathbb{R}$,

$$\varrho_{|\theta}(Z) = \inf_{\lambda \geq 0, \mu} \left\{ \lambda\epsilon + \mu + \mathbb{E}_{\xi|\theta}\big[(\lambda\phi)^*(Z - \mu)\big] \right\}, \tag{2.7}$$

where $\phi^*(y) = \sup_{x \geq 0}\{yx - \phi(x)\}$ is the conjugate of ϕ (recall that the expectation $\mathbb{E}_{\xi|\theta}$ is taken with respect to the reference distribution of ξ). Note that $(\lambda\phi)^*(y) = \lambda\phi^*(y/\lambda)$ for $\lambda > 0$. Hence, the functional (2.2) can be written as

$$\mathfrak{R}(Z) = \mathbb{E}_\theta\Big[\underbrace{\inf_{\lambda > 0, \mu} \mathbb{E}_{\xi|\theta}\big[\lambda\epsilon + \mu + \lambda\phi^*\big((Z - \mu)/\lambda\big)\big]}_{\varrho_{|\theta}(Z)} \Big]. \tag{2.8}$$

The measurability of the infimum in the right hand side of (2.8), considered as a function of θ, can be verified under mild regularity conditions. We can refer to [17, Chapter 14] for a thorough discussion of this measurability issue.

2.1.1 Kullback-Leibler divergence

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

$$D_{KL}(q\|f) := \int_\Xi q(\xi) \ln\big(q(\xi)/f(\xi)\big) d\xi = \int_\Xi (q(\xi)/f(\xi)) \ln\big(q(\xi)/f(\xi)\big) f(\xi) d\xi. \tag{2.9}$$

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

$$\mathbb{E}_{\xi|\theta}\big[e^{tZ}\big] < +\infty \quad \text{for any } t \in \mathbb{R} \text{ and } \theta \in \Theta. \tag{2.10}$$

For the KL-divergence, given $\lambda > 0$ the minimizer over μ in (2.7) is given by $\mu = \lambda \ln \mathbb{E}_{\xi|\theta}\big[e^{Z/\lambda}\big]$, and hence the minimum becomes

$$\varrho_{|\theta}(Z) = \inf_{\lambda>0} \left\{ \lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}\big[e^{Z/\lambda}\big] \right\}. \tag{2.11}$$

Consequently, the DRO problem (2.2) can be written as

$$\min_{x\in\mathcal{X}} \mathbb{E}_{\theta_N} \left[\inf_{\lambda>0} \left\{ \lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}[e^{G_x/\lambda}] \right\} \right]. \tag{2.12}$$

The above optimization problem (2.12) 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 in (2.11) is attained as $\lambda \to +\infty$, and is equal $\mathbb{E}_{\xi|\theta}[Z]$. For $\epsilon > 0$ the optimization problem (2.11) 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 (2.11) for $\theta \in \Theta$ and small $\epsilon > 0$. By condition (2.10), 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 generation function we can write

$$\lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}\big[e^{Z/\lambda}\big] = \lambda\epsilon + \mu + \tfrac{1}{2}\sigma^2/\lambda + O(\lambda^{-2}), \tag{2.13}$$

where[2] $\mu := \mathbb{E}_{\xi|\theta}[Z]$, $\sigma^2 := \mathrm{Var}_{\xi|\theta}(Z)$ By minimizing the right hand side of (2.13) we obtain approximation $\bar\lambda \approx \frac{\sigma}{\sqrt{2\epsilon}}$ of the optimal solution of (2.11), and consequently for small $\epsilon > 0$ the approximation

$$\min_{\lambda>0} \left\{ \lambda\epsilon + \lambda \ln \mathbb{E}_{\xi|\theta}\big[e^{Z/\lambda}\big] \right\} \approx \mu + \sigma\sqrt{2\epsilon}. \tag{2.14}$$

The approximation (2.14) reveals that when the ambiguity set is small, the Bayesian-DRO formulation is approximately equal to a weighted sum of the mean and standard deviation of the performance function under the parametric model, with weight depending on the ambiguity set size ϵ. Similar interpretation has also been observed for empirical DRO, see [7],[10]. Moreover, [11] shows that the empirical DRO can be interpreted as a trade-off between the mean and worst-case sensitivity; whether such an interpretation can be extended to Bayesian-DRO will be left as a future work.

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

2.1.3 Variants of Bayesian-DRO Formulations

Consider the problem

$$\min_{x\in\mathcal{X},\lambda>0} \mathbb{E}_{\theta_N}\left[\lambda\epsilon+\lambda\ln\mathbb{E}_{\xi|\theta}[\exp(G_x/\lambda)]\right]. \tag{2.15}$$

The nested Bayesian-DRO problem (2.12) can be viewed as a relaxation of problem (2.15). In (2.15) the parameter λ is chosen before observing a realization of θ, while in (2.12) the parameter λ is a function of θ. We have that the optimal value of the Bayesian-DRO problem (2.12) is less than or equal to the optimal value of problem (2.15).

In the above derivations we considered the ambiguity with respect to the reference pdf $f(\cdot|\theta)$, and consequently the corresponding Bayesian-DRO problem (1.6). 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 $D_{KL}\left(\mathfrak{p}\|p(\cdot|\boldsymbol{\xi}^{(N)})\right)\le\epsilon$. Let

$$\mathcal{R}(Y):=\sup_{\mathfrak{p}\in\mathcal{M}_\epsilon}\left\{\mathbb{E}_{\mathfrak{p}}[Y]=\int_\Theta Y(\theta)\mathfrak{p}(\theta)d\theta\right\}, \tag{2.16}$$

be the corresponding distributionally robust functional defined on a space of random variables $Y:\Theta\to\mathbb{R}$. Similar to (2.11) we have the following representation of that functional

$$\mathcal{R}(Y)=\inf_{\lambda>0}\left\{\lambda\epsilon+\lambda\ln\mathbb{E}_{\theta_N}\left[e^{Y/\lambda}\right]\right\}. \tag{2.17}$$

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 (2.17) we can write this optimization problem as

$$\min_{x\in\mathcal{X},\lambda>0} \lambda\epsilon+\lambda\ln\mathbb{E}_{\theta_N}\big[\exp\left(\mathbb{E}_{\xi|\theta}[G_x/\lambda]\right)\big]. \tag{2.18}$$

Now by interchanging the expectation $\mathbb{E}_{\theta_N}$ and the supremum in the definition (2.2) of the distributionally robust functional $\mathfrak{R}$, we can consider the functional

$$\Re(Z):=\sup_{Q\in\mathfrak{M}^\theta}\mathbb{E}_{\theta_N}\left[\mathbb{E}_{Q|\theta}[Z]\right], \tag{2.19}$$

and the corresponding Bayesian-DRO problem. We have that $\Re(\cdot)\le\mathfrak{R}(\cdot)$ and the inequality can be strict since the extreme measure Q in (2.2) could depend on θ. The maximization in (2.19) 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 $\xi_1,\dots,\xi_N$ 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. We make the following assumptions.

Assumption 3.1. *Suppose the following holds.* (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

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

Let

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

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 $\theta \in \Theta$

and exponentially fast 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 $N \geq N'$.

Theorem 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

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

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

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

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

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

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

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., [21, 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. □

Remark 3.1. 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 (exponentially fast) as $N \to \infty$.

Note that for an appropriate $\epsilon > 0$, the set $U_\epsilon = \Theta \backslash V_\epsilon$ can be an arbitrarily tight neighborhood of the set Θ^*. Therefore (3.17) implies that 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 $\theta_N \xrightarrow{p} \theta^*$.

Remark 3.2. Convergence of Bayesian posterior distributions has been studied for a long time, dating back to Doob's consistency [6]. We refer the reader to [9] for a nice overview of Bayesian consistency results. Our analysis here resembles the proof and result of Schwartz consistency [19], 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., [19, 9, 13, 23]) 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

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

in the cases of the risk-neutral Bayesian problem (2.1) and the Bayesian-DRO problem (1.6), 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 3.1. *Suppose that Assumption 3.1 holds and $\Theta^* = \{\theta^*\}$ is the singleton. Then for any upper semicontinuous*[4] *function $h : \Theta \to \mathbb{R}$ it follows that*

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

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

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

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

is finite (and nonnegative). Then we can write

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

By (3.3) the term $\int_{V_\epsilon} p(\theta|\boldsymbol{\xi}^{(N)})d\theta$ can be arbitrary small w.p.1 for N large enough. Since $h(\cdot)$ is upper semicontinuous 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 semicontinuous 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 (2.10). Since infimum of a family of continuous functions is upper semicontinuous, it follows that the above $H_x(\cdot)$ is upper semicontinuous 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 semicontinuous on Θ. Then under the assumptions of Lemma 3.1 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 pointwise LLN for random variables $H_x(\theta_N)$. Under mild additional assumptions this pointwise 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., [21, 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., [21, 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., [17, Theorem 7.17]). By using this result it is not difficult to derive the following uniform LLN (e.g., [21, Theorem 7.50]).

Proposition 3.1. *Suppose that Assumption 3.2 is fulfilled and the pointwise 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., [21, 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.*

Proof. For a point $\bar{x} \in \mathcal{X}$, a sequence ν_k of positive numbers converging to zero and $\mathcal{V}_k := \{x \in \mathcal{X} : \|x - \bar{x}\| \le \nu_k\}$, 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 3.1 we have that

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

[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 $\mathcal{X} \subset \mathcal{V}$.

By continuity of $H(\cdot,\theta^*)$, we have that $\Delta_k(\theta^*)$ tends to zero as $k\to\infty$. We also have by Lemma 3.1 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 $x\in\mathcal{V}_k$,

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

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

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

The proof can be completed now exactly in the same way as in the proof of Theorem 7.48 in [21] 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 Determine the ambiguity set size

We consider how to determine the ambiguity set size ϵ in the Bayesian-DRO problem (2.12). Recall that Q_* denotes the true distribution of ξ with q_* denoting its pdf, $\mu:=\mathbb{E}_{\xi|\theta}[Z]$, $\sigma^2:=\mathrm{Var}_{\xi|\theta}(Z)$. The true objective function can be written as

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

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

$$\mathbb{E}_{Q_*}[Z]\le\mu+\sigma\mathbb{E}_{\xi|\theta}\left[\left(\frac{q_0(\xi)-f(\xi|\theta)}{f(\xi|\theta)}\right)^2\right]^{1/2},$$

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 (2.14) we have

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

which implies the objective value of the Bayesian-DRO problem (2.12) 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 the set

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

Hence, $\mathcal{Q}(x)$ is the set of distributions under which x satisfies the first-order optimality condition of the corresponding optimization problem. Let x^* denote the optimal solution to the true problem. Clearly, $Q_*\in\mathcal{Q}(x^*)$. Now we set the ambiguity set size by projecting f_θ onto the set $\mathcal{Q}(x^*)$, i.e., minimizing the KL divergence from f_θ to $\mathcal{Q}(x^*)$:

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

Let $\hat{q}^*$ denote the minimizer of (3.32), and hence $\mathbb{E}_{\hat{q}^*}[\nabla_x G(x^*,\xi)]=0$. 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 $\hat{x}^*\approx x^*+Z/\sqrt{N}$ (cf. [21]) where Z is a standard normal random vector, the approximation error only depends on the data size N and is independent of the dimension of ξ.

4 Numerical Experiments

In this section, we demonstrate the performance of Bayesian-DRO on problems of one-dimension and multi-dimension with non-contaminated and contaminated data respectively. We also compare different methods of determining the ambiguity set size in Bayesian-DRO.

We first consider the classical newsvendor problem of one-dimension, with notations summarized as follows.

- x: ordering amount, assumed to be in $[0, M]$
- ξ: random customer demand
- b: backorder cost per unit, assumed to be a constant
- h: holding cost per unit, assumed to be a constant

The cost function is given by:

$$G(x, \xi) = b(\xi - x)^+ + h(x - \xi)^+,$$

where $(\cdot)^+ = \max(\cdot, 0)$. We set $M = 100, b = 3, h = 4$. Note that the true optimal solution to the original problem (under the true distribution) is given by $x^* = F^{-1}(\frac{b}{h+b})$, where F^{-1} is the inverse cdf for the (true) probability distribution. We assume the true distribution of customer demand is normal distribution with mean 20 and standard deviation 4.

We consider the Bayesian-DRO problem (2.12), which is restated as follows:

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

In implementation, we apply SAA to solve problem (4.1). We generate 100 samples of θ from $p(\theta|\boldsymbol{\xi}^{(N)})$ and 100 samples of ξ from $f(\xi|\theta)$ conditioned on each sample θ. We choose the parametric family $f(\xi|\theta)$ to be the exponential distribution with mean parameter θ. To have closed-form posterior update, we use the conjugate prior of gamma distribution with parameter $(1, 1)$.

We compare the following methods of determining the ambiguity set size ϵ in (4.1). The first three methods have been discussed in Section 3.4, and the last method of cross-validation is a commonly-used empirical method.

1. $\epsilon_1(\theta) = D_{KL}(q_* \| f(\cdot; \theta))$, where the unknown true distribution q_* is replaced by the empirical distribution of the data.

2. $\epsilon_2(\theta) = \frac{1}{2}\epsilon_1(\theta)$. It halves ϵ_1 to reduce the over-estimation, as shown in Section 3.4.

3. $\epsilon_3(\theta)$ is chosen to be the minimum Kullback-Leibler ball which contains a distribution under which the corresponding problem has the same optimal solution as the true problem. That it, solve problem (3.32) with x^* replaced by the empirical optimal solution to the SAA problem $\min_{x \in \mathcal{X}} \mathbb{E}_{\hat{Q}_n}[G(x, \xi)]$.

4. ϵ_4 is chosen by K-folds cross-validation. In particular, we have a predetermined ϵ set. For each ϵ value in the set, we split the data into $K = 5$ groups. For each group, we take the group as the test set, and take the remaining groups as the training set. We solve problem (4.1) with the given ϵ using the training set, evaluate the solution on the test set to obtain the score, which is computed as the summation of sample mean and a coefficient weighted sample standard deviation. We take average of all five scores for each ϵ in the set, and set ϵ_4 to be the one with the smallest average score.

To estimate the KL divergence $\epsilon_1(\theta)$ from the empirical distribution to the reference distribution, we use the estimation method in [15]. Specifically, we compute the empirical 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:

$$\widehat{D}_{\text{KL}}(Q||f(\cdot;\theta)) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\delta P_c(\xi_i)}{\Delta f(\xi_i;\theta)}\right),$$

where n is the number of data, $\{\xi_i\}_{i=1}^n$ are the data points, P_c is the linear interpolation of the empirical cdf, $\delta P_c(\xi_i) = P_c(\xi_i) - P_c(\xi_i - \Delta)$, $\Delta < \min_i\{\xi_i - \xi_{i-1}\}$.

To compute the minimum KL ball $\epsilon_3(\theta)$, we discretize the support in the reference distribution by the discrete set $\mathcal{S} = \{0, 0.1, 0.2, \cdots, 99.9\}$. We replace integration by summation, and minimize the objective function $\sum_{i=1}^{1000} \frac{1}{10} f(s_i|\theta)\log(\frac{f(s_i|\theta)}{q(s_i)})$ with constraints $q(s_i) \geq 0$, $\sum_{i=1}^{1000} \frac{1}{10} q(s_i) = 1$, and $h\sum_{i=1}^{\lceil 10\hat{x}^*\rceil} q(s_i) = b\sum_{i=\lceil 10\hat{x}^*\rceil}^{1000} q(s_i)$. We solve this optimization problem with Gurobi 9.1.

Given the same set data, we obtain the optimal solutions using the algorithms listed above and compare their out-of-sample performance using 10000 independent data points. For comparison, we also consider two other approaches. The first is directly solving the Bayesian average problem (2.1), which is the risk-neutral Bayesian average and is equivalent to letting $\epsilon = 0$ in Bayesian-DRO. The second is the empirical approach, where we solve $\min_{x\in\mathcal{X}} \mathbb{E}_{\hat{Q}_n}[G(x,\xi)]$ with $\hat{Q}_n$ being the empirical distribution. All experiments are run via Jupyter Notebook on MacBook Pro with 1.4 GHz Quad-Core Intel Core i5 processor and 8 GB memory.

4.1 Non-contaminated Data

We first run experiments with non-contaminated data, where the data all come from the true distribution. Figure 1 shows how the out-of-sample mean and out-of-sample standard deviation of Bayesian-DRO solutions vary as ϵ varies. It shows the importance of choosing an appropriate ϵ value. Table 1 to Table 4 show the out-of-sample performance of each approach when data size is $m = 5, 10, 50, 100$ respectively, where variants of the Bayesian-DRO approach with different ϵ values are abbreviated as ϵ_1, ϵ_2, ϵ_3 and ϵ_4 respectively. We compare to the benchmark Bayesian average (abbreviated as $\epsilon = 0$), empirical approach (abbreviated as empirical), and the true problem. In addition to out-of-sample performance, we also show the average ϵ values (with sample standard deviation within the parentheses) and the obtained solutions in the tables. In the cross-validation approach, we set the coefficient to be 1. To better understand the impact of the coefficient on the cross-validation approach, we plot the ϵ_4 values under different coefficients in Figure 2, which shows that ϵ_4 stays relatively constant (ranging from 1.2 to 1.4) when we vary the coefficient from 0 to 2 in the cross-validation approach.

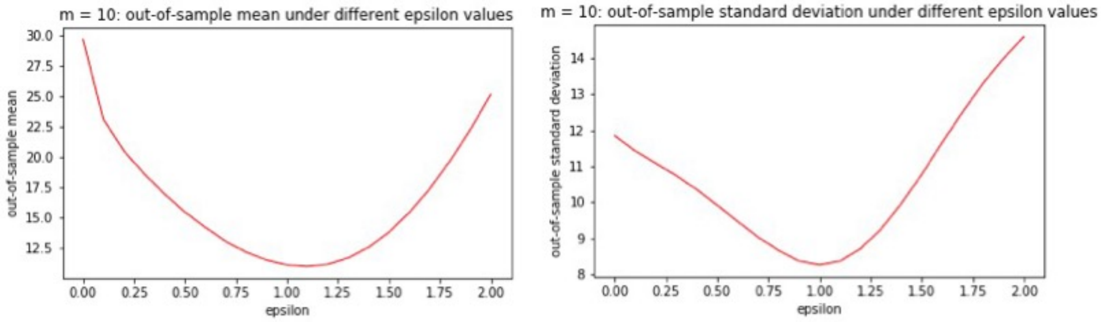

Figure 1: Out-of-sample mean and standard deviation of Bayesian-DRO under different ϵ values. Data size $m = 10$.

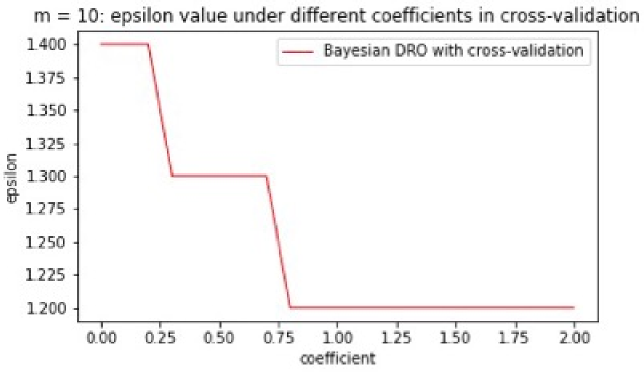

Figure 2: ϵ value under different coefficients in the Bayesian-DRO with cross-validation approach.

m=5	ϵ_1	ϵ_2	ϵ_3	ϵ_4	$\epsilon = 0$	empirical	true
(avg) ϵ value	2.09(0.17)	1.04(0.08)	1.24(0.86)	2.30	-	-	-
solution	26.56	18.29	19.65	28.50	10.15	22.49	19.28
out-of-sample mean	26.70	11.32	10.99	34.03	29.74	14.41	10.95
out-of-sample std	14.85	8.31	8.47	15.58	11.85	11.15	8.33

Table 1: Out-of-sample performance of Bayesian-DRO with non-contaminated data and size 5.

When the data size is small (e.g., $m = 5$ in Table 1), the cross-validation approach performs the worst since the validation set size is small, and the large ambiguity set size ϵ_4 is caused by the variability in the validation data set. From Table 2 to Table 4, we find that the first approach (ϵ_1) among the Bayesian-DRO variants performs the worst, due to the large ambiguity set size. As shown in Section 3.4 the ambiguity set size ϵ_1 is overestimated. When the data size goes to infinity, ϵ_1 converges to the true KL divergence from the true distribution (normal) to the reference distribution (exponential) plus a bias term 0.5772 [15]. Note that when the data size is

m=10	ϵ_1	ϵ_2	ϵ_3	ϵ_4	$\epsilon = 0$	empirical	true
(avg) ϵ value	1.56(0.11)	0.78(0.05)	1.01(0.55)	1.20	-	-	-
solution	22.58	17.28	18.76	20.07	10.16	17.89	19.28
out-of-sample mean	14.61	12.32	11.06	11.13	29.68	11.64	10.95
out-of-sample std	11.26	8.74	8.26	8.71	11.84	8.44	8.33

Table 2: Out-of-sample performance of Bayesian-DRO with non-contaminated data and size 10.

m=50	ϵ_1	ϵ_2	ϵ_3	ϵ_4	$\epsilon = 0$	empirical	true
(avg) ϵ value	1.97(0.02)	0.99(0.01)	0.60(0.22)	1.40	-	-	-
solution	28.83	21.03	18.05	24.20	10.68	18.71	19.28
out-of-sample mean	35.29	11.96	11.50	18.84	28.17	11.08	10.95
out-of-sample std	15.63	9.51	8.37	13.03	11.79	8.25	8.33

Table 3: Out-of-sample performance of Bayesian-DRO with non-contaminated data and size 50.

m=100	ϵ_1	ϵ_2	ϵ_3	ϵ_4	$\epsilon = 0$	empirical	true
(avg) ϵ value	2.01(0.01)	1.01(0.00)	0.65(0.24)	1.30	-	-	-
solution	30.06	21.88	19.02	24.20	10.94	19.23	19.28
out-of-sample mean	40.17	13.24	10.98	18.85	27.39	10.96	10.95
out-of-sample std	15.80	10.44	8.27	13.03	11.76	8.34	8.33

Table 4: Out-of-sample performance of Bayesian-DRO with non-contaminated data and size 100.

small, the bias term cannot be exactly computed, so we do not correct ϵ_1 in the computation. We also find that the third approach (ϵ_3) performs the best among all four approaches, since the out-of-sample mean (note that we consider a cost minimization problem) and out-of-sample standard deviation are both the smallest among the four approaches. Also note that the Bayesian average approach ($\epsilon = 0$) does not perform well, because it does not impose any robustness against model mis-specification (here we choose the exponential distribution as the parametric model, whereas the true distribution is normal). Finally, as expected, with more data the empirical approach yields solution that gets closer to the true optimal solution, and performs better compared to the Bayesian-DRO approach when there is abundant data.

4.2 Contaminated Data

Now we consider a contaminated data model, where we assume 80% data are generated from the true distribution, and 20% data are generated from an arbitrary distribution. In this example, the arbitrary distribution is chosen to be exponential distribution with mean 10. Table 5 and Table 6 show the out-of-sample performance of each approach with data size $m = 50$ and $m = 1000$ respectively. As shown by the tables, when the data are contaminated, the empirical approach does not generate the best solution even when the data size is large (e.g., $m = 1000$), since part of the data are not from the true distribution and possibly become outliers, which degrades the performance of the empirical approach. The third approach (ϵ_3) still performs the best among all four variants of the Bayesian-DRO approach.

m=50	ϵ_1	ϵ_2	ϵ_3	ϵ_4	$\epsilon = 0$	empirical	true
(avg) ϵ value	1.54(0.02)	0.77(0.01)	1.13(0.58)	1.10	-	-	-
solution	22.24	17.02	19.45	19.22	9.45	17.27	19.28
out-of-sample mean	13.90	12.67	10.96	10.96	31.80	12.33	10.95
out-of-sample std	10.86	8.89	8.39	8.32	11.91	8.34	8.33

Table 5: Out-of-sample performance of Bayesian-DRO with contaminated data and size 50.

m=1000	ϵ_1	ϵ_2	ϵ_3	ϵ_4	$\epsilon = 0$	empirical	true
(avg) ϵ value	1.37(0.05)	0.69(0.00)	0.89(0.20)	1.00	-	-	-
solution	22.47	17.49	18.94	19.73	9.77	18.45	19.28
out-of-sample mean	14.40	12.07	11.02	11.03	30.80	11.25	10.95
out-of-sample std	11.14	8.61	8.26	8.51	11.90	8.27	8.33

Table 6: Out-of-sample performance of Bayesian-DRO with contaminated data and size 1000.

4.3 Three-dimensional Example

We consider a three-dimensional news vendor problem with multi-items, where the newsboy sells three kinds of items. Assume the customer demands for each kind of item are independent and follow normal distribution with mean $20, 30, 40$ and standard deviation $4, 5, 6$, respectively. 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 $h_i = 4, b_i = 3$ for $i = 1, 2, 3$. The parametric distribution we choose is the exponential distribution with mean parameter θ_i for each customer demand for item i. To have closed-form posterior update, we use the conjugate prior of gamma distribution with parameter $(1, 1)$ for each customer demand. The other procedures and parameter settings are the same as the one-dimensional problem above.

We first consider non-contaminated data, where data all come from the true distribution. Table 7 and Table 8 show the out-of-sample performance of each approach with data size $m = 10, 1000$ respectively.

Then we consider the contaminated data, where we assume 80% data are generated from the true distribution, and 20% data are generated from exponential distribution with mean 5. Table 9 and Table 10 show the out-of-sample performance of each algorithm with data size $m = 50$ and $m = 1000$ respectively. For both non-contaminated and contaminated cases in the three-dimensional problem, same observations can be made as in the one-dimensional problem.

4.4 Summary of numerical experiments

Finally, we can draw the following conclusions from the numerical experiments conducted above:

- When the model is mis-specified (i.e., the true distribution does not fall in the assumed parametric family), Bayesian-DRO with appropriate ambiguity set size outperforms the risk-neutral Bayesian average approach due to its robustness against model mis-specification.

m=10	ϵ_1	ϵ_2	ϵ_3	ϵ_4	$\epsilon = 0$	empirical	true

(avg) ϵ value	1.56(0.11)
	1.81(0.08)
	3.29(0.07)
		0.78(0.05)
		0.91(0.04)
		1.64(0.03)
			0.97(0.57)
			1.07(0.57)
			1.04(0.43)
				1.40
				1.90
				1.70
					-	-	-

solution	23.11 35.89 75.33	17.43 26.91 47.43	18.74 28.45 38.75	21.89 36.87 48.36	10.51 15.15 21.08	17.89 28.39 39.26	19.28 29.10 38.92
out-of-sample mean	182.98	59.48	41.40	65.59	129.89	42.04	41.19
out-of-sample std	31.87	25.07	18.17	23.09	26.33	18.40	18.28

Table 7: Out-of-sample performance of Bayesian-DRO with non-contaminated data and size 10 in the three-dimensional problem.

m=100	ϵ_1	ϵ_2	ϵ_3	ϵ_4	$\epsilon = 0$	empirical	true

(avg) ϵ value	2.01(0.00)
	1.93(0.00)
	2.02(0.00)
		1.01(0.00)
		0.96(0.00)
		1.01(0.00)
			0.69(0.18)
			0.69(0.19)
			0.63(0.34)
				1.30
				1.30
				1.30
					-	-	-

solution	30.72 45.52 59.63	21.99 32.98 43.73	19.47 29.54 37.64	24.44 37.18 48.21	10.87 17.11 22.12	19.23 30.16 38.62	19.28 29.10 38.92
out-of-sample mean	183.69	53.35	41.77	46.00	119.74	41.66	41.19
out-of-sample std	35.16	24.35	18.47	19.36	26.21	18.65	18.28

Table 8: Out-of-sample performance of Bayesian-DRO with non-contaminated data and size 100 in three-dimensional problem.

m=50	ϵ_1	ϵ_2	ϵ_3	ϵ_4	$\epsilon = 0$	empirical	true

(avg) ϵ value	1.51(0.02)
	1.58(0.02)
	1.37(0.01)
		0.75(0.01)
		0.79(0.01)
		0.69(0.01)
			1.02(0.45)
			1.11(0.32)
			1.13(0.32)
				0.90
				0.80
				0.70
					-	-	-

solution	21.36 32.80 40.45	16.38 25.23 31.63	18.16 28.34 37.18	17.32 25.35 31.81	9.24 14.10 18.36	17.25 28.89 36.26	19.28 29.10 38.92
out-of-sample mean	47.85	55.64	42.39	73.32	144.02	44.29	41.19
out-of-sample std	22.23	21.42	18.20	24.01	26.35	18.72	18.28

Table 9: Out-of-sample performance of Bayesian-DRO with contaminated data and size 50 in three-dimensional problem.

- When the data size is small or when the data are contaminated, the Bayesian-DRO approach outperforms the empirical approach, since the Bayesian estimate reduces the variability from data.

- Based on our analysis in Section 3.4 and numerical results, we recommend to choose the ambiguity set size of Bayesian-DRO to be ϵ_3, i.e., solving problem (3.32) with the true

m=1000	ϵ_1	ϵ_2	ϵ_3	ϵ_4	$\epsilon = 0$	empirical	true

(avg) ϵ value	1.30(0.00)
	1.35(0.00)
	1.53(0.00)
		0.65(0.00)
		0.67(0.00)
		0.76(0.00)
			1.09(0.17)
			1.29(0.09)
			1.13(0.11)
				0.90
				0.70
				0.50
					-	-	-

solution	21.19 30.64 43.46	16.66 24.19 33.22	19.74 30.10 38.08	18.42 24.45 29.61	9.59 13.97 18.44	17.96 27.19 36.20	19.28 29.10 38.92
out-of-sample mean	47.85	55.64	41.61	73.32	144.02	44.29	41.19
out-of-sample std	22.23	21.42	18.57	24.01	26.35	18.72	18.28

Table 10: Out-of-sample performance of Bayesian-DRO with contaminated data and size 1000 in three-dimensional problem.

optimal solution replaced by the empirical optimal solution.

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-Liebler divergence and the size of the set is small, the robustness of Bayesian-DRO can be interpreted as a trade-off between mean and standard deviation of the objective 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 with the appropriate choice of ambiguity set size, Bayesian-DRO has superior out-of-sample performance compared to the Bayesian-average approach and the empirical approach, especially when data are limited and contaminated.

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

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

References

[1] G. Bayraksan and D. K. Love. Data-driven stochastic programming using phi-divergences. *Tutorials in Operations Research, INFORMS*, pages 1563–1581, 2015.

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

[3] J. Blanchet, Y. Kang, and K. Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56:830–857, 2019.

[4] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozls*, 8, 1963.

[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] J.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. *arXiv preprint arXiv:1505.05116*, 2015.

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

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

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

[12] Vishal Gupta. Near-optimal bayesian ambiguity sets for distributionally robust optimization. *Manag. Sci.*, 65:4242–4260, 2019.

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

[14] T. Morimoto. Markov processes and the h-theorem. *J. Phys. Soc. Jap.*, 18(3):328–333, 1963.

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

[16] A. Pichler and A. Shapiro. Mathematical foundations of distributionally robust multistage optimization. *https://arxiv.org/abs/2101.02498*, 2020.

[17] R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis.* Springer, Berlin, 1998.

[18] H. Scarf. A min-max solution of an inventory problem. In *Studies in the Mathematical Theory of Inventory and Production*, pages 201–209. Stanford University Press, 1958.

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

[20] A. Shapiro. Distributionally robust stochastic programming. *SIAM J. Optimization*, 27:2258–2275, 2017.

[21] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory.* SIAM, Philadelphia, 2009.

[22] 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.

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