Accepted to IEEE Journal of Selected Topics in Signal Processing

Learning Against Distributional Uncertainty: On the Trade-off Between Robustness and Specificity

Shixiong Wang, Haowei Wang, Xinke Li, and Jean Honorio

Abstract—**Trustworthy machine learning aims at combating distributional uncertainties in training data distributions compared to population distributions. Typical treatment frameworks include the Bayesian approach, (min-max) distributionally robust optimization (DRO), and regularization. However, three issues have to be raised: 1) the prior distribution in the Bayesian method and the regularizer in the regularization method are difficult to specify; 2) the DRO method tends to be overly conservative; 3) all the three methods are biased estimators of the true optimal cost. This paper studies a new framework that unifies the three approaches and addresses the three challenges above. The asymptotic properties (e.g., consistencies and asymptotic normalities), non-asymptotic properties (e.g., generalization bounds and unbiasedness), and solution methods of the proposed model are studied. The new model reveals the trade-off between the robustness to the unseen data and the specificity to the training data. Experiments on various real-world tasks validate the superiority of the proposed learning framework.**

Index Terms—**Generalization Error, Distributional Robustness, Bayesian Nonparametrics, Regularization.**

I. Introduction

Supervised statistical machine learning can be modeled by the following optimization problem [1], [2]:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \ \mathbb{E}_{\xi\sim\mathbb{P}_0} h(\boldsymbol{x},\xi), \tag{1}$$

in which $\boldsymbol{x}\in\mathcal{X}\subseteq\mathbb{R}^l$ is the decision vector and $\xi\in\Xi\subseteq\mathbb{R}^k$ is the random parameter whose underlying distribution is $\mathbb{P}_0$; the cost function is denoted by $h:\mathcal{X}\times\Xi\to\mathbb{R}$ (particularly $\mathbb{R}_+$). Specifically, hypotheses are parameterized by $\boldsymbol{x}$ and $\xi := (\xi_{\text{in}},\xi_{\text{out}})$ denotes a data pair where ξ_{in} and ξ_{out} denote the feature and expected response, respectively.

In the practice of machine learning, the true population distribution $\mathbb{P}_0$ is unknown, and the empirical distribution $\hat{\mathbb{P}}_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{\xi_i}$, where δ_{ξ_i} is the Dirac distribution concentrated at the point ξ_i, constructed by n independent and identically distributed (i.i.d.) samples $\{\xi_i\}_{i\in[n]}$ is the most common estimate of $\mathbb{P}_0$. As a result, we can use the data-driven **nominal model** [1]

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\xi\sim\hat{\mathbb{P}}_n} h(\boldsymbol{x},\xi) \tag{2}$$

as an approximation to **true model** (1) to find the optimal decision. In the literature, (2) is known as an empirical risk minimization (ERM) model or a sample-average approximation (SAA) model. However, there exists a distributional mismatch (i.e., **distributional uncertainty**) between $\hat{\mathbb{P}}_n$ and $\mathbb{P}_0$ due to scarce data and the approximation error of (2) to (1) vanishes only as $n\to\infty$. Neglecting such distributional uncertainty in $\hat{\mathbb{P}}_n$ may cause significant performance degradation: $\mathbb{E}_{\mathbb{P}_0} h(\hat{\boldsymbol{x}}_n,\xi)$ may be significantly larger than $\min_{\boldsymbol{x}}\mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x},\xi)$ due to overfitting, where $\hat{\boldsymbol{x}}_n$ solves (2). Mitigating the adverse impact resulting from the distributional uncertainty in $\hat{\mathbb{P}}_n$ and controlling the generalization error $\mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}^\star,\xi) - \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}^\star,\xi)$ by selecting a promising decision $\boldsymbol{x}^\star$, in $\mathbb{P}_0^n$-probability or in $\mathbb{P}_0^n$-expectation, lie in the core of trustworthy machine learning, where $\mathbb{P}_0^n$ is the joint distribution of n i.i.d. training samples.

A. Literature Review

Bayesian methods [3], [4] are the first choice to deal with the distributional mismatch in $\hat{\mathbb{P}}_n$. Suppose $\mathcal{C}$ is a family of admissible distributions on the measurable space $(\Xi,\mathcal{B}_\Xi)$ where $\mathcal{B}_\Xi$ denotes the Borel σ-algebra on Ξ; in the literature, $\mathcal{C}$ is also called an **ambiguity set**. For instance, in consideration of the nominal problem (2), $\mathcal{C}$ can be defined as a closed distributional ball with center $\hat{\mathbb{P}}_n$ and radius ϵ_n, that is, $\mathcal{C} := B_{\epsilon_n}(\hat{\mathbb{P}}_n)$. Bayesian approaches attempt to design a probability measure $\mathbb{Q}$ on $(\mathcal{C},\mathcal{B}_\mathcal{C})$, where $\mathcal{B}_\mathcal{C}$ denotes the Borel σ-algebra on $\mathcal{C}$ [5], and the following Bayesian counterpart for the nominal problem (2) is solved:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}\sim\mathbb{Q}}\mathbb{E}_{\xi\sim\mathbb{P}} h(\boldsymbol{x},\xi). \tag{3}$$

In this case, the true population distribution $\mathbb{P}_0$ is expected to be included in $\mathcal{C}$ and an ideal $\mathbb{Q}$ should be the one that lets the distributions in $\mathcal{C}$ concentrate at $\mathbb{P}_0$. Namely, $\mathbb{P}_0$ is the element most likely to be sampled from $\mathcal{C}$ according to $\mathbb{Q}$. Under some mild technical conditions, we can find a point $\mathbb{P}'\in\mathcal{C}$ satisfying $\mathbb{E}_\mathbb{Q}\mathbb{E}_\mathbb{P} h(\boldsymbol{x},\xi) = \mathbb{E}_{\mathbb{P}'} h(\boldsymbol{x},\xi)$ for all $\boldsymbol{x}$ (see Lemma 1). Hence, essentially, Bayesian methods tell us how to locate the "best" candidate in $\mathcal{C}$. If $\mathbb{P}'$ is closer to $\mathbb{P}_0$ than $\hat{\mathbb{P}}_n$ to $\mathbb{P}_0$, the Bayesian method (3) would have a smaller approximation error for (1) than the nominal method (2) would have; examples and justifications can be accessed in, e.g., [6], [7]. Note that either (resp. both) $\mathbb{Q}$ or (resp. and) $\mathbb{P}$ can be parametric distributions.

S. Wang is with the Institute of Data Science, National University of Singapore, Singapore 117602 (E-mail: s.wang@u.nus.edu).

H. Wang is with Rice-Rick Digitalization, Singapore 179098 (E-mail: haowei_wang@ricerick.com).

X. Li is with the Department of Data Science, City University of Hong Kong, Kowloon, Hong Kong. (E-mail: xinkeli@cityu.edu.hk).

J. Honorio is with School of Computing and Information Systems, The University of Melbourne, and ARC Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA) (Email: jean.honorio@unimelb.edu.au).

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2RP-2020-018).

Corresponding Author: Haowei Wang.

Regularization approaches are another promising choice to hedge against the distributional uncertainty in $\hat{\mathbb{P}}_n$ [8, Sec. A1.3]. To be specific, a regularization term $f(\boldsymbol{x})$ is employed and the regularized counterpart

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\xi\sim\hat{\mathbb{P}}_n} h(\boldsymbol{x},\xi) + \lambda f(\boldsymbol{x}) \tag{4}$$

for the nominal empirical risk minimization problem (2) is studied, in which $\lambda \geq 0$ is a balancing coefficient. For example, the regularizer $f(\boldsymbol{x})$ can be a proper norm $\|\boldsymbol{x}\|$ on $\mathcal{X}$ [9, Chap. 7], and λ may depend on the sample size n. Regularization methods are believed to be able to work against "overfitting" and reduce generalization errors in a great number of learning problems; one may reminisce about the "bias-variance trade-off" in the machine learning literature [10, Sec. 2.9]. The rationale of the regularization methods can also be quantitatively justified from many other perspectives such as the measure concentration inequalities [11], the stability properties of the learning algorithms [12, Sec. 5.2], and the PAC-Bayesian learning [13, Sec. 2], to name a few. The key point is that the regularized SAA cost $\mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x},\xi) + \lambda f(\boldsymbol{x})$ can be an upper bound of the unknown true cost $\mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x},\xi)$ for all $\boldsymbol{x}$, and therefore, by minimizing the regularized SAA cost, the unknown true cost can also be controlled. However, the SAA cost $\mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x},\xi)$ solely cannot serve as an upper bound of the true cost.

The (min-max) distributionally robust optimization (DRO) counterpart

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\mathbb{P}\in\mathcal{C}} \mathbb{E}_{\xi\sim\mathbb{P}} h(\boldsymbol{x},\xi) \tag{5}$$

for the nominal model (2) is another potential approach to handle the distributional uncertainty in $\hat{\mathbb{P}}_n$ [14], [15], [16]. If the distributional family $\mathcal{C}$ contains the true distribution $\mathbb{P}_0$, then the inequality $\mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x},\xi) \leq \max_{\mathbb{P}\in\mathcal{C}} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x},\xi)$ holds for all $\boldsymbol{x}$, and therefore, by minimizing the robust cost $\max_{\mathbb{P}\in\mathcal{C}} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x},\xi)$, the unknown true cost can also be controlled; for more interpretations and justifications of the DRO method, see [1], [15]. According to, e.g., [17], [18], we can find a point $\mathbb{P}'$ in $\mathcal{C}$ such that $\max_{\mathbb{P}\in\mathcal{C}} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x},\xi) = \mathbb{E}_{\mathbb{P}'} h(\boldsymbol{x},\xi)$, for all $\boldsymbol{x}$, if some mild technical conditions on the function h can be satisfied. Therefore, as an alternative to the Bayesian approach (3), the DRO approach (5) chooses the "best" candidate $\mathbb{P}'$ in $\mathcal{C}$ from another perspective. However, in the practice of DRO methods, elegantly specifying the size parameter ϵ_n of the employed ambiguity set $\mathcal{C} := B_{\epsilon_n}(\hat{\mathbb{P}}_n)$ is not easy because the radius can be neither too large nor too small. A small radius cannot guarantee $\mathbb{P}_0$ to be included in $\mathcal{C}$. Consequently, the worst-case cost $\max_{\mathbb{P}\in\mathcal{C}} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x},\xi)$ cannot provide an upper bound for the unknown true cost. Conversely, if the radius is too large, the DRO methods would become overly conservative and the upper bound of the true cost specified by $\max_{\mathbb{P}\in\mathcal{C}} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x},\xi)$ may be extremely loose. In the DRO literature, typical design methods for ϵ_n and their drawbacks are as follows.

1) The measure concentration bounds in, e.g., [19] and [1], are just theoretical results, far away from practical utilization, because the involved constants depend on the true underlying distributions, which are unknown. In addition, measure concentration bounds are not tight. Third, measure concentration bounds are dependent on the dimension of ξ, and therefore, they may face the curse of dimensionality [20].
2) Practical methods such as cross-validation [18, p. 156] and bootstrap [18, p. 158] are reliable if and only if the data size n is sufficiently large. When n is small, they may not work well [21], [22].
3) Statistical inference methods presented in [23], [24] also require n to be large because the optimality of the presented methods is established in the asymptotic sense (i.e., when $n \to \infty$).

According to, e.g., [1, Thm. 10], [2], under some technical conditions, the DRO approach (5) amounts to a regularized empirical risk minimization method (4), which also advocates why the DRO approach (5) is able to combat overfitting and provide excellent generalization performance.

B. Research Gaps and Motivations

It is practically uneasy to specify prior distribution $\mathbb{Q}$ in Bayesian method (3), regularizer $f(\boldsymbol{x})$ in regularization method (4), and radius ϵ_n of distributional ball $B_{\epsilon_n}(\hat{\mathbb{P}}_n)$ in DRO method (5). The three quantities cannot be arbitrarily specified, otherwise, the performances of the three associated methods cannot be guaranteed. For example, as explained before, ϵ_n can be neither too large nor too small. Therefore, the first motivation of this work is to design a new framework that frees us from the elaborate selection of prior distribution $\mathbb{Q}$, regularizer $f(\boldsymbol{x})$, and radius ϵ_n.

In addition, the DRO approach, SAA approach, and regularized SAA approach are biased estimators of the true optimal objective value (1) when n is finite; the biases only vanish asymptotically (i.e., as $n \to \infty$). Hence, the second motivation of this work is to design a new model that is able to be unbiased for finite n, which brings the asymptotic statistical property to finite-sample learning.

C. Contributions

The contributions of this paper can be summarized as follows.

1) A new framework that can combat the distributional uncertainty in $\hat{\mathbb{P}}_n$ is designed; see Section III, and Models (9) and (10). The framework generalizes Bayesian method (3), regularization method (4), and DRO method (5) and suggests the instructions in designing $\mathbb{Q}$ and $f(\boldsymbol{x})$; see Remark 1. In addition, the framework reveals the trade-off between the robustness to the unseen data (i.e., the adverse distributional uncertainty in $\hat{\mathbb{P}}_n$) and the specificity to the training data (i.e., the exploitable empirical information in $\hat{\mathbb{P}}_n$); see Remark 2. Moreover, the framework can diminish the conservatism, and therefore improve the performance, of the DRO method; see Theorem 2, Remark 4, and Examples 1 and 2. Statistical properties of the new learning model such as consistencies, asymptotic normalities, generalization bounds, and unbiasedness are established; see Theorems 1, 2, and 3.
2) The proposed new model is specifically studied under the ϕ-divergence and Wasserstein distributional balls, and

respective solution methods are derived; see Section V. In particular, the solutions disclose two important insights from the perspective of data augmentation (see Examples 3 and 4), which intuitively explain the flexibility of the proposed learning model.

II. Notations and Preliminaries

Notations used in this paper are summarized in Appendix A-A. Necessary DRO theories are reviewed in Appendices A-B and A-C. Statistical concepts including Glivenko–Cantelli class, Donsker class, and Brownian bridge are presented in Appendix A-D. In this section, we focus on a reformulation of the Bayesian model (3). We start with the concept of mean distribution.

Definition 1 (Mean Distribution). *A distribution $\bar{\mathbb{P}}$ satisfying $\bar{\mathbb{P}}(E) = \int_{\mathbb{R}} \mathbb{P}(E)\mathbb{Q}(\mathrm{d}\mathbb{P}(E))$, $\forall E \in \mathcal{B}_{\Xi}$ is a mean distribution of $\mathbb{P}$ under $\mathbb{Q}$.* □

Namely, the mean distribution is a mixture of distributions in $\mathcal{C}$ with weights determined by $\mathbb{Q}$. To be specific, for an event E in $\mathcal{B}_{\Xi}$, $\mathbb{P}(E)$ is a random variable taking values on $\mathbb{R}_+$ and its distribution is specified by $\mathbb{Q}$. This definition can transform Bayesian model (3).

Lemma 1 ([25]). *If $\bar{\mathbb{P}}$ is the mean distribution of $\mathbb{P}$ under $\mathbb{Q}$ and $\mathbb{E}_{\mathbb{Q}}\mathbb{E}_{\mathbb{P}}|h(\boldsymbol{x}, \xi)| < \infty$, then $\mathbb{E}_{\mathbb{Q}}\mathbb{E}_{\mathbb{P}}h(\boldsymbol{x}, \xi) = \mathbb{E}_{\bar{\mathbb{P}}}h(\boldsymbol{x}, \xi)$ for every $\boldsymbol{x}$.* □

In terms of model (3), the most popular choice for a non-parametric prior distribution $\mathbb{Q}$ of $\mathbb{P}$, in Bayesian nonparametrics, is the Dirichlet-process prior. Furthermore, when the n-sample empirical distribution $\hat{\mathbb{P}}_n$ is considered, the posterior non-parametric distribution of $\mathbb{P}$ is still a Dirichlet process whose mean distribution is $\frac{\alpha}{\alpha+n}\hat{\mathbb{P}} + \frac{n}{\alpha+n}\hat{\mathbb{P}}_n$, where $\hat{\mathbb{P}}$ is *a priori* knowledge of $\mathbb{P}_0$ and $\alpha \geq 0$ is employed to quantify the trust level towards $\hat{\mathbb{P}}$ [3], [4, Chap. 3]. Specifically, if we trust the prior $\hat{\mathbb{P}}$ more than the empirical distribution $\hat{\mathbb{P}}_n$, α should be large. When the Dirichlet-process prior is utilized, as a result of Lemma 1, the Bayesian model (3) becomes

$$\min_{\boldsymbol{x}} \frac{\alpha}{\alpha+n}\mathbb{E}_{\hat{\mathbb{P}}}h(\boldsymbol{x}, \xi) + \frac{n}{\alpha+n}\mathbb{E}_{\hat{\mathbb{P}}_n}h(\boldsymbol{x}, \xi). \tag{6}$$

It can be generalized into

$$\min_{\boldsymbol{x}} \beta_n \mathbb{E}_{\hat{\mathbb{P}}}h(\boldsymbol{x}, \xi) + (1-\beta_n)\mathbb{E}_{\hat{\mathbb{P}}_n}h(\boldsymbol{x}, \xi) \tag{7}$$

where the weight $\beta_n \in [0, 1]$ depends on sample size n; β_n can be an arbitrary sequence satisfying $\beta_n \to 0$ as $n \to \infty$. Model (7) serves as a foundation for the new machine learning framework that we propose subsequently.

III. New Framework: Bayesian Distributionally Robust Learning

In real-world operation, it is often difficult to specify an exact (non-parametric Bayesian) prior $\mathbb{Q}$ for a Bayesian model (3). This motivates us to study the second-order min-max (or worst-case) Bayesian distributionally robust optimization counterpart for the nominal model (2)

$$\min_{\boldsymbol{x}} \max_{\mathbb{Q}} \mathbb{E}_{\mathbb{P}\sim\mathbb{Q}}\mathbb{E}_{\xi\sim\mathbb{P}}h(\boldsymbol{x}, \xi), \tag{8}$$

which is a robustified version of the Bayesian model. In particular, model (8) is a combination of a Frequentist and a Bayesian method: The random measure $\mathbb{P}$ follows the second-order probability measure $\mathbb{Q}$, and therefore, in terms of $\mathbb{P}$, (8) is a Bayesian method; the admissible values of $\mathbb{Q}$ are only assumed to lie in an ambiguity set (which is not explicitly specified here), and therefore, in terms of $\mathbb{Q}$, (8) is a Frequentist method.

Inspired by (8), we shall study the worst-case version of (7):

$$\min_{\boldsymbol{x}\in\mathcal{X}} \beta_n \max_{\mathbb{P}\in B_{\epsilon}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{P}}h(\boldsymbol{x}, \xi) + (1-\beta_n)\mathbb{E}_{\hat{\mathbb{P}}_n}h(\boldsymbol{x}, \xi). \tag{9}$$

Note that the uncertainty in $\mathbb{Q}$ is reflected by the uncertainty in the prior estimate $\hat{\mathbb{P}}$ because $\hat{\mathbb{P}}_n$ is completely determined given samples $\{\xi_i\}_{i\in[n]}$.

Remark 1 (Interpretation of Model (9)). *Model* (9) *is a Bayesian non-parametric model in terms of the data distribution $\mathbb{P}$ and also a Frequentist distributionally robust optimization model in terms of the distribution $\mathbb{Q}$ of the data distribution; cf.* (8). *Since* (9) *is equivalent to* $\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}_n}h(\boldsymbol{x}, \xi) + \frac{\beta_n}{1-\beta_n}\max_{\mathbb{P}\in B_{\epsilon}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{P}}h(\boldsymbol{x}, \xi)$, *by letting* $\lambda_n := \frac{\beta_n}{1-\beta_n}$ *and*

$$f(\boldsymbol{x}) := \max_{\mathbb{P}\in B_{\epsilon}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{P}}h(\boldsymbol{x}, \xi),$$

(9) *can be rewritten as* $\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\hat{\mathbb{P}}_n}h(\boldsymbol{x}, \xi) + \lambda_n f(\boldsymbol{x})$, *which is a regularized SAA model* (4). *Also, when* $\beta_n := 1$, (9) *reduces to a DRO model* (5)*; when* $\beta_n := 0$, (9) *reduces to a SAA model* (2). *Hence, the new model* (9) *is a generalized model that unifies the SAA model* (2), *the Bayesian model* (3), *the regularized SAA model* (4), *and the DRO model* (5). *The benefit is that* (9) *suggests how to design $\mathbb{Q}$ in the Bayesian method* (3) *and $f(\boldsymbol{x})$ in the regularization method* (4). □

In practice, it is uneasy to specify $\hat{\mathbb{P}}$. Alternatively, if the distributional ambiguity set is constructed around $\hat{\mathbb{P}}_n$ rather than $\hat{\mathbb{P}}$, the model (9) becomes completely data-driven:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \beta_n \max_{\mathbb{P}\in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}}h(\boldsymbol{x}, \xi) + (1-\beta_n)\mathbb{E}_{\hat{\mathbb{P}}_n}h(\boldsymbol{x}, \xi). \tag{10}$$

This is a change-of-center trick for the employed distributional ambiguity set: Non-rigorously speaking, we are assuming $\hat{\mathbb{P}}$ is contained in $B_{\epsilon_{n,1}}(\hat{\mathbb{P}}_n)$ and $\hat{\mathbb{P}}_n$ is contained in $B_{\epsilon_{n,2}}(\hat{\mathbb{P}})$ for some radii $\epsilon_{n,1}, \epsilon_{n,2} \geq 0$. We call (10) a **Bayesian distributionally robust** (BDR) optimization.

Remark 2 (Robustness-Specificity Trade-off). *Since the objective of* (10) *balances the worst-case cost specified by DRO and the nominal cost specified by SAA, the new model* (10) *reveals the trade-off between the robustness to the distributional uncertainty (i.e., unseen data) and the specificity to the empirical information (i.e., training data).* □

In the following, we use a linear regression example with Gaussian data distribution to intuitively explain the BDR learning framework. Consider the data generating distribution $\xi := [\xi_{\text{in}}; \xi_{\text{out}}] \sim N(\mathbf{0}, \mathbf{\Sigma}_0)$ and the linear regression model $\xi_{\text{out}} = \boldsymbol{x}^\top\xi_{\text{in}} + e$ where $e \in \mathbb{R}$ denotes the regression residual. The true optimization problem $\min_{\boldsymbol{x}}[\boldsymbol{x}^\top, -1]\mathbb{E}_{\mathbb{P}_0}\xi\xi^\top[\boldsymbol{x}; -1]$ admits

$$\min_{\boldsymbol{x}\in\mathcal{X}} [\boldsymbol{x}^\top, -1] \cdot \mathbf{\Sigma}_0 \cdot [\boldsymbol{x}; -1]. \tag{True}$$

Denoting $\hat{\boldsymbol{\Sigma}}_n$ as the sample-estimate of $\boldsymbol{\Sigma}_0$, the SAA counterpart $\min_{\boldsymbol{x}}[\boldsymbol{x}^\top, -1]\mathbb{E}_{\hat{\mathbb{P}}_n}\xi\xi^\top[\boldsymbol{x}; -1]$ is

$$\min_{\boldsymbol{x}\in\mathcal{X}}[\boldsymbol{x}^\top, -1]\cdot\hat{\boldsymbol{\Sigma}}_n\cdot[\boldsymbol{x}; -1]. \tag{SAA}$$

The DRO counterpart $\min_{\boldsymbol{x}}\max_{\mathbb{P}}[\boldsymbol{x}^\top, -1]\mathbb{E}_{\mathbb{P}}\xi\xi^\top[\boldsymbol{x}; -1]$ under the order-2 Wasserstein ball $W_2(\mathbb{P}, \hat{\mathbb{P}}_n) \le \epsilon_n$ is

$$\begin{array}{ll} \min_{\boldsymbol{x}}\max_{\boldsymbol{\Sigma}} & [\boldsymbol{x}^\top, -1]\boldsymbol{\Sigma}[\boldsymbol{x}; -1] \\ \text{s.t.} & \operatorname{Tr}[\boldsymbol{\Sigma} + \hat{\boldsymbol{\Sigma}}_n - 2(\boldsymbol{\Sigma}^{1/2}\hat{\boldsymbol{\Sigma}}_n\boldsymbol{\Sigma}^{1/2})^{1/2}] \le \epsilon_n^2, \end{array}$$

for which the von Neumann's minimax theorem holds. If $\boldsymbol{\Sigma}_n^*$ solves the above display ($\boldsymbol{\Sigma}_n^*$ may depend on $\boldsymbol{x}$), the DRO problem becomes

$$\min_{\boldsymbol{x}\in\mathcal{X}}[\boldsymbol{x}^\top, -1]\cdot\boldsymbol{\Sigma}_n^*\cdot[\boldsymbol{x}; -1]. \tag{DRO}$$

As a result, the BDR counterpart is

$$\min_{\boldsymbol{x}\in\mathcal{X}}[\boldsymbol{x}^\top, -1]\cdot[\beta_n\boldsymbol{\Sigma}_n^* + (1-\beta_n)\hat{\boldsymbol{\Sigma}}_n]\cdot[\boldsymbol{x}; -1]. \tag{BDR}$$

IV. Statistical Properties of BDR Model (10)

This subsection studies the asymptotic and non-asymptotic statistical properties of the new BDR model (10) under any appropriate distributional ball $B_{\epsilon_n}(\hat{\mathbb{P}}_n)$, for example, the ϕ-divergence ball or the Wasserstein ball, whose mathematical definitions can be found in Appendix A-B. Statistical concepts such as Glivenko–Cantelli class, Donsker class, and Brownian bridge can be found in Appendix A-D; see also [26, Chap. 19]. The key notations in this subsection are given in Table I.

TABLE I
Notation list. ("Opt. Sln." stands for Optimal Solution.)

Notation	Definition	Mathematical Form
$v(\boldsymbol{x})$	True Cost	$\mathbb{E}_{\mathbb{P}_0}h(\boldsymbol{x},\xi)$
$v_n(\boldsymbol{x})$	SAA Cost	$\mathbb{E}_{\hat{\mathbb{P}}_n}h(\boldsymbol{x},\xi)$
$v_{r,n}(\boldsymbol{x})$	DRO Cost	$\max_{\mathbb{P}\in B_{\epsilon_n}(\hat{\mathbb{P}}_n)}\mathbb{E}_{\mathbb{P}}h(\boldsymbol{x},\xi)$
$v_{b,n}(\boldsymbol{x})$	BDR Cost	$\beta_n\max_{\mathbb{P}\in B_{\epsilon_n}(\hat{\mathbb{P}}_n)}\mathbb{E}_{\mathbb{P}}h(\boldsymbol{x},\xi)$ $+(1-\beta_n)\mathbb{E}_{\hat{\mathbb{P}}_n}h(\boldsymbol{x},\xi)$
$\mathcal{X}_0$	True Opt. Sln. Set	$\operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} v(\boldsymbol{x})$
$\hat{\mathcal{X}}_n$	SAA Opt. Sln. Set	$\operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} v_n(\boldsymbol{x})$
$\hat{\mathcal{X}}_{r,n}$	DRO Opt. Sln. Set	$\operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} v_{r,n}(\boldsymbol{x})$
$\hat{\mathcal{X}}_{b,n}$	BDR Opt. Sln. Set	$\operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} v_{b,n}(\boldsymbol{x})$
$\boldsymbol{x}_0$	True Opt. Sln.	$\boldsymbol{x}_0 \in \mathcal{X}_0$
$\hat{\boldsymbol{x}}_n$	SAA Opt. Sln.	$\hat{\boldsymbol{x}}_n \in \hat{\mathcal{X}}_n$
$\hat{\boldsymbol{x}}_{r,n}$	DRO Opt. Sln.	$\hat{\boldsymbol{x}}_{r,n} \in \hat{\mathcal{X}}_{r,n}$
$\hat{\boldsymbol{x}}_{b,n}$	BDR Opt. Sln.	$\hat{\boldsymbol{x}}_{b,n} \in \hat{\mathcal{X}}_{b,n}$

A. Asymptotic Properties of (10)

We consider the $\boldsymbol{x}$-parametric function class

$$\mathcal{H} := \{h(\boldsymbol{x},\cdot) : \Xi \to \mathbb{R} | \boldsymbol{x} \in \mathcal{X}\} \tag{11}$$

indexed by $\mathcal{X}$. The asymptotic properties of Bayesian distributionally robust model (10) are given below, which illustrate the learning effectiveness when the sample size becomes infinitely large, as the generalization error approaches zero.

Theorem 1 (Asymptotic Properties of (10))**.** *Consider the nominal problem* (2) *and its Bayesian distributionally robust counterpart* (10)*. If the following conditions hold*

C1) The DRO objective $v_{r,n}(\boldsymbol{x})$ is bounded in $\mathbb{P}_0^n$-probability and attainable for $\boldsymbol{x} \in \mathcal{X}' \subseteq \mathcal{X}$;

C2) The weight coefficient $\beta_n \in [0, 1]$ for every n and $\sqrt{n}\beta_n \to 0$ as $n \to \infty$;

C3) The function class $\mathcal{H}$ in (11) *is $\mathbb{P}_0$-Glivenko–Cantelli;*

C4) At least one of the following properties holds for the function $v(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}_0}h(\boldsymbol{x},\xi)$:

C4a) $v(\boldsymbol{x})$ is continuous on $\mathcal{X}'$;

C4b) $v(\boldsymbol{x})$ has the unique global minimizer $\boldsymbol{x}_0$ on $\mathcal{X}'$;

C5) The function class $\mathcal{H}$ in (11) *is $\mathbb{P}_0$-Donsker;*

C6) $\mathbb{E}_{\mathbb{P}_0}[h(\hat{\boldsymbol{x}}_n,\xi) - h(\boldsymbol{x}_0,\xi)]^2 \xrightarrow{p} 0$ as $\hat{\boldsymbol{x}}_n \xrightarrow{p} \boldsymbol{x}_0$,[1]

then the following statements are true.

S1) (Point-Wise Consistency of Objective Function.) *For every $\boldsymbol{x} \in \mathcal{X}'$, we have $v_{b,n}(\boldsymbol{x}) \xrightarrow{p} v(\boldsymbol{x})$ as $n \to \infty$.*

S2) (Consistency of Optimal Value.) *For every $\hat{\boldsymbol{x}}_{b,n} \in \hat{\mathcal{X}}_{b,n} \subseteq \mathcal{X}'$ and every $\boldsymbol{x}_0 \in \mathcal{X}_0 \subseteq \mathcal{X}'$, we have $v_{b,n}(\hat{\boldsymbol{x}}_{b,n}) \xrightarrow{p} v(\boldsymbol{x}_0)$ as $n \to \infty$. In other words, $\min_{\boldsymbol{x}} v_{b,n}(\boldsymbol{x}) \xrightarrow{p} \min_{\boldsymbol{x}} v(\boldsymbol{x})$ as $n \to \infty$.*

S3) (Consistency of Optimal Solution.) *The limit point of any solution sequence $\{\hat{\boldsymbol{x}}_{b,n}\}$ of* (10) *is a solution of the true problem* (1) *in $\mathbb{P}_0^n$-probability: $\mathbb{P}_0^n\{\hat{\mathcal{X}}_{b,n} \subseteq \mathcal{X}_0\} \to 1$ as $n \to \infty$.*

S4) (Point-Wise Asymptotic Normality of Objective Function.) *For every $\boldsymbol{x} \subseteq \mathcal{X}'$, we have $\sqrt{n}[v_{b,n}(\boldsymbol{x}) - v(\boldsymbol{x})] \xrightarrow{d} N(0, V_{v,\boldsymbol{x}})$ as $n \to \infty$, where $V_{v,\boldsymbol{x}} := \mathbb{D}_{\mathbb{P}_0}h(\boldsymbol{x},\xi)$ denotes the variance of $h(\boldsymbol{x},\xi)$ under $\mathbb{P}_0$.*

S5) (Asymptotic Normality of Optimal Value.) *For every $\hat{\boldsymbol{x}}_{b,n} \in \hat{\mathcal{X}}_{b,n} \subseteq \mathcal{X}'$ and every $\boldsymbol{x}_0 \in \mathcal{X}_0 \subseteq \mathcal{X}'$, if $\hat{\boldsymbol{x}}_{b,n} \xrightarrow{p} \boldsymbol{x}_0$, we have $\sqrt{n}[v_{b,n}(\hat{\boldsymbol{x}}_{b,n}) - v(\boldsymbol{x}_0)] \xrightarrow{d} N(0, V_v)$ as $n \to \infty$, where $V_v := \mathbb{D}_{\mathbb{P}_0}h(\boldsymbol{x}_0,\xi)$.*

Proof. See Appendix D-A in the supplementary materials. □

Remark 3 (Practicability of Conditions)**.** *The conditions C1)-C6) stipulated in Theorem 1 are not restrictive, as they can be easily fulfilled in practice; concrete examples can be found in Appendix B.* □

Note that in conducting minimization over $\boldsymbol{x}$, it is sufficient to only consider the subset $\mathcal{X}'$ where objective functions are finite-valued. Note also that when the DRO objective $v_{r,n}(\boldsymbol{x})$ is finite at $\boldsymbol{x}$, the SAA objective $v_n(\boldsymbol{x})$ and the true objective $v(\boldsymbol{x})$ will be finite as well because $\hat{\mathbb{P}}_n$ and $\mathbb{P}_0$ are included in $B_{\epsilon_n}(\hat{\mathbb{P}}_n)$ for sufficiently large ϵ_n. The asymptotic normality of the optimal solution of the BDR model (10), which requires stronger and therefore more restrictive technical conditions, is deferred to Appendix D-B in the supplementary materials.

B. Non-Asymptotic Properties of (10)

First, we discuss the one-sided generalization bound, which is a crucial non-asymptotic property in machine learning.

DRO learning has better generalization ability than traditional ERM learning because by reducing DRO cost $v_{r,n}(\boldsymbol{x})$, true cost $v(\boldsymbol{x})$ can also be diminished; however, ERM cost $v_n(\boldsymbol{x})$ cannot upper bound $v(\boldsymbol{x})$. Nevertheless, DRO learning is usually

[1] The notations $\xrightarrow{p}$ and $\xrightarrow{d}$ mean the convergence in probability and distribution, respectively.

criticized for its conservatism. Specifically, to guarantee that the true distribution $\mathbb{P}_0$ is included in the distributional ball, the radius ϵ_n of the ball should be sufficiently large (cf. Appendix A-B2), which leads to that for every $\boldsymbol{x}$, the upper bound $v_{r,n}(\boldsymbol{x})$ may be extremely loose. In what follows, we show that BDR model (10) can be less conservative than the DRO model when the same distributional ball (with the same radius ϵ_n) is shared.

Theorem 2 (Generalization Bound of (10))**.** *For every $\eta \in (0,1]$ and every $\beta_n \in [\beta_n^*, 1]$, if $\mathbb{P}_0^n[\mathbb{P}_0 \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)] \geq 1-\eta$, then the true cost $v(\boldsymbol{x})$ is upper bounded, with $\mathbb{P}_0^n$-probability at least $1-\eta$, by the BDR cost $v_{b,n}(\boldsymbol{x})$:*

$$v(\boldsymbol{x}) \leq \beta_n v_{r,n}(\boldsymbol{x}) + (1-\beta_n) v_n(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{12}$$

where the smallest (i.e., best) value β_n^* of β_n satisfying the above display is

$$\beta_n^* := \max\left\{\max_{\boldsymbol{x}\in\mathcal{X}} \frac{v(\boldsymbol{x}) - v_n(\boldsymbol{x})}{v_{r,n}(\boldsymbol{x}) - v_n(\boldsymbol{x})},\ 0\right\} \tag{13}$$

which takes values on $[0,1]$ and we assume that $0/0 = 0$; in addition, $\beta_n^* < 1$ if one of the following conditions holds:

C1) $v_{r,n}(\boldsymbol{x}) > v(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathcal{X}$;

C2) $v_n(\boldsymbol{x}) = v(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$ such that $v_{r,n}(\boldsymbol{x}) = v(\boldsymbol{x})$.

Proof. See Appendix D-C in the supplementary materials. □

Remark 4. *In Theorem 2, the best value β_n^* depends on the unknown true distribution $\mathbb{P}_0$ [via the true cost function $v(\boldsymbol{x})$], which cannot be obtained in practice. This is reminiscent of the practical limitation of the DRO theory where the best radius ϵ_n^* also depends on the unknown true distribution $\mathbb{P}_0$; see Appendix A-B2, especially* (22)*. Hence, both DRO and BDR require empirical parameter tuning in real-world operation. However, Theorem 2 suggests that whenever DRO is empirically perfectly tuned, it is possible to further improve performance by tuning the BDR parameter β_n; recall that DRO and BDR share the same distributional ball (with the same ϵ_n).* □

Theorem 2 justifies the rationale of the BDR learning (10) from the perspective of generalization theory. The BDR generalization bound $v_{b,n}(\boldsymbol{x})$ in (12) tightens the DRO generalization bound $v_{r,n}(\boldsymbol{x})$ for every distributional ball $B_{\epsilon_n}(\hat{\mathbb{P}}_n)$ such that $\mathbb{P}_0 \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)$ because $v_{r,n}(\boldsymbol{x}) \geq v_n(\boldsymbol{x})$. To clarify further, suppose that ϵ_n^* is the smallest value of ϵ_n such that $\mathbb{P}_0 \in B_{\epsilon_n^*}(\hat{\mathbb{P}}_n)$. According to the DRO theory, the DRO cost $v_{r,n}(\boldsymbol{x})$ with $\epsilon_n = \epsilon_n^*$ is the tightest upper bound for the true cost $v(\boldsymbol{x})$. However, Theorem 2 indicates that this DRO bound $v_{r,n}(\boldsymbol{x})$ can be further refined to the BDR bound $v_{b,n}(\boldsymbol{x})$ even when $\epsilon_n = \epsilon_n^*$. The refinement is non-trivial (i.e., $\beta_n^* < 1$) if one of the conditions in Theorem 2 holds, which is the case, e.g., when Ξ is a subspace of $\mathbb{R}^k$. To be specific, see [18, Thm. 6.3] and [2] for $\bar{v}_{r,n}(\boldsymbol{x}) > v(\boldsymbol{x})$ when $\Xi \neq \mathbb{R}^k$, where $\bar{v}_{r,n}(\boldsymbol{x})$ is a computational surrogate (i.e., finite-dimensional reformulation) of $v_{r,n}(\boldsymbol{x})$. To avoid the conservatism of the DRO method, [27] introduces an alternative modeling framework known as robust satisfying. However, [27] is not rooted in DRO, and therefore, most existing DRO-based machine-learning methods cannot be directly upgraded.

Another concrete example for Theorem 2 is as follows.

Example 1. *According to [18, Thm. 6.3], if the cost function h is convex in ξ on $\Xi = \mathbb{R}^k$, the support set Ξ of ξ is a closed and convex set, the order p of the Wasserstein distance is set to $p := 1$, and the employed metric d in the Wasserstein distance is specified by a proper norm $\|\cdot\|$ on Ξ, then the distributionally robust optimization objective exactly equals to a regularized SAA objective, point-wisely for every $\boldsymbol{x} \in \mathbb{R}^l$: i.e., for every $\boldsymbol{x} \in \mathbb{R}^l$, we have*

$$v_{r,n}(\boldsymbol{x}) := \max_{\mathbb{P}: W_p(\mathbb{P}, \hat{\mathbb{P}}_n) \leq \epsilon_n} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) = v_n(\boldsymbol{x}) + \epsilon_n \cdot f(\boldsymbol{x})$$

where $f(\boldsymbol{x}) := \max_{\boldsymbol{\theta}\in\Xi}\{\|\boldsymbol{\theta}\|_* : h^*(\boldsymbol{x}, \boldsymbol{\theta}) < \infty\}$ is a regularization term,[2] $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$, and $h^*(\boldsymbol{x}, \boldsymbol{\theta})$ denotes the Fenchel convex conjugate of $h(\boldsymbol{x}, \xi)$ point-wisely for every given $\boldsymbol{x} \in \mathbb{R}^l$. As a result, the generalization bound specified by the DRO model is

$$v(\boldsymbol{x}) \leq v_{r,n}(\boldsymbol{x}) = v_n(\boldsymbol{x}) + \epsilon_n f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathbb{R}^l.$$

However, Theorem 2 supports that the bound above can be tightened to

$$v(\boldsymbol{x}) \leq v_{b,n}(\boldsymbol{x}) \leq v_n(\boldsymbol{x}) + \beta_n \epsilon_n f(\boldsymbol{x}),\ \forall \boldsymbol{x} \in \mathbb{R}^l,\ \exists \beta_n \in [0,1].$$

The best value of β_n is $\beta_n^* := \max_{\boldsymbol{x}\in\mathcal{X}} \frac{v(\boldsymbol{x}) - v_n(\boldsymbol{x})}{\epsilon_n f(\boldsymbol{x})} \leq 1$. The inequality is strict if 1) the radius ϵ_n is large; or 2) $\mathcal{X}$ is a specified subspace of $\mathbb{R}^l$ on which $v(\boldsymbol{x}) < v_n(\boldsymbol{x}) + \epsilon_n f(\boldsymbol{x})$. One may interpret $\beta_n \epsilon_n$ as the radius of a new distributional ball that may not include $\mathbb{P}_0$ in the DRO sense. However, the true cost can still be upper-bounded, indicating that the conventional DRO bound is not sufficiently tight on the focused region $\mathcal{X}$, although it may be tight on the whole space $\mathbb{R}^l$. □

A specific instance of Example 1 is given below.

Example 2 (1-norm Linear Regression)**.** *Let the data vector be $\xi := [\xi_{in}; \xi_{out}]$ and the true data generating model be $\xi_{out} = \boldsymbol{x}_0^\top \xi_{in} + e$, where $\xi_{in} \sim N(\boldsymbol{0}, \boldsymbol{E}_{k-1})$ denotes the feature vector, $\boldsymbol{E}_{k-1}$ denotes the $(k-1)$-dimensional identity matrix, the standard Gaussian variable $e \in \mathbb{R}$ denotes the regression residual (uncorrelated with ξ_{in}), and $\xi_{out} \in \mathbb{R}$ denotes the response. Consider the 1-norm linear regression problem. Supposing that $\xi \sim \mathbb{P}_0$, we have*

$$v(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}_0} |\xi_{out} - \boldsymbol{x}^\top \xi_{in}|,$$
$$v_n(\boldsymbol{x}) = \mathbb{E}_{\hat{\mathbb{P}}_n} |\xi_{out} - \boldsymbol{x}^\top \xi_{in}|,$$

and according to Example 1 and [29, Eq. (4.5)],

$$\begin{aligned} v_{r,n}(\boldsymbol{x}) &= \max\nolimits_{\mathbb{P}\in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} |\xi_{out} - \boldsymbol{x}^\top \xi_{in}| \\ &= \mathbb{E}_{\hat{\mathbb{P}}_n} |\xi_{out} - \boldsymbol{x}^\top \xi_{in}| + \epsilon_n \|(-\boldsymbol{x}, 1)\|_*. \end{aligned}$$

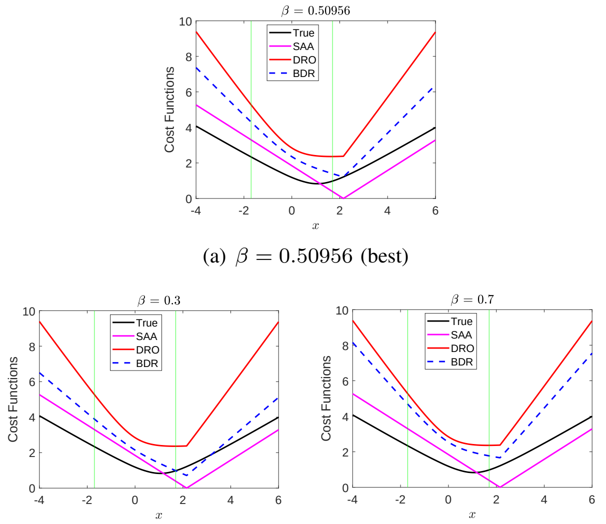
As a demonstration, we particular $\|\cdot\|_*$ into the vector 2-norm. Therefore, the best value β_n^* is

$$\beta_n^* = \max_{\boldsymbol{x}} \frac{\mathbb{E}_{\mathbb{P}_0} |\xi_{out} - \boldsymbol{x}^\top \xi_{in}| - \mathbb{E}_{\hat{\mathbb{P}}_n} |\xi_{out} - \boldsymbol{x}^\top \xi_{in}|}{\epsilon_n \|(-\boldsymbol{x}, 1)\|_2}.$$

To visualize, we examine a one-dimensional case. We set the true parameter be $x_0 = 1$, the sample size $n = 1$, and the radius $\epsilon_n = 1$. Under one realization of $\hat{\mathbb{P}}_n$, the true, SAA,

[2] Similar results are reported in, e.g., [23], [28], [20], [2], where $f(\boldsymbol{x})$ may be of different forms.

DRO, and BDR costs are shown in Fig. 1, where we assume that x takes grid values on $[-4, 6]$ with step size of 0.01. As we can see, if the feasible region of the decision variable x is required to be $[-1.7, 1.7]$, the BDR bound in Fig. 1(a) is no longer tight but the BDR bound in Fig. 1(b) becomes tight. □



(a) $\beta = 0.50956$ (best)

(b) $\beta = 0.3$

(c) $\beta = 0.7$

Fig. 1. Cost functions; the SAA cost cannot upper bound the true cost. (a): when $\beta = 0.50956$, the BDR cost function provides a tight upper bound for the true cost function; (b): when $\beta < 0.50956$, the BDR cost function cannot upper bound the true cost function; (c): when $\beta > 0.50956$, the BDR cost function provides a loose upper bound for the true cost function. If the feasible region of the decision variable x is required to be $[-1.7, 1.7]$ rather than $\mathbb{R}$, the BDR bound in (a) is no longer tight but that in (b) becomes tight. (Source Codes: https://github.com/Spratm-Asleaf/Robustness-Specificity.)

As a result of Theorem 2, focusing on the Bayesian distributionally robust solution $\hat{\boldsymbol{x}}_{b,n}$, the true cost $v(\hat{\boldsymbol{x}}_{b,n})$ of the BDR model is upper bounded, with $\mathbb{P}_0^n$-probability at least $1-\eta$, as $v(\hat{\boldsymbol{x}}_{b,n}) \le \beta_n v_{r,n}(\hat{\boldsymbol{x}}_{b,n}) + (1-\beta_n) v_n(\hat{\boldsymbol{x}}_{b,n})$.

Next, we discuss the unbiasedness of the BDR model. The DRO model is always an upward (i.e., positively) biased estimator of the true optimal cost for all radius $\epsilon_n \ge 0$ such that $\mathbb{P}_0 \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)$, while the SAA model is always a downward (i.e., negatively) biased estimator.[3] However, the BDR model can be unbiased with a proper β_n.

Theorem 3 (Unbiasedness)**.** *For every n, there exists $\beta_n \in [0, 1]$ such that the BDR-estimated cost $v_{b,n}(\hat{\boldsymbol{x}}_{b,n})$ is an unbiased estimate of the true optimal cost $v(\boldsymbol{x}_0)$.*

Proof (sketch). We first show that the DRO model is an upward (positively) biased model and the SAA model is a downward (negatively) biased model. Then, the BDR model is proved to be unbiased. For details, see Appendix D-D in the supplementary materials. □

The BDR model's unbiasedness indicates that achieving asymptotic statistical property is possible in finite-sample learning; note that this result is theoretically impossible for DRO and SAA models. However, a β_n satisfying Theorem 2 [i.e., (12)] does not necessarily satisfy Theorem 3 for unbiasedness, and vice versa. The finite-sample unbiasedness shows the statistical superiority of BDR over SAA and DRO.

V. Solution Method of BDR Model (10)

To solve BDR model (10), the key is to reformulate the DRO sub-problem $\max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi)$ under a specified distributional ball $B_{\epsilon_n}(\hat{\mathbb{P}}_n)$. This paper examines the ϕ-divergence and Wasserstein distributional balls; see Appendix A-B.

A. ϕ-Divergence

We start with the ϕ-divergence ball whose mathematical definition is available in Appendix A-B1; this case is practical if the underlying true data-generating distribution $\mathbb{P}_0$ is discrete.

Theorem 4. *Consider the ϕ-divergence distributional ball $B_{\epsilon_n,\phi}(\hat{\mathbb{P}}_n)$ induced by the ϕ-divergence. The DRO sub-problem $\max_{\mathbb{P} \in B_{\epsilon_n,\phi}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi)$ can be reformulated to*

$$\max_{\boldsymbol{\mu} \in \mathbb{R}^n} \sum_{i=1}^n \mu_i \cdot h(\boldsymbol{x}, \xi_i), \qquad \textit{s.t.} \qquad F_\phi(\boldsymbol{\mu} \| \bar{\boldsymbol{\mu}}) \le \epsilon_n, \tag{14}$$

where $F_\phi(\boldsymbol{\mu} \| \bar{\boldsymbol{\mu}})$ defines the ϕ-divergence of the discrete distribution $\boldsymbol{\mu} := [\mu_1, \mu_2, \ldots, \mu_n]$ from the nominal distribution $\bar{\boldsymbol{\mu}} := [1/n, 1/n, \ldots, 1/n]$; note that $\boldsymbol{\mu}, \bar{\boldsymbol{\mu}} \in \mathbb{R}^n$.

Proof. From Appendix A-B1, we know that distributions $\mathbb{P}$ in $B_{\epsilon_n,\phi}(\hat{\mathbb{P}}_n)$ have the same support as $\hat{\mathbb{P}}_n$. Hence, distributions in $B_{\epsilon_n,\phi}(\hat{\mathbb{P}}_n)$ can be characterized as $\mathbb{P} = \sum_{i=1}^n \mu_i \delta_{\xi_i}$, which completes the proof. □

A concrete example of the constraint in (14) can be obtained using the Kullback–Leibler (KL) divergence: that is,

$$F_\phi(\boldsymbol{\mu} \| \bar{\boldsymbol{\mu}}) := \sum_{i=1}^n \mu_i \cdot \log(\mu_i / \bar{\mu}_i) = \sum_{i=1}^n \mu_i \cdot \log(n\mu_i) \le \epsilon_n.$$

As a result, the solution of BDR method (10) is given in the corollary below.

Corollary 1 (Solution of BDR Method (10) Under ϕ-Divergence Ball)**.** *The BDR model* (10) *under the ϕ-divergence ball can be reformulated into*

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathcal{X}} \quad & \beta_n \max_{\boldsymbol{\mu} \in \mathbb{R}^n} \sum_{i=1}^n \mu_i \cdot h(\boldsymbol{x}, \xi_i) + (1-\beta_n) \sum_{i=1}^n \frac{1}{n} \cdot h(\boldsymbol{x}, \xi_i) \\ \textit{s.t.} \quad & F_\phi(\boldsymbol{\mu} \| \bar{\boldsymbol{\mu}}) \le \epsilon_n, \end{aligned} \tag{15}$$

which is a finite-dimensional optimization. □

B. Wasserstein Distance

We then study the Wasserstein distributional ball whose mathematical definition is available in Appendix A-B2.

Theorem 5. *Consider the Wasserstein distributional ball $B_{\epsilon_n,p}(\hat{\mathbb{P}}_n)$ induced by the order-p Wasserstein distance. Suppose one of the following conditions holds: 1) For every $\boldsymbol{x}$, $h(\boldsymbol{x}, \xi)$ is continuous in ξ on Ξ; 2) For every $\boldsymbol{x}$, $h(\boldsymbol{x}, \xi)$ is concave in ξ on Ξ. Then, the DRO sub-problem $\max_{\mathbb{P} \in B_{\epsilon_n,p}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi)$ can be reformulated to*

$$\max_{\{\zeta_j\}_{j \in [n]}} \frac{1}{n} \sum_{j=1}^n h(\boldsymbol{x}, \zeta_j), \qquad \textit{s.t.} \quad \frac{1}{n} \sum_{j=1}^n d^p(\xi_j, \zeta_j) \le \epsilon_n^p, \tag{16}$$

[3] For technical details, see the proof of Theorem 3.

where d is a distance on Ξ.

Proof. See Appendix E in the supplementary materials. □

A concrete example of the constraint in (16) can be obtained using the 2-norm on Ξ and $p := 1$, that is,

$$\tfrac{1}{n}\sum_{j=1}^{n} \|\xi_j - \zeta_j\|_2 \leq \epsilon_n.$$

As a result, the solution of the BDR method (10) is given in the corollary below.

Corollary 2 (Solution of BDR Method (10) Under Wasserstein Ball)**.** *The BDR model* (10) *under the Wasserstein ball can be reformulated into*

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathcal{X}} \quad & \beta_n \max_{\{\zeta_j\}_{j\in[n]}} \frac{1}{n}\sum_{j=1}^{n} h(\boldsymbol{x},\zeta_j) + (1-\beta_n)\sum_{i=1}^{n}\frac{1}{n}\cdot h(\boldsymbol{x},\xi_i) \\ \textit{s.t.} \quad & \frac{1}{n}\sum_{j=1}^{n} d^p(\xi_j,\zeta_j) \leq \epsilon_n^p, \end{aligned} \tag{17}$$

which is a finite-dimensional optimization. □

C. Numerical Solution

The algorithm below, adapted from stochastic gradient descent (SGD) [30], provides a numerically iterative method to solve (15) and (17) for gradient-based learning (e.g., neural networks).

Algorithm 1 (BDR-GD to Solve (15) and (17))**.** *With probability β_n we use the gradient of the DRO term $\max_{\boldsymbol{\mu}}\sum_{i=1}^{n}\mu_i\cdot h(\boldsymbol{x},\xi_i)$ or $\max_{\{\zeta_j\}_{j\in[n]}}\frac{1}{n}\sum_{j=1}^{n}h(\boldsymbol{x},\zeta_j)$, and with probability $1-\beta_n$ we use the gradient of the SAA term $\frac{1}{n}\sum_{i=1}^{n}h(\boldsymbol{x},\xi_i)$. For example, in the t-th iteration step, ξ_t is sampled from $\hat{\mathbb{P}}_n$ and p_t is sampled from the uniform distribution $\mathbb{U}_{(0,1]}$. Then the stochastic gradient, with respect to $\boldsymbol{x}$,*

$$\boldsymbol{g}_{\boldsymbol{x},t} = \begin{cases} \nabla_{\boldsymbol{x}} h(\boldsymbol{x},\xi_t), & \beta_n \leq p_t, \\ \nabla_{\boldsymbol{x}} \max_{\zeta_t} h(\boldsymbol{x},\zeta_t) \ \textit{s.t.}\ d^p(\xi_t,\zeta_t) < \epsilon^p, & \beta_n > p_t, \end{cases}$$

is calculated to update the hypothesis parameter $\boldsymbol{x}$. □

D. Hyper-Parameter Tuning

As demonstrated by the statistical properties in Theorem 2, the generalization performance of BDR learning is significantly influenced by the value of the hyper-parameter β_n. However, as highlighted in Remark 4, the optimal value β_n^* for β_n cannot be theoretically determined due to its dependence on the unknown true distribution $\mathbb{P}_0$. Therefore, in practice, β_n can be empirically tuned using, e.g., grid search, cross-validation, and bootstrapping. This is a common practice of hyperparameter searching in, e.g., regularized SAA learning (4) and DRO learning (5). Experiments in Section VII show that it is computationally lightweight to find some β_n such that BDR can outperform both DRO and SAA.

VI. Practical Insights from BDR Learning

Suppose that $\boldsymbol{\mu}^*$ solves (15) and $\{\zeta_j^*\}_{j\in[n]}$ solves (17). Corollaries 1 and 2 motivate two important insights in Examples 3 and 4, respectively.

Example 3 (Sample Weight Modification)**.** *In ERM learning* (2)*, we work on equal-weighted n samples $\{\xi_i\}_{i\in[n]}$, while in DRO learning $\min_{\boldsymbol{x}\in\mathcal{X}}\max_{\mathbb{P}\in B_{\epsilon_n,\phi}(\hat{\mathbb{P}}_n)}\mathbb{E}_{\mathbb{P}}h(\boldsymbol{x},\xi)$ with the ϕ-divergence ball, the weights of samples $\{\xi_i\}_{i\in[n]}$ are modified into $\boldsymbol{\mu}^*$. However, in BDR learning* (15)*, the weight of ξ_i is given by $\beta_n\mu_i^* + (1-\beta_n)/n$.* □

An application of Example 3 is "hard sample mining" [31], where β_n balances worst-case weight μ_i^* and homogeneous weight $1/n$ for sample ξ_i.

Example 4 (Data Augmentation)**.** *ERM learning* (2) *works on equal-weighted n nominal samples $\{\xi_i\}_{i\in[n]}$, while DRO learning $\min_{\boldsymbol{x}\in\mathcal{X}}\max_{\mathbb{P}\in B_{\epsilon_n,p}(\hat{\mathbb{P}}_n)}\mathbb{E}_{\mathbb{P}}h(\boldsymbol{x},\xi)$ with the Wasserstein ball constructs equal-weighted n adversarial samples $\{\zeta_j^*\}_{j\in[n]}$. In contrast, BDR learning* (17) *leverages $2n$ samples $\{\zeta_j^*\}_{j\in[n]}\cup\{\xi_i\}_{i\in[n]}$ with weight β_n/n for adversarial samples $\{\zeta_j^*\}_{j\in[n]}$ and weight $(1-\beta_n)/n$ for nominal samples $\{\xi_i\}_{i\in[n]}$: it enables data augmentation by combining DRO-generated adversarial samples and the nominal samples in SAA.* □

In robust deep learning, DRO-based adversarial training is widely used but infamous for its poor performance due to conservatism [32]. BDR learning, however, can mitigate this issue by incorporating SAA learning, which is shown by experiments in Subsection VII-B.

VII. Applications and Experiments

We show the practical benefits of the BDR learning framework through experimental results on real-world tasks such as 2D image and 3D point cloud classifications. Support vector machines and deep neural networks are specifically leveraged. All the source codes are available online at GitHub: https://github.com/Spratm-Asleaf/Robustness-Specificity.

A. Linear Model: BDR Support Vector Machine

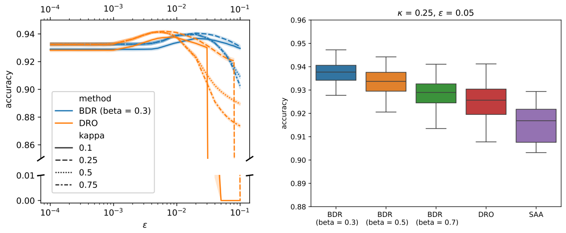
We consider the binary classification problem on MNIST dataset [33] to distinguish similar handwritten digits 4 and 9. We adopt the support vector machine (SVM) as the classification algorithm and solve the problem under the frameworks of BDR, DRO, and SAA, respectively. Denote the i-th image's pixel vector as $\boldsymbol{I}_i \in \mathbb{R}^{784}$ and its label as $Y_i \in \{-1, 1\}$, i.e., $\xi_i := (\boldsymbol{I}_i, Y_i)$. We choose the order-1 Wasserstein distance to define a distributional ball under the metric [2]

$$d(\xi_i,\xi_j) := \|\boldsymbol{I}_i - \boldsymbol{I}_j\|_\infty + \kappa\cdot\mathbb{1}_{\{Y_i\neq Y_j\}}, \tag{18}$$

where $\|\cdot\|_\infty$ denotes the ∞-norm and κ quantifies the cost of reversing a label. Hinge loss is used in SVM, i.e.

$$h(\boldsymbol{x},\xi) = h(\boldsymbol{x},(\boldsymbol{I},Y)) := \max\{1 - Y\cdot\langle\boldsymbol{x},\boldsymbol{I}\rangle,\ 0\}.$$

It can be derived from (17) and [2, Cor. 15] that the BDR formulation is a linear program; see Appendix C-A1 for technical details. We conduct 100 independent trials, in each of which, 80% of the images are randomly selected to train the model and the remaining 20% images are used for testing. For BDR, we choose β from $\{0.3, 0.5, 0.7\}$. For BDR and DRO, radius ϵ is chosen from $\{a \times 10^{-b} \mid a = 1, \cdots, 9, b = 4, 3, 2\}$ and κ is chosen from $\{0.1, 0.25, 0.5, 0.75\}$. The results are shown in Fig. 2.



(a) Accuracy against ϵ & κ

(b) Box plot of accuracy

Fig. 2. Average test accuracy for 4 vs 9 over 100 trials. Averaged CPU times (seconds): BDR = 68, DRO = 66, and SAA = 7.

It can be seen in Fig. 2(a) that the performances of BDR and DRO are significantly affected by ϵ and κ. The test accuracy first increases when ϵ increases but drops afterwards; the peak occurs in the range of $\epsilon \in [0.005, 0.05]$. This phenomenon agrees with our claim that the radii of the ambiguity sets can neither be too large nor too small: If the ambiguity sets are too small, robust methods cannot provide sufficient robustness; however, if the ambiguity sets are too large, robust methods are too conservative. Among different κ, $\kappa = 0.25$ works best for both BDR and DRO. Fig. 2(b) shows an accuracy comparison among BDR (with different β), DRO, and SAA, under $\kappa = 0.25$ and $\epsilon = 0.05$ as selected above, where BDR with $\beta = 0.3$ has higher accuracy compared to that with $\beta = 0.5$ and $\beta = 0.7$. Fig. 2(a) also supports our claim that BDR is less conservative than DRO—To be specific, DRO is sensitive to the choice of ϵ because a slight change of ϵ can lead to a large change in accuracy (especially around $\epsilon = 0.07$); in contrast, BDR is more robust to the choice of ϵ.

For more experimental results of BDR SVM on MNIST and UCI data sets, as well as running times, see Appendix C-A.

B. Deep Learning Model: BDR Learning

We present an implementation of deep BDR learning (DBDRL) and demonstrate the potential of our BDR model in enhancing the performance of deep models on various tasks.

Tasks: We apply the proposed BDR model to 2D image classification tasks using MNIST [33], CIFAR-10, and CIFAR-100 [34] datasets, as well as 3D point cloud classification utilizing ModelNet40 [35] dataset. To evaluate the generalization capacity of our method, we perform experiments under a low-shot data setting; that is, the model is learned on a subset of the training dataset. This setup means that a learning model yielding higher testing performance on a small training dataset has a better generalization capability.

Implementation: We consider the objective of DBDRL as presented in (17). Specifically, we employ the convex cross-entropy loss [36] as the function h for our learning. Additionally, we implement the BDR-GD in Algorithm 1 for DBDRL. The DRO term in BDR-GD is actualized through Adversarial Training (AT) techniques, with the employment of a specific Projected Gradient Descent (PGD) method [37] to perform the maximization and construct adversarial samples. For PGD implementation, we use the order-2 distance for the constraints; using notations in (18), an example is given by

$$d\left(\xi_i, \xi_j\right) = \left\|\boldsymbol{I}_i - \boldsymbol{I}_j\right\|_2 .$$

We follow the official implementation to train our models in both 2D and 3D tasks except for the low-shot data setting and BDR-GD utilization. Further details, such as the parameters of training and PGD, are put in Appendix C-B.

Results of MINIST: We implement the WideResNet-28 (WRN) [38] for our 2D experiments. We first demonstrate the capability of DBDRL with varying β values on the MNIST dataset. As depicted in Fig. 3, the best β^*, which is an estimation of β_n^* in Theorem 2, diminishes as the volume of training data escalates, corroborating the property of (7). Moreover, we observe that the best BDR models consistently outperform both their DRO and SAA counterparts; the advantage of BDR is especially obvious with smaller training data set. This is consistent with the theoretical analyses in Section IV.

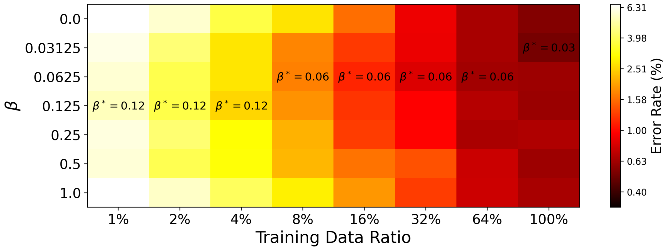


Fig. 3. Error rate of models trained by partial training sets on MNIST test set. Various β values are used during training: $\beta = 0$ for SAA learning, $\beta = 1$ for DRO learning, and β^* indicating the best value among various β for BDR learning.

Parameter Tuning: To obtain β^* during training, we adopt a validation-based search strategy: we leverage a subset of the training dataset (20% in our setting) as a validation set to search for a decent β. We highlight that the search cost is not significantly high, as it is found that a low precision of estimation can still enhance performance in practice. In later experiments, we restrict our search of β^* to a smaller set, *i.e.*, $\{0.5, 0.1, 0.05, 0.01\}$, and employ early stopping techniques to expedite the search process.

Main Results of 2D and 3D Classification: With the same tuning strategy for β_n, we showcase the superiority of our methods on CIFAR datasets in Table II. The used model is WRN-28 which is the same as MNIST experiments. We also employ DBDRL in 3D point cloud classification by implementing two models, PointNet [39] and DGCNN [40]. We utilize the above search method of β^* and demonstrate the consistent best performances of our BDR methods in Table III. Notably, DBDRL can improve the model performance from both the SAA learning and DRO learning across all tasks.

Additionally, we note that β^* estimation may not be accurate, as our search is limited to only a small set $\{0.01, 0.05, 0.1, 0.5\}$. However, high estimation accuracy of β^* is not critical in practice because, as depicted in Fig. 4, a wide range of β values can make BDR outperform the DRO and SAA. Overall, it is computationally lightweight to search for decent βs that enable BDR to outperform DRO and SAA. To illustrate this, we provide a detailed complexity analysis in Appendix C-B4 to show that the above search process can be done with trivial effort while achieving better performance.

TABLE II
ACCURACY (%) OF IMAGE CLASSIFICATION ON CIFAR-10 & CIFAR-100 UNDER LOW-SHOT DATA (10% OR 50% TRAINING DATA) SETTING.

Method	CIFAR-10		CIFAR-100	
	10%	50%	10%	50%
DRO	64.9	86.3	26.2	61.3
SAA	63.5	87.0	24.1	61.6
BDR	**66.5**	**87.3**	**26.9**	**63.4**
(β^*)	(0.05)	(0.05)	(0.1)	(0.05)

TABLE III
ACCURACY (%) OF POINT CLOUD CLASSIFICATION ON MODELNET40 BY DIFFERENT LEARNING METHODS. DIFFERENT TRAINING DATA RATIOS ARE UTILIZED. THE ESTIMATED β^* FOR EACH BDR LEARNING IS ALSO GIVEN.

Model	Data ratio	Method			β^*
		DRO	SAA	BDR	
PointNet	5%	72.9	72.3	**72.9**	0.5
	10%	79.6	79.4	**80.6**	0.1
	100%	88.7	89.1	**89.8**	0.05
DGCNN	5%	78.4	77.1	**79.86**	0.1
	10%	85.1	84.3	**85.8**	0.1
	100%	91.9	92.1	**92.8**	0.01

VIII. CONCLUSIONS

This paper proposes the Bayesian distributionally robust learning framework (9) or (10) that generalizes the Bayesian method, distributionally robust optimization method, and regularization method; see Remark 1. The new framework reveals that there exists a trade-off between the robustness to the distributional uncertainty and the specificity to the empirical information; see Remark 2. The new framework also suggests the design methods of the prior distribution $\mathbb{Q}$ in the Bayesian method (3) and the regularizer $f(\boldsymbol{x})$ in the regularization method (4) (see Remark 1), and shows that BDR learning can be less conservative than DRO learning (see Theorem 2, Remark 4, Examples 1 and 2, and Figs. 2, 3, and 4). The asymptotic (i.e., consistencies and asymptotic normalities in Theorem 1) and non-asymptotic (i.e., generalization bounds in Theorem 2 and unbiasedness in Theorem 3) properties, and the solution method (i.e., Corollaries 1 and 2) of the new framework are studied. In addition, the BDR learning framework reveals important insights from the perspective of data augmentation; see Examples 3 and 4. Experiments on diverse real-world datasets demonstrate the practical usefulness of the proposed BDR model.

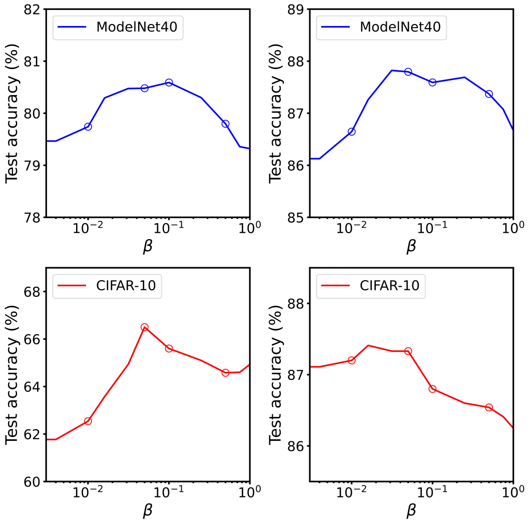


Fig. 4. Test set accuracy *v.s.* β across various tasks. Upper Panel: PointNet on ModelNet40 with 10% (left) and 50% (right) training data. Lower Panel: WRN-18 on CIFAR-10 with 10% (left) and 50% (right) training data. The marker "○" stands for searching set of β: i.e., $\{0.01, 0.05, 0.1, 0.5\}$. (NB: $\beta = 0$ for SAA learning, $\beta = 1$ for DRO learning.)

The future research direction is to study alternatives for the Dirichlet-process priors for the second-order probability measure $\mathbb{Q}$ in the Bayesian model (3), which possibly motivates other new robust learning models than the proposed BDR models in (9) and (10). Possible replacements are Dirichlet-process mixture priors [4, Chap. 5], tail-free process priors [4, Sec. 3.6], among many others.

APPENDIX A
APPENDICES OF SECTION II

A. Notations

Notations used in this paper are summarized in Table IV.

B. Similarity Measures of Distributions and Distributional Balls

1) ϕ-Divergence: Suppose $\mathbb{P}$ is absolutely continuous with respect to $\bar{\mathbb{P}}$. Let $\phi : \mathbb{R}_+ \to \{\mathbb{R} \cup +\infty\}$ denote a convex function that satisfies $\phi(1) = 0$ and $0\phi(0/0) = 0$. The ϕ-divergence (i.e., f-divergence) of $\mathbb{P}$ from $\bar{\mathbb{P}}$, generated by ϕ, is defined as

$$F_\phi(\mathbb{P}\|\bar{\mathbb{P}}) = \int_\Xi \phi\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\bar{\mathbb{P}}}\right)\bar{\mathbb{P}}(\mathrm{d}\xi), \tag{19}$$

where $\mathrm{d}\mathbb{P}/\mathrm{d}\bar{\mathbb{P}}$ is the Radon–Nikodym derivative of $\mathbb{P}$ with respect to $\bar{\mathbb{P}}$. When $\phi(t) := t\ln t$ for all $t > 0$, the ϕ-divergence specifies the well-known Kullback–Leibler divergence; cf. [41, Table 2].

A ϕ-divergence distributional ball with radius $\epsilon \geq 0$ and center $\bar{\mathbb{P}}$ is defined as

$$B_{\epsilon,\phi}(\bar{\mathbb{P}}) := \{\mathbb{P} \in \mathcal{M}(\Xi) | F_\phi(\mathbb{P}\|\bar{\mathbb{P}}) \leq \epsilon\}.$$

If $\epsilon = 0$, the ball $B_{\epsilon,\phi}(\bar{\mathbb{P}})$ reduces to the singleton that contains only $\bar{\mathbb{P}}$. In some literature, the ball is also defined as $B_{\epsilon,\phi}(\bar{\mathbb{P}}) :=$

TABLE IV
FULL NOTATION LIST

Symbol	**Interpretation**
$\mathcal{M}(\Xi)$	all distributions on $(\Xi, \mathcal{B}_\Xi)$ where $\mathcal{B}_\Xi$ is the Borel σ-algebra on Ξ
$\mathcal{B}_{\mathcal{M}(\Xi)}$	Borel σ-algebra on $\mathcal{M}(\Xi)$
$\mathbb{P}_0$	true population distribution
$\hat{\mathbb{P}}_n$	empirical distribution supported on n i.i.d. samples
$\hat{\mathbb{P}}$	a prior estimate of $\mathbb{P}_0$ based on prior knowledge
$\bar{\mathbb{P}}$	reference distribution working as a proper estimate of $\mathbb{P}_0$, which can be the empirical $\hat{\mathbb{P}}_n$ or the prior $\hat{\mathbb{P}}$, among many others
$\mathbb{P}_0^n$	n-fold product measure induced by $\mathbb{P}_0$ (i.e., joint distribution of n i.i.d. samples)
$[n]$	$[n] := \{1, 2, \ldots, n\}$, the running index set
$\Delta(\mathbb{P}, \hat{\mathbb{P}}_n)$	statistical similarity measure between $\mathbb{P}$ and $\hat{\mathbb{P}}_n$; Δ can be any possible divergences or statistical distances
$B_\epsilon(\hat{\mathbb{P}}_n)$	$:= \{\mathbb{P} \in \mathcal{M}(\Xi) \| \Delta(\mathbb{P}, \hat{\mathbb{P}}_n) \le \epsilon\}$, closed distributional ball with radius ϵ and center $\hat{\mathbb{P}}_n$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$\xrightarrow{a.s.}$	converges almost surely
$\xrightarrow{p}$	converges in probability
$\xrightarrow{d}$	converges in distribution
$o_p(1)$	if a sequence $a_n = o_p(1)$, then a_n converges to zero in probability
$d(\boldsymbol{x}, \boldsymbol{y})$	distance between two points $\boldsymbol{x}$ and $\boldsymbol{y}$
$d(\boldsymbol{x}, \mathcal{X})$	$:= \inf_{\boldsymbol{y} \in \mathcal{X}} \\|\boldsymbol{x} - \boldsymbol{y}\\|$, distance between the point $\boldsymbol{x}$ and the set $\mathcal{X}$
$d(\mathcal{X}, \mathcal{Y})$	$:= \sup_{\boldsymbol{x} \in \mathcal{X}} d(\boldsymbol{x}, \mathcal{Y})$, distance between the two sets $\mathcal{X}$ and $\mathcal{Y}$
$\mathbb{E}_{\mathbb{P}}[\cdot], \mathbb{D}_{\mathbb{P}}[\cdot]$	the expectation operator and the covariance operator, respectively, with respect to the distribution $\mathbb{P}$
$\nabla_{\boldsymbol{x}} h(\boldsymbol{x}_0, \xi)$	the gradient, i.e., Jacobian, of $h(\boldsymbol{x}, \xi)$ with respect to $\boldsymbol{x}$ evaluated at $\boldsymbol{x}_0$
$\nabla^2_{\boldsymbol{x}} h(\boldsymbol{x}_0, \xi)$	the second-order gradient, i.e., Hessian, of $h(\boldsymbol{x}, \xi)$ with respect to $\boldsymbol{x}$ evaluated at $\boldsymbol{x}_0$
$\boldsymbol{V}^{-\top} := [\boldsymbol{V}^{-1}]^\top$	the transpose of the inverse of the matrix $\boldsymbol{V}$
$[\boldsymbol{a}, \boldsymbol{b}]$ and $[\boldsymbol{a}; \boldsymbol{b}]$	MATLAB notation for row and column concatenation of $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively

$\{\mathbb{P} \in \mathcal{M}(\Xi) | F_\phi(\bar{\mathbb{P}} \| \mathbb{P}) \le \epsilon\}$ where $\mathbb{P}$ and $\bar{\mathbb{P}}$ are swapped. The two versions are not equivalent because the ϕ-divergence is not guaranteed to be symmetric in general.

2) Wasserstein Distance: The order-p Wasserstein distance between the two distributions $\mathbb{P}$ and $\bar{\mathbb{P}}$ is defined as

$$\begin{aligned} W_p(\mathbb{P}, \bar{\mathbb{P}}) &= \left[\inf_{\pi \in \mathcal{M}(\Xi \times \Xi)} \mathbb{E}_\pi d^p(\xi_1, \xi_2)\right]^{\frac{1}{p}} \\ &= \left[\inf_{\pi \in \mathcal{M}(\Xi \times \Xi)} \int_{\Xi \times \Xi} d^p(\xi_1, \xi_2) \pi(\mathrm{d}\xi_1, \mathrm{d}\xi_2)\right]^{\frac{1}{p}}, \end{aligned} \tag{20}$$

where d is a distance on Ξ, $p \ge 1$, and π is a joint distribution on $\Xi \times \Xi$ with marginals $\mathbb{P}$ and $\bar{\mathbb{P}}$.

An order-p Wasserstein distributional ball with radius ϵ and center $\bar{\mathbb{P}}$ is defined as

$$B_{\epsilon,p}(\bar{\mathbb{P}}) := \{\mathbb{P} \in \mathcal{M}(\Xi) | W_p(\mathbb{P}, \bar{\mathbb{P}}) \le \epsilon\}.$$

If $\epsilon = 0$, the ball $B_{\epsilon,p}(\bar{\mathbb{P}})$ reduces to the singleton that contains only $\bar{\mathbb{P}}$.

Wasserstein balls admit the following concentration properties. Suppose the true population distribution $\mathbb{P}_0$ has a light tail: That is, there exist $\alpha > p \ge 1$ (but $p \ne k/2$) and finite $A > 0$ such that $\mathbb{E}_{\mathbb{P}_0}[\exp(\|\xi\|^\alpha)] \le A$ (recall that k is the dimension of ξ). Then, there exist constants $c_1, c_2 > 0$ such that

$$\mathbb{P}_0^n \left[\mathbb{P}_0 \in B_{\epsilon_n,p}(\hat{\mathbb{P}}_n)\right] \ge 1 - \eta \tag{21}$$

holds, for any $\eta \in (0, 1]$, when

$$\epsilon_n \ge \begin{cases} \left(\frac{\log(c_1/\eta)}{c_2 n}\right)^{\min\{1/k, 1/2\}} & \text{if } n \ge \frac{\log(c_1/\eta)}{c_2}, \\ \left(\frac{\log(c_1/\eta)}{c_2 n}\right)^{1/\alpha} & \text{if } n < \frac{\log(c_1/\eta)}{c_2}. \end{cases} \tag{22}$$

Note that c_1 and c_2 are determined by α, A, and k. This result is attributed to [1, Thm. 18]. The difficulty of applying this result in practice is that the involved constants α and A cannot be exactly obtained because the population distribution $\mathbb{P}_0$ is unknown, and so are c_1 and c_2.

When the support set Ξ is finite and bounded (i.e., $\mathbb{P}_0$ is discrete), there exist concentration properties of $\hat{\mathbb{P}}_n$ with respect to the Wasserstein distance that do not depend on unknown constants; see, e.g., [29, pp. 42].

C. Wasserstein DRO Models

1) Existence of The Solution of Wasserstein DRO Models: Suppose (Ξ, d) is a proper,[4] complete, and separable metric space, $h(\boldsymbol{x}, \xi)$ is upper semi-continuous in ξ on Ξ and $\mathbb{E}_{\bar{\mathbb{P}}} |h(\boldsymbol{x}, \xi)| < \infty$ for every $\boldsymbol{x}$, and $\bar{\mathbb{P}}$ has a finite p-th moment: That is, for every $\xi_0 \in \Xi$, we have $\int_\Xi d^p(\xi, \xi_0) \bar{\mathbb{P}}(\mathrm{d}\xi) < \infty$. Then, for every $\boldsymbol{x}$, the optimal value of the Wasserstein DRO problem

$$\max_{\mathbb{P}: W_p(\mathbb{P}, \bar{\mathbb{P}}) \le \epsilon} \int_\Xi h(\boldsymbol{x}, \xi) \mathbb{P}(\mathrm{d}\xi) \tag{23}$$

is finite if and only if there exist $\xi_0 \in \Xi$ and $c_1(\boldsymbol{x}) > 0$ such that

$$h(\boldsymbol{x}, \xi) \le c_1(\boldsymbol{x})[1 + d^p(\xi, \xi_0)], \quad \forall \xi \in \Xi. \tag{24}$$

In addition, the optimal value is attainable (by one $\mathbb{P}^*$ such that $W_p(\mathbb{P}^*, \bar{\mathbb{P}}) \le \epsilon$) if there exist $\xi_0 \in \Xi$, $c_1(\boldsymbol{x}) > 0$, and $c_2 \in (0, p)$ such that

$$h(\boldsymbol{x}, \xi) \le c_1(\boldsymbol{x})[1 + d^{c_2}(\xi, \xi_0)], \quad \forall \xi \in \Xi. \tag{25}$$

The results above can be seen in, e.g., [17], [29]. Note that (24) is in analogy to the Lipschitz continuity which limits the "change rate" of a function. To clarify further, for example, by letting $p := 1$ and $d := \|\cdot\|$ (i.e., the metric d is induced by a norm $\|\cdot\|$), we can see that (24) is in analogy to $h(\boldsymbol{x}, \xi) \le h(\boldsymbol{x}, \xi_0) + L(\boldsymbol{x}) \|\xi - \xi_0\|$, for every $\xi, \xi_0 \in \Xi$, where $L(\boldsymbol{x}) > 0$ is the Lipschitz constant. For this reason, in literature, e.g.,

[4] A metric space (Ξ, d) is proper if for any $\epsilon > 0$ and $\xi_0 \in \Xi$, the closed ϵ-ball $B_\epsilon(\xi_0) := \{\xi \in \Xi | d(\xi, \xi_0) \le \epsilon\}$, is compact.

[29], [42], (24) is called the "finite-growth-rate" condition for the function h.

In this paper, for practicality, we consistently assume that the condition (25) is satisfied so that it is safe to replace the supremum with the maximum in the DRO model.

2) Reformulation of Wasserstein DRO Models: According to, e.g., [24, Thm. 1] and [42, Thm. 1],[5] the Wasserstein DRO problem (23) is equivalent to its Lagrangian dual:[6]

$$\min_{\lambda\geq 0}\left\{\lambda\epsilon^p + \int_{\Xi}\max_{\xi\in\Xi}\left\{h(\boldsymbol{x},\xi) - \lambda\cdot d^p(\xi,\bar{\xi})\right\}\bar{\mathbb{P}}(\mathrm{d}\bar{\xi})\right\}. \quad (26)$$

If $\bar{\mathbb{P}} = \sum_{i=1}^{n}\bar{\mu}_i\delta_{\xi_i}$ is a discrete distribution, e.g., an empirical distribution, supported on n points $\{\xi_i\}_{i\in[n]}$, then (26) becomes

$$\min_{\lambda\geq 0}\left\{\lambda\epsilon^p + \sum_{i=1}^{n}\bar{\mu}_i\max_{\xi\in\Xi}\left\{h(\boldsymbol{x},\xi) - \lambda\cdot d^p(\xi,\xi_i)\right\}\right\}. \quad (27)$$

3) Support Set of Worst-Case Distributions: If $\bar{\mathbb{P}}$ is supported on n points in Ξ, then the worst-case distribution solving (23) is supported on at most $n+1$ points in Ξ; see [17, Thm. 4], [42, Cor. 2].

Special cases when h is concave or piece-wise linear in ξ or when $\bar{\mathbb{P}} := \hat{\mathbb{P}}_n$ are discussed in, e.g., [18], [2], [1], [43].

D. Glivenko–Cantelli Class, Donsker Class, and Brownian Bridge

Consider a function class $\mathcal{F} := \{f : \Xi \to \mathbb{R}\}$.

Definition 2 (Glivenko–Cantelli Class)**.** *Suppose for every $f \in \mathcal{F}$, $\mathbb{E}_{\mathbb{P}_0}f(\xi)$ is defined[7] and finite; that is, f is $\mathbb{P}_0$-integrable. The function class $\mathcal{F}$ is called $\mathbb{P}_0$-Glivenko–Cantelli if*

$$\sup_{f\in\mathcal{F}}|\mathbb{E}_{\hat{\mathbb{P}}_n}f(\xi) - \mathbb{E}_{\mathbb{P}_0}f(\xi)| \xrightarrow{a.s.} 0. \quad (28)$$

Intuitively, if $\mathcal{F}$ is a Glivenko–Cantelli class, then the uniform strong law of large numbers holds on $\mathcal{F}$. □

Definition 3 (Donsker Class)**.** *Consider an empirical process*

$$\mathbb{G}_n(f) := \sqrt{n}[\mathbb{E}_{\hat{\mathbb{P}}_n}f(\xi) - \mathbb{E}_{\mathbb{P}_0}f(\xi)], \quad \forall f\in\mathcal{F} \quad (29)$$

indexed by the function class $\mathcal{F}$. That is, $\{\mathbb{G}_n(f)|f\in\mathcal{F}\}$ in (29) *is a stochastic process indexed by $\mathcal{F}$; the randomness comes from the (random) empirical measure $\hat{\mathbb{P}}_n$. Suppose for every $f\in\mathcal{F}$, $\mathbb{D}_{\mathbb{P}_0}f(\xi)$ is defined and finite; that is, f is $\mathbb{P}_0$-square-integrable. The function class $\mathcal{F}$ is called $\mathbb{P}_0$-Donsker if the empirical (stochastic) process $\mathbb{G}_n$ converges in distribution to a Brownian bridge (stochastic) process:*

$$\mathbb{G}_n \xrightarrow{d} \mathbb{G}_{\mathbb{P}_0}, \quad (30)$$

where $\mathbb{G}_{\mathbb{P}_0}$ is a zero-mean $\mathbb{P}_0$-Brownian bridge on $\mathcal{F}$ with uniformly continuous sample paths with respect to the semi-metric $\sqrt{\mathbb{D}_{\mathbb{P}_0}[f_1(\xi) - f_2(\xi)]}$ between $f_1\in\mathcal{F}$ and $f_2\in\mathcal{F}$; in addition, $\mathbb{G}_{\mathbb{P}_0}(f)$ is tight for every $f\in\mathcal{F}$; i.e., $\sup_{f\in\mathcal{F}}|\mathbb{G}_{\mathbb{P}_0}(f)| < \infty$ in $\mathbb{P}_0$-probability. Intuitively, if $\mathcal{F}$ is a Donsker class, then the uniform central limit theorem[8] holds on $\mathcal{F}$. □

A zero-mean $\mathbb{P}_0$-***Brownian bridge*** $\mathbb{G}_{\mathbb{P}_0}$ on $\mathcal{F}$ is a Gaussian process on $\mathcal{F}$ satisfying the following two conditions:

1) For every $f\in\mathcal{F}$, $\mathbb{G}_{\mathbb{P}_0}(f)$ is a random variable with mean of zero and variance of $\mathbb{D}_{\mathbb{P}_0}(f)$.
2) For every integer r and every possible collection of functions $\{f_1, f_2, \ldots, f_r\}$ taken from $\mathcal{F}$, the random vector $[\mathbb{G}_{\mathbb{P}_0}(f_1), \mathbb{G}_{\mathbb{P}_0}(f_2), \ldots, \mathbb{G}_{\mathbb{P}_0}(f_r)]^\top$follows a r-dimensional multivariate Gaussian distribution with covariance between $\mathbb{G}_{\mathbb{P}_0}(f_i)$ and $\mathbb{G}_{\mathbb{P}_0}(f_j)$ being defined as $\mathbb{E}_{\mathbb{P}_0}f_i\cdot f_j - \mathbb{E}_{\mathbb{P}_0}f_i\cdot\mathbb{E}_{\mathbb{P}_0}f_j$, for every $i,j\in[r]$.

Since the values of the Gaussian process $\mathbb{G}_{\mathbb{P}_0}$ at some functions $f\in\mathcal{F}$ are strictly zeros, without any randomness, the Gaussian process $\mathbb{G}_{\mathbb{P}_0}$ is called a Brownian bridge because some values are tied, for example, when f is $\mathbb{P}_0$-almost everywhere constant.

Appendix B
Examples Satisfying The Conditions in Theorem 1

The conditions C1)-C6) in Theorem 1 are not practically restrictive as they are standard for the DRO model (5) [1], [18], [17], [20] and the SAA model (2) [44], [26, Chap. 19], [45, Chap. 5]. The only new requirement is Condition C2); i.e., $\sqrt{n}\beta_n \to 0$, which is also mild. Some specific situations where the conditions C1)-C6) in Theorem 1 hold are given below.

Condition C1) holds if, for example, (25) is satisfied;

Condition C2) holds if, for example, $\beta_n := \frac{\alpha}{n+\alpha}$, for every n, where $\alpha \geq 0$ is a constant;[9]

Condition C3) holds if, for example, one of the following is satisfied:

a) The function class $\mathcal{H}$ is finite and every element of $\mathcal{H}$ is $\mathbb{P}_0$-integrable;
b) The parameter space $\mathcal{X}$ is bounded, every element of $\mathcal{H}$ is $\mathbb{P}_0$-integrable, and there exists a $\mathbb{P}_0$-integrable function $m(\xi)$ such that
$$|h(\boldsymbol{x}_1,\xi) - h(\boldsymbol{x}_2,\xi)| \leq m(\xi)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \quad \forall\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{X}, \quad (31)$$
is satisfied $\mathbb{P}_0$-almost surely.
c) The parameter space $\mathcal{X}$ is compact, every element of $\mathcal{H}$ is $\mathbb{P}_0$-integrable, every element $\boldsymbol{x}\mapsto h(\boldsymbol{x},\xi)$ in $\mathcal{H}$ is continuous on $\mathcal{X}$ $\mathbb{P}_0$-almost-surely, and there exists a $\mathbb{P}_0$-integrable envelop $m(\xi)$ such that
$$\sup_{\boldsymbol{x}\in\mathcal{X}}|h(\boldsymbol{x},\xi)| \leq m(\xi) \quad (32)$$
is satisfied $\mathbb{P}_0$-almost surely.
d) Every element in $\mathcal{H}$ is a finite linear combination of other $\mathbb{P}_0$-integrable functions; that is,
$$\mathcal{H} := \left\{\sum_{i=1}^{l}x_i f_i(\xi)\,\middle|\, \boldsymbol{x}\in\mathcal{X}\subseteq\mathbb{R}^l,\ \mathbb{E}_{\mathbb{P}_0}f_i(\xi) < \infty\right\}. \quad (33)$$

[5] The finite growth-rate assumption for the function h in [42] is equivalent to require (24); see Lemma 2 therein.

[6] λ is the dual variable for the constraint in (23).

[7] At least one of the positive part and the negative part of f has finite integral.

[8] The uniform central limit theorem is also known as the functional central limit theorem as a random function(al) sequence (i.e., the empirical process) converges to a random function(al) (i.e., a Brownian bridge).

[9] Recall from (6) that this rule is used in the Dirichlet process prior for a Bayesian non-parametric model.

This type of $\mathcal{H}$ is popular in machine learning, for example, when the hypothesis class $\mathcal{H}$ is a well-designed reproducing kernel Hilbert space.

e) The function class $\mathcal{H}$ is a Vapnik–Chervonenkis (VC) class; that is, the VC index of $\mathcal{H}$ is finite. For example, the function class in (33) is a VC class.

Condition C4) holds if, for example, one of the following is satisfied:

a) For any $\boldsymbol{x} \in \mathcal{X}'$, if the function $h(\boldsymbol{x}, \xi)$ is continuous at $\boldsymbol{x}$, $\mathbb{P}_0$-almost surely, and the function $h(\boldsymbol{x}_n, \xi)$ is dominated by a $\mathbb{P}_0$-integrable envelop function $m(\xi)$ for every n, then $v(\boldsymbol{x})$ is continuous on $\mathcal{X}'$. This is by the dominated convergence theorem.

b) The fact that $\boldsymbol{x}_n \to \boldsymbol{x}$ implies $v(\boldsymbol{x}_n) \to v(\boldsymbol{x})$, for every $\boldsymbol{x}_n, \boldsymbol{x} \in \mathcal{X}'$. This means that $v(\boldsymbol{x})$ is continuous on $\mathcal{X}'$.

c) The function $h(\boldsymbol{x}, \xi)$ is convex in $\boldsymbol{x}$, $\mathbb{P}_0$-almost surely, so that $v(\boldsymbol{x})$ is convex and therefore continuous in the interior of $\mathcal{X}'$. Note that the convexity of $v(\boldsymbol{x})$ on $\mathcal{X}'$ implies its continuity in the interior of $\mathcal{X}'$.

d) The function $h(\boldsymbol{x}, \xi)$ is strictly convex (resp. strongly convex) in $\boldsymbol{x}$, $\mathbb{P}_0$-almost surely, so that $v(\boldsymbol{x})$ is strictly convex (resp. strongly convex). This means that $v(\boldsymbol{x})$ is has a unique global minimizer on $\mathcal{X}'$.

Condition C5) holds if, for example, one of the following is satisfied:

a) The function class $\mathcal{H}$ is finite and every element of $\mathcal{H}$ is $\mathbb{P}_0$-square-integrable;

b) The parameter space $\mathcal{X}$ is bounded, every element of $\mathcal{H}$ is $\mathbb{P}_0$-square-integrable, and there exists a $\mathbb{P}_0$-square-integrable function $m(\xi)$ such that

$$|h(\boldsymbol{x}_1, \xi) - h(\boldsymbol{x}_2, \xi)| \le m(\xi) \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X},$$

is satisfied $\mathbb{P}_0$-almost surely.

c) Every element in $\mathcal{H}$ is a finite linear combination of other $\mathbb{P}_0$-square-integrable functions; that is,

$$\mathcal{H} := \left\{ \sum_{i=1}^{l} x_i f_i(\xi) \,\middle|\, \boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^l, \ \mathbb{E}_{\mathbb{P}_0}[f_i(\xi)]^2 < \infty \right\}. \tag{34}$$

This type of $\mathcal{H}$ is popular in machine learning, for example, when the hypothesis class $\mathcal{H}$ is a well-designed reproducing kernel Hilbert space.

d) The function class $\mathcal{H}$ is a Vapnik–Chervonenkis (VC) class; that is, the VC index of $\mathcal{H}$ is finite.

Condition C6) holds if, for example, one of the following is satisfied:

a) There exists a $\mathbb{P}_0$-square-integrable function $m(\xi)$ such that

$$|h(\boldsymbol{x}_1, \xi) - h(\boldsymbol{x}_2, \xi)| \le m(\xi) \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X},$$

is satisfied $\mathbb{P}_0$-almost surely.

b) Every element in $\mathcal{H}$ is a finite linear combination of other $\mathbb{P}_0$-square-integrable functions; that is, (34). This is because $\mathbb{E}_{\mathbb{P}_0}[\sum_{i=1}^{l} (\hat{x}_{n,i} - x_{0,i}) f_i(\xi)]^2 \le \sum_{i=1}^{l} (\hat{x}_{n,i} - x_{0,i})^2 \cdot \sum_{i=1}^{l} \mathbb{E}_{\mathbb{P}_0}[f_i(\xi)]^2 \xrightarrow{p} 0$, as $\hat{\boldsymbol{x}}_n \xrightarrow{p} \boldsymbol{x}_0$.

Therefore, if we assume the pointwise $m(\xi)-$Lipschitz continuity of the function $h(\boldsymbol{x}, \xi)$ where $m(\xi)$ is $\mathbb{P}_0$-square-integrable, then Conditions C3-C6 in Theorem 1 are simultaneously satisfied. In addition, if $\mathcal{H}$ takes the form as in (34), then Conditions C3-C6 in Theorem 1 are simultaneously satisfied as well. This two situations are sufficient for most of practical machine learning hypothesis classes.

Appendix C
Appdices of Section VII

A. BDR Support Vector Machine

The BDR SVM classifier is derived in Appendix C-A1. Additional experimental results of the BDR SVM on the MNIST dataset are shown in Appendix C-A2. Experimental tests of the BDR SVM on the UCI datasets [46], i.e., the Ionosphere dataset, the Breast Cancer dataset, and the Adult dataset are reported in Appendices C-A3, C-A4, and C-A5, respectively. The tuning method of hyperparameters is the same as that used on the MNIST dataset; see Subsection VII-A in the main body of the paper. The average computational times (averaged over 100 independent trials) for the experiments are provided in Table V, which shows that BDR and DRO are computationally comparable.

TABLE V
Averaged CPU Times (Unit: Seconds)

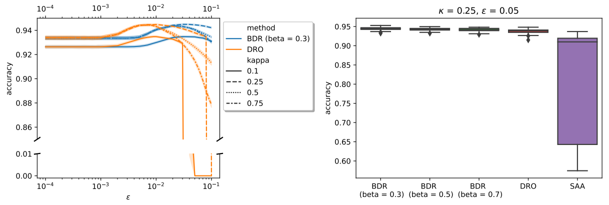
	BDR	DRO	SAA
MNIST (4 vs 9)	68	66	7
Ionosphere	4.0	3.9	0.3
Breast Cancer	4.2	3.5	0.4
Adult (a1a)	4.0	3.8	0.3

1) BDR Formulation of SVM: The BDR formulation of the SVM classification problem can be solved with a linear program

$$\begin{array}{lll}
\displaystyle\min_{\boldsymbol{x},\, \boldsymbol{\lambda},\, \boldsymbol{s}} & \displaystyle\beta_n \epsilon \lambda_0 + \frac{1}{n} \sum_{i=1}^{n} \lambda_i & \\
s.t. & 1 - Y_i \cdot \langle \boldsymbol{x}, \boldsymbol{I}_i \rangle \le \lambda_i, & \forall i \in [n], \\
& 1 + Y_i \cdot \langle \boldsymbol{x}, \boldsymbol{I}_i \rangle - \kappa \lambda_0 \le \lambda_i, & \forall i \in [n], \\
& 0 \le \lambda_i, & \forall i \in \{0\} \cup [n], \\
& \sum_{j=1}^{l} s_j \le \lambda_0, & \\
& x_j \le s_j, -x_j \le s_j, 0 \le s_j, & \forall j \in [l], \\
& \boldsymbol{x} \in \mathbb{R}^l, \ \boldsymbol{\lambda} \in \mathbb{R}^{n+1}, \ \boldsymbol{s} \in \mathbb{R}^l, &
\end{array} \tag{35}$$

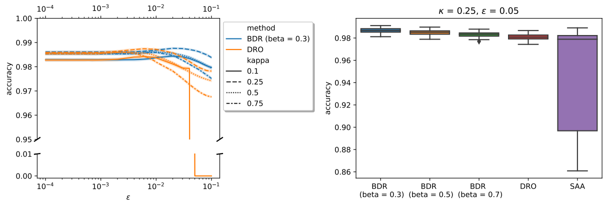
where n is the size of training samples and $\boldsymbol{\lambda} := (\lambda_0, \lambda_1, \ldots, \lambda_n)$. The derivation process is trivial and therefore omitted here. Just note that the dual norm of the ∞-norm is the 1-norm, and in [2, Eq. (19)] we have $\boldsymbol{C} = \boldsymbol{0}$ and $\boldsymbol{d} = \boldsymbol{0}$ (i.e., $\mathcal{X} := \mathbb{R}^l$). In this special case, BDR amounts to DRO, where BDR just employs a β_n-shrunken radius $\beta_n \epsilon$ for the distributional uncertainty ball. However, this motivational relation no longer holds for complicated learning tasks such as BDR deep learning.

2) Additional Experiments on The MNIST Dataset: Experimental results of the average out-of-sample accuracy on the MNIST dataset for 3 vs 8 over 100 independent trials are shown in Fig. 5, while for 1 vs 7 are in Fig. 6. From the two figures, we can see that the conclusions are consistent with those given in the main body of the paper (i.e., Subsection VII-A): For example, BDR is more robust than DRO to the choice of the radius ϵ of the distributional uncertainty ball.



(a) Mean accuracy against ϵ & κ (b) Box plot of accuracy

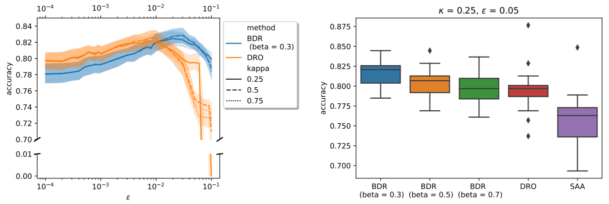
Fig. 5. Average out-of-sample accuracy on the MNIST dataset for 3 vs 8 over 100 independent trials.



(a) Mean accuracy against ϵ & κ (b) Box plot of accuracy

Fig. 6. Average out-of-sample accuracy on the MNIST dataset for 1 vs 7 over 100 independent trials.

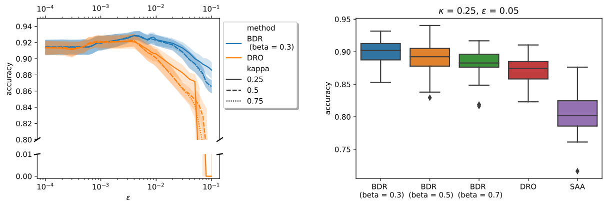
3) Experiments on The UCI Ionosphere Dataset: The results on the UCI Ionosphere dataset are shown in Fig. 7.



(a) Mean accuracy against ϵ & κ (b) Box plot of accuracy

Fig. 7. Average out-of-sample accuracy on the UCI Ionosphere dataset over 100 independent trials.

4) Experiments on The UCI Breast Cancer Dataset: The results on the UCI Breast Cancer dataset are shown in Fig. 8.



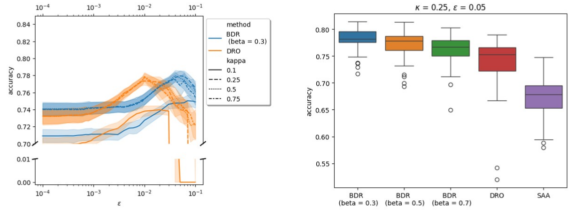
(a) Mean accuracy against ϵ & κ (b) Box plot of accuracy

Fig. 8. Average out-of-sample accuracy on the UCI Breast Cancer dataset over 100 independent trials.

5) Experiments on The UCI Adult Dataset: The results on the UCI Adult dataset are shown in Figs. 9-13.

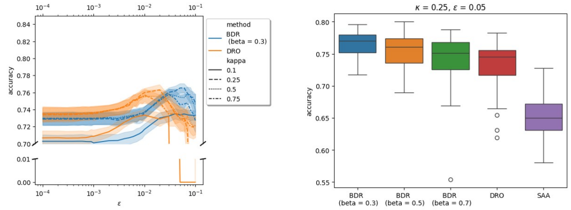
B. Deep BDR Learning

1) Dataset Overview: We provide the numerical details of the utilized datasets in Table VI.



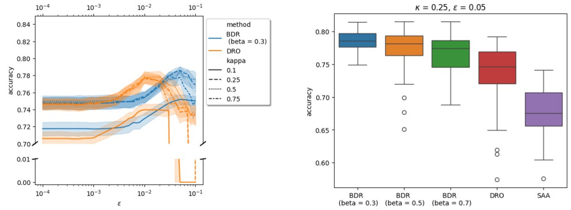
(a) Mean accuracy against ϵ & κ (b) Box plot of accuracy

Fig. 9. Average out-of-sample accuracy on the UCI Adult dataset (a1a) over 100 independent trials.



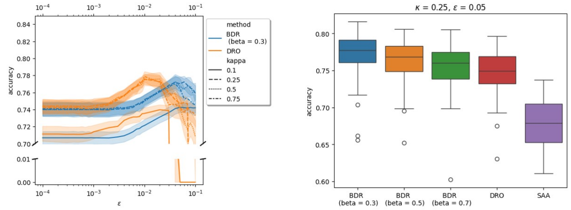
(a) Mean accuracy against ϵ & κ (b) Box plot of accuracy

Fig. 10. Average out-of-sample accuracy on the UCI Adult dataset (a2a) over 100 independent trials.



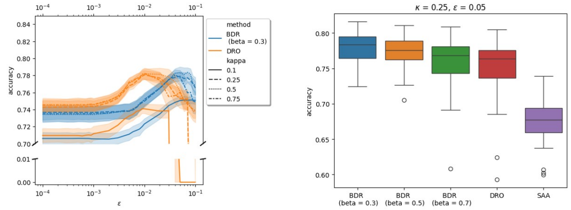
(a) Mean accuracy against ϵ & κ (b) Box plot of accuracy

Fig. 11. Average out-of-sample accuracy on the UCI Adult dataset (a3a) over 100 independent trials.



(a) Mean accuracy against ϵ & κ (b) Box plot of accuracy

Fig. 12. Average out-of-sample accuracy on the UCI Adult dataset (a4a) over 100 independent trials.



(a) Mean accuracy against ϵ & κ (b) Box plot of accuracy

Fig. 13. Average out-of-sample accuracy on the UCI Adult dataset (a5a) over 100 independent trials.

TABLE VI
SUMMARY OF DATASETS

Dataset Name	Train Data Size	Test Data Size	Categories
MNIST	60,000	10,000	10
CIFAR-10	50,000	10,000	10
CIFAR-100	50,000	10,000	100
ModelNet40	9,843	2,468	40

2) Training Details: All experiments are executed using Python 3.9, PyTorch 1.2, on a NVIDIA TITAN V GPU, ensuring a stable computing environment for deep learning tasks. We also note that in practice we implement the mini-batch version of Algorithm 1 which is just similar to the mini-batch SGD.

In 2D image classification, the WideResNet-28 model's training on CIFAR-10 and CIFAR-100 utilizes a batch size of 128 across 200 epochs. The learning rate is initially 0.1, adjusted down to 0.01 at epoch 100 and further to 0.001 at epoch 150. We employ SGD with momentum for optimization, setting weight decay at 0.0005. Furthermore, the training of model on MNIST uses Adam optimizer with learning rate 0.001 without decaying. There is no extra data augmentation strategy except for the DRO-based adversarial sample construction.

In 3D point cloud classification, we sample 1,024 points of the 2048-point data as the input. PointNet training setup includes a batch size of 32, up to 250 epochs, and an initial learning rate of 0.001, adjusted by a decay mechanism. The Adam optimizer is used for training. The learning rate's decay step is set to 200,000, with a decay rate of 0.7. On the other hand, DGCNN training specifies a batch size of 32, 250 epochs, and a learning rate of 0.1 with SGD (momentum 0.9). It includes a cosine annealing for adaptive learning rate adjustments. There is no extra data augmentation strategy except for the DRO-based adversarial sample construction.

3) Adversarial Training Details: In our study, the PGD attack [37] within a 2-norm ball is implemented to generate adversarial samples based on DRO cost, of which the parameters are carefully chosen to ensure an effective yet subtle modification of data.

In the image classification, the epsilon ϵ, defining the maximum perturbation limit per pixel, is set to 0.03, to maintain the visual similarity of the adversarial images to their originals. The step size α, determining the granularity of each update towards the adversarial direction, is chosen as 0.008. This fine-grained approach allows for precise control over the perturbation process. We iterate this process for 10 iterations to achieve a balance between perturbation invisibility and the success rate of the attack. In the point cloud classification, we set ϵ to 0.05, α to 0.01, and the iteration number to 7.

4) Search Complexity Analysis: Here we provide the quantitative analysis of the time complexity. Let t_{SAA}, t_{DRO} and t_{BDR} denote the training time per epoch for SAA, DRO and BDR method; t_{BDR} can be divided into two phases: β-searching phase by cross-validation and BDR training phase, formally,

$$t_{\text{BDR}} = t_{\text{Search}} + t_{\text{Train}}. \tag{36}$$

For a selected β, we have $t_{\text{Train}} = \beta t_{\text{DRO}} + (1-\beta)t_{\text{SAA}}$ due to Algorithm 1. Suppose the set of candidate β is $\{\beta_1, \cdots, \beta_k\}$, we implement the training on each β_i as

$$t_{\text{Search}} = r\sum_{i=1}^{k} t_{\text{Train with } \beta_i} = r\sum_{i=1}^{k}(\beta_i t_{\text{DRO}} + (1-\beta_i)t_{\text{SAA}}) \tag{37}$$

where $r \in (0, 1]$ is a factor standing for the effect of early stop for cross-validation. Thus, the total time when β^* is selected as the optimal one is

$$\begin{aligned} t_{\text{BDR}} &= r\textstyle\sum_{i=1}^{k} t_{\text{Train with } \beta_i} + t_{\text{Train}} \\ &= r\textstyle\sum_{i=1}^{k}(\beta_i t_{\text{DRO}} + (1-\beta_i)t_{\text{SAA}}) + \\ &\beta^* t_{\text{DRO}} + (1-\beta^*)t_{\text{SAA}} \end{aligned} \tag{38}$$

if we consider the upper bound of total time when $r = 1$, we have

$$t_{\text{BDR}} \leq (\max_i \beta_i + \sum_{i=1}^{k}\beta_i)t_{\text{DRO}} + (k+1-\sum_{i=1}^{k}\beta_i - \min_i \beta_i)t_{\text{SAA}}. \tag{39}$$

Considering the search set $\{0.5, 0.1, 0.05, 0.01\}$, we have

$$t_{\text{BDR}} \leq 1.16t_{\text{DRO}} + 3.83t_{\text{SAA}} = 1.6t_{\text{DRO}} \tag{40}$$

The upper bound of t_{BDR} is $1.6t_{\text{DRO}}$ when $t_{\text{DRO}} \backslash t_{\text{SAA}} = 9$ in practice. However, since the usage of early stop (i.e., $r < 1$) and $t_{\text{Train}} < \max_i \beta_i t_{\text{DRO}} + (1 - \min_i \beta_i)t_{\text{SAA}}$. In practice, we always get $t_{\text{BDR}} \approx t_{\text{DRO}}$, as shown in Table VII. Table VII provides the time used per epoch for DRO, SAA, and BDR learning for CIFAR-10 and 50% data experiment. Our searching time is the equivalent time used per epoch for all β validation training.

TABLE VII
AVERAGED TIME USED PER EPOCH FOR VARIOUS METHODS IN CIFAR-10 AND 50% DATA EXPERIMENT

Method	Time (s)
DRO	888.0 ± 7.3
SAA	101.1 ± 1.5
BDR (β=0.01)	111.2 ± 5.6
BDR (β=0.05)	143.3 ± 12.5
BDR (β=0.1)	184.4 ± 16.5
BDR (β=0.5)	539.3 ± 29.3
BDR search time	694.0

As the estimated β^* is 0.05, the total time of the BDR method is 837.3s, which is less than the DRO method. The rationale behind this is that DRO optimization often requires significantly more time than SAA. Algorithm 1, by integrating the two, effectively reduces the overall time required for DRO optimization even though we have to conduct the searching by cross-validation.

REFERENCES

[1] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations Research & Management Science in the Age of Analytics*. INFORMS, 2019, pp. 130–166.

[2] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Regularization via mass transportation," *Journal of Machine Learning Research*, vol. 20, no. 103, pp. 1–68, 2019.

[3] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, pp. 209–230, 1973.

[4] S. Ghosal and A. Van der Vaart, *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017, vol. 44.

[5] M. Gaudard and D. Hadwin, "Sigma-algebras on spaces of probability measures," *Scandinavian Journal of Statistics*, pp. 169–175, 1989.

[6] D. Wu, H. Zhu, and E. Zhou, "A Bayesian risk approach to data-driven stochastic optimization: Formulations and asymptotics," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1588–1612, 2018.

[7] E. Anderson and H. Nguyen, "When can we improve on sample average approximation for stochastic optimization?" *Operations Research Letters*, vol. 48, no. 5, pp. 566–572, 2020.

[8] V. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[10] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: Data mining, inference, and prediction," *Springer New York*, 2009.

[11] H. Zhang and S. Chen, "Concentration inequalities for statistical inference," *Communications in Mathematical Research*, vol. 37, no. 1, pp. 1–85, 2021.

[12] O. Bousquet and A. Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[13] P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien, "PAC-Bayesian theory meets Bayesian inference," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[14] H. Rahimian and S. Mehrotra, "Frameworks and results in distributionally robust optimization," *Open Journal of Mathematical Optimization*, vol. 3, pp. 1–85, 2022.

[15] D. Kuhn, S. Shafiee, and W. Wiesemann, "Distributionally robust optimization," *Acta Numerica*, Nov 2024.

[16] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford, "Large-scale methods for distributionally robust optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8847–8860, 2020.

[17] M.-C. Yue, D. Kuhn, and W. Wiesemann, "On linear optimization over Wasserstein balls," *Mathematical Programming*, pp. 1–16, 2021.

[18] P. Mohajerin Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, vol. 171, no. 1, pp. 115–166, 2018.

[19] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 707–738, 2015.

[20] R. Gao, "Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality," *Operations Research*, 2022.

[21] M. R. Chernick, *Bootstrap methods: A Guide for Practitioners and Researchers*. John Wiley & Sons, 2011.

[22] G. Varoquaux, "Cross-validation failure: Small sample sizes lead to large error bars," *Neuroimage*, vol. 180, pp. 68–77, 2018.

[23] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," *Journal of Applied Probability*, vol. 56, no. 3, pp. 830–857, 2019.

[24] J. Blanchet, K. Murthy, and V. A. Nguyen, "Statistical analysis of Wasserstein distributionally robust estimators," in *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*. INFORMS, 2021, pp. 227–254.

[25] S. Wang and H. Wang, "Distributional robustness bounds generalization errors," 2022. [Online]. Available: https://arxiv.org/abs/2212.09962

[26] A. W. Van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 1998.

[27] D. Z. Long, M. Sim, and M. Zhou, "Robust satisficing," *Operations Research*, vol. 71, no. 1, pp. 61–82, 2023.

[28] R. Gao, X. Chen, and A. J. Kleywegt, "Wasserstein distributionally robust optimization and variation regularization," *Operations Research*, 2022.

[29] R. Chen, I. C. Paschalidis *et al.*, "Distributionally robust learning," *Foundations and Trends® in Optimization*, vol. 4, no. 1-2, pp. 1–243, 2020.

[30] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, pp. 462–466, 1952.

[31] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769.

[32] A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang, "Adversarial training can hurt generalization," in *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019.

[33] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[34] A. Krizhevsky, "Learning multiple layers of features from tiny images," University of Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[35] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1912–1920.

[36] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning," *Annals of Data Science*, pp. 1–26, 2020.

[37] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[38] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

[39] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.

[40] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics*, vol. 38, no. 5, pp. 1–12, 2019.

[41] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science*, vol. 59, no. 2, pp. 341–357, 2013.

[42] R. Gao and A. Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," *Mathematics of Operations Research*, 2022.

[43] C. Zhao and Y. Guan, "Data-driven risk-averse stochastic optimization with Wasserstein metric," *Operations Research Letters*, vol. 46, no. 2, pp. 262–267, 2018.

[44] A. W. Van Der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. Springer, 1996.

[45] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[46] D. Dua and C. Graff, "UCI Machine Learning Repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

Supplementary Materials

APPENDIX D
APPENDICES OF SECTION IV

A. Proof of Theorem 1

Proof. For every $\boldsymbol{x} \in \mathcal{X}'$ such that the SAA and DRO objectives are bounded in $\mathbb{P}_0^n$-probability, we have

$$\begin{aligned} v_{b,n}(\boldsymbol{x}) - v_n(\boldsymbol{x}) \quad &= \beta_n \max\nolimits_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + \\ &\quad (1-\beta_n)\mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) - \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) \\ &\xrightarrow{p} 0, \end{aligned}$$

because $\beta_n \to 0$. Hence, by Slutsky's theorem, $v_{b,n}(\boldsymbol{x})$ shares the same asymptotic properties with $v_n(\boldsymbol{x})$, for every $\boldsymbol{x} \in \mathcal{X}'$. As a result, Statement S1) and S4) are immediate due to the conventional strong law of large numbers, i.e.,

$$v_n(\boldsymbol{x}) = \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) \xrightarrow{a.s.} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) = v(\boldsymbol{x}), \quad \forall x \in \mathcal{X}',$$

and the conventional central limit theorem

$$\sqrt{n}[\mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) - \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi)] \xrightarrow{d} N(0, \mathbb{D}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi)), \quad \forall x \in \mathcal{X}',$$

respectively.

Suppose the DRO sub-problem is solved by $\bar{\mathbb{P}}_n$ such that

$$\mathbb{E}_{\bar{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) = \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi),$$

for $\boldsymbol{x} \in \mathcal{X}'$. Note that this assumption is reasonable due to Condition C1). We have

$$\begin{aligned} &\sup_{\boldsymbol{x} \in \mathcal{X}'} |v_{b,n}(\boldsymbol{x}) - v(\boldsymbol{x})| \\ &\quad = \sup_{\boldsymbol{x} \in \mathcal{X}'} \left| \mathbb{E}_{\beta_n \bar{\mathbb{P}}_n + (1-\beta_n)\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) - \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) \right| \\ &\quad = \sup_{\boldsymbol{x} \in \mathcal{X}'} \left| \beta_n \left[\mathbb{E}_{\bar{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) - \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) \right] + \right. \\ &\qquad\quad \left. (1-\beta_n) \left[\mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) - \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) \right] \right| \\ &\quad \le \beta_n \sup_{\boldsymbol{x} \in \mathcal{X}'} \left| \mathbb{E}_{\bar{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) - \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) \right| + \\ &\qquad\quad (1-\beta_n) \sup\nolimits_{\boldsymbol{x} \in \mathcal{X}'} \left| \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) - \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) \right| \\ &\quad \xrightarrow{p} 0. \end{aligned}$$

The first term vanishes because β_n approaches zero and $\sup_{\boldsymbol{x} \in \mathcal{X}'} \left| \mathbb{E}_{\bar{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) - \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) \right|$ is finite on $\mathcal{X}'$, whereas the second term decays because $\mathcal{H}$ is $\mathbb{P}_0$-Glivenko–Cantelli. As a result, $\min_{\boldsymbol{x}} v_{b,n} \xrightarrow{p} \min_{\boldsymbol{x}} v(\boldsymbol{x})$, as $n \to \infty$, because

$$\begin{aligned} |v_{b,n}(\hat{\boldsymbol{x}}_{b,n}) - v(\boldsymbol{x}_0)| \quad &= |\min_{\boldsymbol{x} \in \mathcal{X}'} v_{b,n}(\boldsymbol{x}) - \min_{\boldsymbol{x} \in \mathcal{X}'} v(\boldsymbol{x})| \\ &\le \sup_{\boldsymbol{x} \in \mathcal{X}'} |v_{b,n}(\boldsymbol{x}) - v(\boldsymbol{x})| \\ &\xrightarrow{p} 0. \end{aligned}$$

This is Statement S2).

For every $\hat{\boldsymbol{x}}_{b,n} \in \hat{\mathcal{X}}_{b,n}$, we have

$$\begin{aligned} &|v(\hat{\boldsymbol{x}}_{b,n}) - \min_{\boldsymbol{x} \in \mathcal{X}'} v(\boldsymbol{x})| \\ &\quad \le |v(\hat{\boldsymbol{x}}_{b,n}) - v_{b,n}(\hat{\boldsymbol{x}}_{b,n})| + |v_{b,n}(\hat{\boldsymbol{x}}_{b,n}) - \min_{\boldsymbol{x} \in \mathcal{X}'} v(\boldsymbol{x})| \\ &\quad \le \sup_{\boldsymbol{x} \in \mathcal{X}'} |v_{b,n}(\boldsymbol{x}) - v(\boldsymbol{x})| + |v_{b,n}(\hat{\boldsymbol{x}}_{b,n}) - \min_{\boldsymbol{x} \in \mathcal{X}'} v(\boldsymbol{x})| \\ &\quad \xrightarrow{p} 0. \end{aligned}$$

Therefore, due to Condition C4), there exists $\boldsymbol{x}_0 \in \mathcal{X}_0$ such that $\hat{\boldsymbol{x}}_{b,n} \xrightarrow{p} \boldsymbol{x}_0$, which proves Statement S3). (One may use a contradiction, by assuming that the limit point of $\hat{\boldsymbol{x}}_{b,n}$ is not in $\mathcal{X}_0$, to verify this claim.)

By Conditions C5) and C6), we have $\mathbb{G}_n h(\hat{\boldsymbol{x}}_n, \xi) \xrightarrow{d} \mathbb{G}_{\mathbb{P}_0} h(\boldsymbol{x}_0, \xi) \sim N(0, \mathbb{D}_{\mathbb{P}_0} h(\boldsymbol{x}_0, \xi))$; see [26, Lemma 19.24]. On the one hand, we have

$$\begin{aligned} &\sqrt{n}[v_{b,n}(\hat{\boldsymbol{x}}_{b,n}) - v(\boldsymbol{x}_0)] \\ &\quad = \mathbb{E}_{\sqrt{n}[\beta_n \bar{\mathbb{P}}_n + (1-\beta_n)\hat{\mathbb{P}}_n]} h(\hat{\boldsymbol{x}}_{b,n}, \xi) - \mathbb{E}_{\sqrt{n}\mathbb{P}_0} h(\boldsymbol{x}_0, \xi) \\ &\quad \le \mathbb{E}_{\sqrt{n}[\beta_n \bar{\mathbb{P}}_n + (1-\beta_n)\hat{\mathbb{P}}_n]} h(\boldsymbol{x}_0, \xi) - \mathbb{E}_{\sqrt{n}\mathbb{P}_0} h(\boldsymbol{x}_0, \xi) \\ &\quad = \sqrt{n}[\mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}_0, \xi) - \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}_0, \xi)] + o_p(1) \\ &\quad \xrightarrow{d} \mathbb{G}_{\mathbb{P}_0} h(\boldsymbol{x}_0, \xi), \end{aligned}$$

where the second equality is because $\sqrt{n}\beta_n \to 0$ and $o_p(1)$ in the second equality denotes the "small-Oh" notation (i.e., $a_n = o_p(1)$ implies that the sequence $\{a_n\}$ converges in probability to zero as $n \to \infty$), and the convergence in distribution is due to the fact that $h(\boldsymbol{x}_0, \cdot)$ is in $\mathcal{H}$ and $\mathcal{H}$ is $\mathbb{P}_0$-Donsker. By Slutsky's theorem, it implies that

$$\begin{aligned} &\mathbb{E}_{\sqrt{n}[\beta_n \bar{\mathbb{P}}_n + (1-\beta_n)\hat{\mathbb{P}}_n]} h(\boldsymbol{x}_0, \xi) - \mathbb{E}_{\sqrt{n}\mathbb{P}_0} h(\boldsymbol{x}_0, \xi) \\ &\quad \xrightarrow{d} \mathbb{G}_{\mathbb{P}_0} h(\boldsymbol{x}_0, \xi). \end{aligned}$$

On the other hand, we have

$$\begin{aligned} &\sqrt{n}[v_{b,n}(\hat{\boldsymbol{x}}_{b,n}) - v(\boldsymbol{x}_0)] \\ &\quad = \mathbb{E}_{\sqrt{n}[\beta_n \bar{\mathbb{P}}_n + (1-\beta_n)\hat{\mathbb{P}}_n]} h(\hat{\boldsymbol{x}}_{b,n}, \xi) - \mathbb{E}_{\sqrt{n}\mathbb{P}_0} h(\boldsymbol{x}_0, \xi) \\ &\quad \ge \mathbb{E}_{\sqrt{n}[\beta_n \bar{\mathbb{P}}_n + (1-\beta_n)\hat{\mathbb{P}}_n]} h(\hat{\boldsymbol{x}}_{b,n}, \xi) - \mathbb{E}_{\sqrt{n}\mathbb{P}_0} h(\hat{\boldsymbol{x}}_{b,n}, \xi) \\ &\quad = \sqrt{n}[\mathbb{E}_{\hat{\mathbb{P}}_n} h(\hat{\boldsymbol{x}}_{b,n}, \xi) - \mathbb{E}_{\mathbb{P}_0} h(\hat{\boldsymbol{x}}_{b,n}, \xi)] + o_p(1) \\ &\quad \xrightarrow{d} \mathbb{G}_{\mathbb{P}_0} h(\hat{\boldsymbol{x}}_{b,n}, \xi) \\ &\quad = \mathbb{G}_{\mathbb{P}_0} h(\boldsymbol{x}_0, \xi) + o_p(1), \end{aligned}$$

where the second equality is because $\sqrt{n}\beta_n \to 0$, the convergence in distribution is due to the fact that $h(\hat{\boldsymbol{x}}_{b,n}, \cdot)$ is in $\mathcal{H}$ and $\mathcal{H}$ is $\mathbb{P}_0$-Donsker, and the third equality is because the function $\mathbb{G}_{\mathbb{P}_0} h(\cdot, \xi)$ is (uniformly) continuous[10] so that the continuous mapping theorem applies. By Slutsky's theorem, it implies that

$$\begin{aligned} &\mathbb{E}_{\sqrt{n}[\beta_n \bar{\mathbb{P}}_n + (1-\beta_n)\hat{\mathbb{P}}_n]} h(\hat{\boldsymbol{x}}_{b,n}, \xi) - \mathbb{E}_{\sqrt{n}\mathbb{P}_0} h(\hat{\boldsymbol{x}}_{b,n}, \xi) \\ &\quad \xrightarrow{d} \mathbb{G}_{\mathbb{P}_0} h(\boldsymbol{x}_0, \xi). \end{aligned}$$

Therefore, by the squeeze theorem, we have

$$\sqrt{n}[v_{b,n}(\hat{\boldsymbol{x}}_{b,n}) - v(\boldsymbol{x}_0)] \xrightarrow{d} \mathbb{G}_{\mathbb{P}_0} h(\boldsymbol{x}_0, \xi) \sim N(0, \mathbb{D}_{\mathbb{P}_0} h(\boldsymbol{x}_0, \xi)),$$

because the cumulative distribution function of $N(0, \mathbb{D}_{\mathbb{P}_0} h(\boldsymbol{x}_0, \xi))$ is continuous everywhere on $\mathbb{R}$. This completes the proof. □

B. Asymptotic Normality of the Optimal Solution

The asymptotic normality of the optimal solution is established below.

Proposition 1 (Asymptotic Normality of Optimal Solution). *For every $\hat{\boldsymbol{x}}_{b,n} \in \hat{\mathcal{X}}_{b,n}$ and every $\boldsymbol{x}_0 \in \mathcal{X}_0$, if Conditions C1) and C2) in Theorem 1 hold, $\hat{\boldsymbol{x}}_{b,n} \xrightarrow{p} \boldsymbol{x}_0$, the Jacobian $\nabla_{\boldsymbol{x}} h(\boldsymbol{x}_0, \xi)$ exists and is $\mathbb{P}_0$-square-integrable such that*

$$|h(\boldsymbol{x}_1, \xi) - h(\boldsymbol{x}_2, \xi)| \le \nabla_{\boldsymbol{x}} h(\boldsymbol{x}_0, \xi) \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X},$$

and the Hessian $\nabla^2_{\boldsymbol{x}} h(\boldsymbol{x}_0, \xi)$ exists and is nonsingular and $\mathbb{P}_0$-integrable, then we have $\sqrt{n}(\hat{\boldsymbol{x}}_{b,n} - \boldsymbol{x}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{x}_0})$ as $n \to \infty$, where

$$\begin{aligned} \boldsymbol{V}_{\boldsymbol{x}_0} := {} & [\mathbb{E}_{\mathbb{P}_0} \nabla^2_{\boldsymbol{x}} h(\boldsymbol{x}_0, \xi)]^{-1} \cdot \\ & \mathbb{E}_{\mathbb{P}_0}[\nabla_{\boldsymbol{x}} h(\boldsymbol{x}_0, \xi) \nabla^\top_{\boldsymbol{x}} h(\boldsymbol{x}_0, \xi)] \cdot [\mathbb{E}_{\mathbb{P}_0} \nabla^2_{\boldsymbol{x}} h(\boldsymbol{x}_0, \xi)]^{-\top}. \end{aligned}$$ □

[10] Almost all sample paths $f \mapsto \mathbb{G}_{\mathbb{P}_0}(f), \forall f \in \mathcal{F}$ of the $\mathbb{P}_0$-Brownian bridge process $\mathbb{G}_{\mathbb{P}_0}$ are uniformly continuous on the semi-metric space $(\mathcal{F}, d)$ where d is a semi-metric on $\mathcal{F}$; see [26, Lemma 18.15].

Proof. The proof is routine in light of proofs of [26, Thm. 5.23] and Theorem 1, and thus, omitted. Just note that a $\mathbb{P}_0$-square-integrable function is bounded in $\mathbb{P}_0$-probability. □

C. *Proof of Theorem 2*

Proof. For every given $\boldsymbol{x}$, if $v_n(\boldsymbol{x}) \geq v(\boldsymbol{x})$, the first inequality holds for all $\beta_{n,\boldsymbol{x}} \in [0,1]$ because $v_{r,n}(\boldsymbol{x}) \geq v(\boldsymbol{x})$ and $v_{r,n}(\boldsymbol{x}) \geq v_n(\boldsymbol{x})$; note that $\beta_{n,\boldsymbol{x}}$ depends on $\boldsymbol{x}$; if $v_n(\boldsymbol{x}) < v(\boldsymbol{x})$, the first inequality holds for some $\beta_{n,\boldsymbol{x}} \in [0,1]$. Therefore, for every $\boldsymbol{x}$, there exists $\beta_{n,\boldsymbol{x}} \in [0,1]$ such that the inequality

$$v(\boldsymbol{x}) \leq \beta_{n,\boldsymbol{x}} v_{r,n}(\boldsymbol{x}) + (1-\beta_{n,\boldsymbol{x}}) v_n(\boldsymbol{x})$$

holds $\boldsymbol{x}$-point-wisely. Let $\beta^*_{n,\boldsymbol{x}}$ denote the smallest value of $\beta_{n,\boldsymbol{x}}$ that satisfies the above display. Because $v_n(\boldsymbol{x}) \leq v_{r,n}(\boldsymbol{x})$, by letting $\beta_n \geq \beta^*_n := \max_{\boldsymbol{x}} \beta^*_{n,\boldsymbol{x}}$, the inequality

$$v(\boldsymbol{x}) \leq \beta_n v_{r,n}(\boldsymbol{x}) + (1-\beta_n) v_n(\boldsymbol{x})$$

holds uniformly for all $\boldsymbol{x}$; note that

$$\beta^*_{n,\boldsymbol{x}} v_{r,n}(\boldsymbol{x}) + (1-\beta^*_{n,\boldsymbol{x}}) v_n(\boldsymbol{x}) \leq \beta_n v_{r,n}(\boldsymbol{x}) + (1-\beta_n) v_n(\boldsymbol{x}).$$

Since

$$\beta^*_{n,\boldsymbol{x}} = \frac{v(\boldsymbol{x}) - v_n(\boldsymbol{x})}{v_{r,n}(\boldsymbol{x}) - v_n(\boldsymbol{x})}, \quad \forall \boldsymbol{x},$$

β^*_n equals the largest value of the right-hand side of the above display. This completes the proof. □

D. *Proof of Theorem 3*

Proof. For the DRO problem, if $\mathbb{P}_0 \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)$, as is the case in (21), we have

$$\begin{aligned} \min_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) &\leq \mathbb{E}_{\mathbb{P}_0} h(\hat{\boldsymbol{x}}_{r,n}, \xi) \\ &\leq \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\hat{\boldsymbol{x}}_{r,n}, \xi) \\ &= \min_{\boldsymbol{x}} \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi). \end{aligned}$$

The above display implies that $\min_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) \leq \mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi)\right]$. Therefore, the DRO model $\min_{\boldsymbol{x}} \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi)$ is always a positively biased estimator of $\min_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi)$, for every n such that $\mathbb{P}_0 \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)$. On the other hand, $\mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right] \leq \min_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi)$, that is, the SAA model $\min_{\boldsymbol{x}} \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)$ is always a negatively biased estimator of $\min_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi)$, for every n.

As for the BDR model, we have

$$\begin{aligned} &\min_{\boldsymbol{x}} \left[\beta_n \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + (1-\beta_n) \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right] \\ &\leq \beta_n \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + (1-\beta_n) \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) \\ &\leq \beta_n \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + (1-\beta_n) \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) \\ &= \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi), \end{aligned}$$

and therefore,

$$\begin{aligned} &\min_{\boldsymbol{x}} \left[\beta_n \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + (1-\beta_n) \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right] \\ &\leq \min_{\boldsymbol{x}} \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi). \end{aligned}$$

The above implies that, for every n, the BDR model gives a smaller estimate than the DRO model. (Since the DRO model is always positively biased, this is a desired property of the BDR model.) Furthermore, this means that

$$\begin{aligned} &\mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \left[\beta_n \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + (1-\beta_n) \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right]\right] \\ &\leq \mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi)\right], \end{aligned}$$

that is, the BDR model tends to have a smaller bias than the DRO model. In addition,

$$\begin{aligned} &\min_{\boldsymbol{x}} \left[\beta_n \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + (1-\beta_n) \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right] \\ &\geq \min_{\boldsymbol{x}} \left[\beta_n \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi) + (1-\beta_n) \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right] \\ &= \min_{\boldsymbol{x}} \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi). \end{aligned}$$

Hence, for every n, the BDR model gives a larger estimate than the SAA model. (Since the SAA model is always negatively biased, this is also a desired property of the BDR model.) Furthermore, this means that

$$\begin{aligned} &\mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \left[\beta_n \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + (1-\beta_n) \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right]\right] \\ &\geq \mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right], \end{aligned}$$

that is, the BDR model tends to have a smaller bias than the SAA model.

Since for every $\boldsymbol{x}$,

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right] &\leq \min_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) \\ &\leq \mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi)\right], \end{aligned}$$

there exists $\overline{\beta}_n \in [0,1]$ such that

$$\begin{aligned} \min_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) &= \overline{\beta}_n \cdot \mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi)\right] + \\ &\qquad (1-\overline{\beta}_n) \cdot \mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right] \\ &\leq \mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \left[\overline{\beta}_n \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + \right.\right. \\ &\qquad \left.\left. (1-\overline{\beta}_n) \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right]\right]. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} &\mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \left[0 \cdot \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + (1-0) \cdot \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right]\right] \\ &= \mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right] \\ &\leq \min_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi). \end{aligned}$$

Therefore, there exists $\beta_n \in [0, \overline{\beta}_n]$ such that

$$\begin{aligned} &\min_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi) = \\ &\mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \left[\beta_n \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + (1-\beta_n) \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right]\right], \end{aligned}$$

that is, the BDR model is an unbiased estimator of $\min_{\boldsymbol{x}} \mathbb{E}_{\mathbb{P}_0} h(\boldsymbol{x}, \xi)$, because the function

$$\beta \mapsto \mathbb{E}_{\mathbb{P}_0^n}\left[\min_{\boldsymbol{x}} \left[\beta \cdot \max_{\mathbb{P} \in B_{\epsilon_n}(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{P}} h(\boldsymbol{x}, \xi) + (1-\beta) \cdot \mathbb{E}_{\hat{\mathbb{P}}_n} h(\boldsymbol{x}, \xi)\right]\right]$$

is increasing and continuous in $\beta \in [0,1]$. □

Appendix E
Appendices of Section V: Proof of Theorem 5

Before we provide the formal proof of Theorem 5 in Appendix E-C, we prepare with preliminary results in Appendices E-A∼E-B. The key is to reformulate the infinite-dimensional DRO sub-problem into a finite-dimensional optimization.

A. Monte–Carlo Approximation

In the literature, the DRO problem

$$\begin{array}{lll}\min\limits_{\boldsymbol{x}\in\mathcal{X}} \max\limits_{\mathbb{P}} & \mathbb{E}_{\mathbb{P}}h(\boldsymbol{x},\xi) \\ s.t. & \Delta(\mathbb{P},\bar{\mathbb{P}}) \le \epsilon\end{array} \tag{41}$$

can be reformulated to a non-linear finite-dimensional optimization. For details, see Appendix A-C.

In this subsection, we propose to use a novel Monte–Carlo-based method to solve (41). Suppose $\mathbb{P} \approx \sum_{j=1}^{m} \mu_j \delta_{\zeta_j}$, where $\{\zeta_j\}_{j\in[m]}$ are samples from $\mathbb{P}$, δ_{ζ_j} is the Dirac measure at ζ_j, and the weights $\{\mu_j\}_{j\in[m]}$ can be determined by, e.g., importance sampling through using an appropriate proposal distribution (e.g., uniform distribution).[11] Likewise, we suppose that the set of observations $\{\xi_i\}_{i\in[n]}$ are sampled from $\bar{\mathbb{P}}$ and their weights are $\{\bar{\mu}_i\}_{i\in[n]}$ and therefore $\bar{\mathbb{P}} \approx \sum_{i=1}^{n} \bar{\mu}_i \delta_{\xi_i}$. As a result, all integrals in (41), i.e., $\mathbb{E}_{\mathbb{P}}h(\boldsymbol{x},\xi)$ and those involved in Δ if any,[12] can be approximated by weighted sums; the approximations are exact in the weak convergence sense (i.e., sums converge to integrals) if $\min\{n, m\} \to \infty$ due to the law of large numbers. In practice, we may choose large enough values for n and m, which however depends on specific problems. As a result, (41) transforms to

$$\begin{array}{lll}\min\limits_{\boldsymbol{x}} \max\limits_{\{\mu_j,\zeta_j\}_{j\in[m]}} & \sum_{j=1}^{m} \mu_j h(\boldsymbol{x},\zeta_j) \\ s.t. & \Delta(\mathbb{P},\bar{\mathbb{P}}) \le \epsilon.\end{array} \tag{42}$$

When Δ is the Wasserstein distance, (42) transforms to

$$\begin{array}{lll}\min\limits_{\boldsymbol{x}} \max\limits_{\boldsymbol{P},\boldsymbol{\mu},\{\zeta_j\}} & \sum_{j=1}^{m} \mu_j h(\boldsymbol{x},\zeta_j) & \\ \text{s.t.} & \sum_{i=1}^{n}\sum_{j=1}^{m} d^p(\xi_i,\zeta_j)\cdot P_{ij} \le \epsilon^p & \\ & \sum_{i=1}^{n} P_{ij} = \mu_j, & \forall j \in [m] \\ & \sum_{j=1}^{m} P_{ij} = \bar{\mu}_i, & \forall i \in [n] \\ & P_{ij} \ge 0, & \forall i \in [n], \forall j \in [m],\end{array} \tag{43}$$

where $\boldsymbol{P} := \{P_{ij}\}, \forall i \in [n], \forall j \in [m]$ can be seen as a joint distribution whose marginals are $\boldsymbol{\mu}$ and $\bar{\boldsymbol{\mu}}$, respectively. By eliminating $\boldsymbol{\mu}$, (43) is equivalent to

$$\begin{array}{lll}\min\limits_{\boldsymbol{x}} \max\limits_{\boldsymbol{P},\{\zeta_j\}} & \sum_{j=1}^{m}\sum_{i=1}^{n} h(\boldsymbol{x},\zeta_j)\cdot P_{ij} & \\ s.t. & \sum_{j=1}^{m}\sum_{i=1}^{n} d^p(\xi_i,\zeta_j)\cdot P_{ij} \le \epsilon^p & \\ & \sum_{j=1}^{m} P_{ij} = \bar{\mu}_i, & \forall i \in [n] \\ & P_{ij} \ge 0, & \forall i \in [n], \forall j \in [m].\end{array} \tag{44}$$

The worst-case distribution solving (44) is given below.

Theorem 6. *For every given $\boldsymbol{x}$, the worst-case distribution $\mathbb{P}^*$ solving* (44) *is supported on at most $n+1$ points in Ξ, that is, there exist $\{\mu_j,\zeta_j\}_{j\in[n+1]}$ such that $\mathbb{P}^* = \sum_{j=1}^{n+1} \mu_j \delta_{\zeta_j}$. Moreover, the discrete worst-case distribution $\mathbb{P}^*$ has the following structure*

$$\mathbb{P}^* = \bar{\mu}_{i_0} \cdot [q\delta_{\xi_{i_0,1}} + (1-q)\delta_{\xi_{i_0,2}}] + \sum_{j=1, j\ne i_0}^{n} \bar{\mu}_j \delta_{\zeta_j}, \tag{45}$$

for one $i_0 \in [n]$, where $0 \le q \le 1$ and $\{\zeta_j\}_{j\in[n+1]} = \{\xi_{i_{0,1}}\}\bigcup\{\xi_{i_{0,2}}\}\bigcup\{\xi_i\}_{i\in[n]-i_0}$. To be specific, at most one weight $\bar{\mu}_{i_0}$ of $\bar{\mathbb{P}}$ is split into two weights of $\mathbb{P}^*$ (N.B.: q is the splitting weight), and the other $n-1$ weights of $\bar{\mathbb{P}}$ (i.e., $\{\bar{\mu}_i\}_{i\in[n]-i_0}$) are directly inherited by $\mathbb{P}^*$.

Proof. See Appendix E-B. □

Theorem 6 implies that although $\mathbb{P}^*$ and $\bar{\mathbb{P}}$ have slightly different support sets, $\mathbb{P}^*$ is almost determined by the discrete reference distribution $\bar{\mathbb{P}}$.

Data-Driven Case: If $\bar{\mathbb{P}} := \hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}$ and $\bar{\mu}_j = 1/n, \forall j \in [n]$, (44) gives

$$\begin{array}{lll}\min\limits_{\boldsymbol{x}} \max\limits_{\boldsymbol{P},\{\zeta_j\}} & \displaystyle\sum_{j=1}^{m}\sum_{i=1}^{n} h(\boldsymbol{x},\zeta_j)\cdot P_{ij} & \\ s.t. & \sum_{j=1}^{m}\sum_{i=1}^{n} d^p(\xi_i,\zeta_j)\cdot P_{ij} \le \epsilon^p & \\ & \sum_{j=1}^{m} P_{ij} = \frac{1}{n}, & \forall i \in [n] \\ & P_{ij} \ge 0, & \forall i \in [n], \forall j \in [m].\end{array} \tag{46}$$

According to Theorem 6, when conducting the optimization (46), it is safe to let $m := n+1$. Note that the solution method (46) includes several existing duality-based methods as special cases, e.g., Corollary 3.3.1 in [29], Section 2.2 in [1].

B. Proof of Theorem 6

Proof. In (44), we have $n+1$ constraints, and therefore, at most $n+1$ components in $\boldsymbol{P}$ is non-zero. This further implies that, given $m \ge n+1$, at most $n+1$ components of $\boldsymbol{\mu}$ can be non-zero. In other words, the worst-case distribution solving (44) is supported on at most $n+1$ points for every $m \ge n+1$. The structure of $\mathbb{P}^*$ is straightforward to be verified by contradiction: If there exist two weights of $\bar{\mathbb{P}}$ to be split, then $\mathbb{P}^*$ needs to be supported on at least $n+2$ points, which contradicts the fact that $\mathbb{P}^*$ is supported on at most $n+1$ points. □

C. Proof of Theorem 5

Proof. For every $\boldsymbol{x}$, suppose $h(\boldsymbol{x},\xi)$ is continuous in ξ on Ξ. Then, for every $\{\mu_j\}_{j\in[m]}$ and $\{\zeta_j\}_{j\in[m]}$, there exists $\mu'_j = 1/m, \forall j \in [m]$ and $\{\xi'_j\}_{j\in[m]}$ such that

$$\sum_{j=1}^{m} \mu_j h(\boldsymbol{x},\zeta_j) = \sum_{j=1}^{m} \frac{1}{m} h(\boldsymbol{x},\xi'_j). \tag{47}$$

This is due to the intermediate value theorem of a continuous function. Hence, using the representation (47) of the weighted sum in the objective of (43), according to Theorem 6, we must have $m = n$ and $\mu_i = \bar{\mu}_i = 1/n$, for every $i \in [n]$. As a result, (46) reduces to (16). Note that in the interior of Ξ, concavity implies continuity. This completes the proof. □

Remark 5. *The proof process above recovers a well-known reformulation for* (46) *in [29, Cor. 3.3.1], which requires the concavity of $h(\boldsymbol{x},\cdot)$, for every $\boldsymbol{x}$. However, we can relax the concavity to the continuity.* □

[11] C. M. Bishop and N. M. Nasrabadi, Pattern Recognition and Machine Learning, pp. 532. Springer, 2006.

[12] Recall the case where Δ is the Wasserstein distance defined in (20).