Mathematical Programming manuscript No.
(will be inserted by the editor)

Residuals-based distributionally robust optimization with covariate information

Rohit Kannan · Güzin Bayraksan · James R. Luedtke

Received: date / Accepted: date

Abstract We present a flexible framework for incorporating covariate information in stochastic optimization when we only have access to limited joint observations of uncertain parameters and covariates. Building on our previous work, we consider data-driven approaches that integrate a general machine learning prediction model within a distributionally robust optimization (DRO) framework. We investigate the asymptotic and finite sample properties of solutions obtained using Wasserstein and phi-divergence-based ambiguity sets within our DRO formulations. We also explore three cross-validation approaches for sizing these ambiguity sets. Through numerical experiments, we validate our theoretical results, study the effectiveness of our approaches for sizing ambiguity sets, and illustrate the benefits of our DRO formulations in the limited data regime even when the prediction model is misspecified.

Keywords Data-driven stochastic programming · distributionally robust optimization · Wasserstein distance · phi-divergences · covariates · machine learning · convergence rate · large deviations

Mathematics Subject Classification (2010) 90C15 · 90C47

1 Introduction

Stochastic programming [40] is a powerful modeling framework for decision-making under uncertainty that finds applications in engineering, operations research, and economics. A standard formulation of a stochastic program is

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z, Y)\right], \tag{1}$$

where z denotes the decision vector, $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ is the decision space, Y denotes the uncertain model parameters, and $c : \mathbb{R}^{d_z} \times \mathbb{R}^{d_y} \to \overline{\mathbb{R}}$ is an extended real-valued objective function. Because the distribution of the random vector Y is typically unknown, popular data-driven approaches for solving problem (1), such as sample average approximation (SAA) [34, 40], only assume access to a finite sample of Y. Often, in real-world applications, the random vector Y (e.g., demand for a new product) can be predicted using knowledge of covariates X (e.g., web chatter and historical demands for similar existing products). In our previous work [28], we investigated extensions of SAA that can incorporate covariate information in problem (1) and studied the asymptotic and finite sample properties of the resulting solutions (see

This research is supported by the Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Contract Number DE-AC02-06CH11357.

R. Kannan
Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA.
E-mail: rohitk@alum.mit.edu

G. Bayraksan
Department of Integrated Systems Engineering, The Ohio State University, Columbus, OH, USA.
E-mail: bayraksan.1@osu.edu

J. R. Luedtke
Department of Industrial & Systems Engineering and Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA.
E-mail: jim.luedtke@wisc.edu

Section 2.2). Despite its favorable theoretical guarantees [28, 34, 40], a potentially key limitation of the SAA approach is that its solutions may exhibit disappointing out-of-sample performance in the small sample size regime [7, 19].

Distributionally robust optimization (DRO) [37] is a systematic modeling framework for addressing ambiguity in the distribution of Y. The DRO counterpart of problem (1) can be formulated as

$$\min_{z\in\mathcal{Z}} \sup_{Q\in\hat{\mathcal{P}}} \mathbb{E}_{Y\sim Q}\left[c(z,Y)\right], \tag{2}$$

where we minimize the worst-case expected objective over an ambiguity set $\hat{\mathcal{P}}$ of distributions. Several studies have shown that if the ambiguity set $\hat{\mathcal{P}}$ is chosen wisely, the DRO problem (2) can regularize a small-sample SAA of problem (1) and its solutions can mitigate the out-of-sample disappointment of decisions determined using the SAA approach (see the reviews [29, 37]).

In this work, we introduce a DRO framework for decision-making under uncertainty in the presence of covariate information and study its theoretical properties. To enable this, we first consider the setup in [2, 8, 39] for incorporating covariate information in problem (1). Suppose we have access to joint observations of the random vector Y and random covariates X. Given a new random observation $X = x$, our goal is to approximate the solution to the *conditional stochastic program*

$$v^*(x) := \min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z,Y) \mid X = x\right]. \tag{SP}$$

Some applications of this framework include shipment planning under demand uncertainty [8, 9], where products' demands can be predicted using past demands, location, and web search results before making production and inventory decisions, and portfolio optimization under market uncertainty [16], where stock prices can be predicted using economic indicators and historical stock data before making investment decisions.

Motivated by applications where we may only have access to limited data, we consider data-driven DRO formulations that incorporate a statistical or machine learning model within a DRO framework in a bid to construct estimators for (SP) with better out-of-sample performance. We formally define our DRO frameworks in Section 3, and analyze their asymptotic and finite sample properties in Sections 4 and 5. Section 4 focuses on ambiguity sets defined using Wasserstein distances, whereas Section 5 studies a family of ambiguity sets with discrete support. We investigate data-driven methods for choosing the radii of these ambiguity sets in the presence of covariate observations in Section 6. Numerical experiments in Section 7 demonstrate the potential benefits of our data-driven DRO estimators in the limited data regime.

1.1 Related work

We begin by reviewing related work that aims to solve the conditional stochastic program (SP) without using DRO. Ban and Rudin [2] and Bertsimas and Kallus [8] study policy-based empirical risk minimization and nonparametric regression-based reweighted SAA approaches for solving (SP). Bertsimas and Kallus [8] establish asymptotic optimality of their data-driven decisions, whereas Ban and Rudin [2] also present finite sample guarantees in the context of the data-driven newsvendor problem. Bazier-Matte and Delage [5] explore linear decision rules for a regularized portfolio selection problem given side information. They derive finite sample and suboptimality performance guarantees for their solutions. Ban et al. [1] and Sen and Deng [39] use parametric regression methods along with their empirical residuals to generate scenarios of the random variables given covariate information. Ban et al. [1] prove asymptotic optimality of their decisions for their particular application. Kannan et al. [28] introduce two new SAA formulations that use leave-one-out residuals. They identify conditions under which solutions to their data-driven SAAs possess asymptotic and finite sample guarantees. Kannan et al. [28] review other data-driven approximations to (SP) that do not use DRO.

Solutions to the above approximations to (SP) might display poor out-of-sample performance when we only have access to limited joint data on the random variables and covariates. DRO offers a structured framework for determining solutions with better out-of-sample performance in such situations. Next, we review related work that attempts to solve (SP) using DRO.

Hanasusanto and Kuhn [26] study multi-stage stochastic programs with time series data. They propose a χ^2-distance-based DRO formulation that uses Nadaraya-Watson regression estimates to approximate value functions, and solve it using an approximate dynamic programming method. Bertsimas et al. [9] consider a multi-stage DRO extension of the approach in [8] using the sample robust optimization

method of [10]. They demonstrate asymptotic optimality of their decisions and develop an approximate solution method using linear decision rules. Bertsimas and Van Parys [12] propose a notion of 'bootstrap robustness'. They define DRO extensions of the Nadaraya-Watson and k-nearest neighbors formulations in [8] using ambiguity sets based on discrepancy measures and study their theoretical properties.

Blanchet et al. [14] and Nguyen et al. [35] consider Wasserstein DRO formulations of single-stage stochastic programs arising in statistics or machine learning applications. Blanchet et al. [14] study how to optimally size their ambiguity sets. Dou and Anitescu [16] consider a tailored Wasserstein DRO formulation of single-stage stochastic convex programs when the data obeys a linear vector autoregressive model and derive its tractable dual. Finally, Esteban-Pérez and Morales [20] use nonparametric regression methods to construct a Wasserstein DRO extension of (SP) by linking trimmings of probability distributions with the partial mass transportation problem. They also allow for the available data to be contaminated, and establish asymptotic and finite sample guarantees for their solutions.

We consider a flexible data-driven DRO extension of (SP) that integrates a machine learning prediction model within a DRO framework. Our work is similar in spirit to [16, 20], but we consider more general formulations (SP), including two-stage stochastic programs, generic prediction models, and more general DRO setups, including ones based on Wasserstein distances, sample robust optimization, and phi-divergences. A key difference between our Wasserstein DRO formulation in Section 4 and the formulation in [16] is that we consider an ambiguity set for the residuals of the prediction model, but do not consider one for its coefficients. We investigate the theoretical properties of our residuals-based DRO formulations in Sections 4 and 5 and explore data-driven approaches for sizing our ambiguity sets in Section 6. The case study in Section 7 demonstrates the benefit of the modularity of our formulations.

1.2 Summary of main contributions

The following summarizes the main contributions of this paper:

1. We introduce a general residuals-based DRO framework for approximating the solution to problem (SP) based on the data-driven SAA framework in [28]. Our DRO framework is flexible and seamlessly extends existing DRO formulations that do not utilize covariate information.
2. We study asymptotic optimality, rates of convergence, and finite sample guarantees of solutions determined using Wasserstein ambiguity sets.
3. We consider a family of ambiguity sets with only discrete distributions and study the asymptotic and finite sample properties of resulting solutions.
4. We investigate three data-driven approaches for choosing the radii of ambiguity sets for our residuals-based DRO formulations.
5. Finally, our numerical experiments investigate the effectiveness of proposed approaches for sizing ambiguity sets, validate our theoretical results, and demonstrate the advantages of our data-driven DRO formulations in the limited data regime even when the prediction model is misspecified.

Notation. Let $[n] := \{1, \cdots, n\}$, $\|\cdot\|_p$ denote the ℓ_p-norm for $p \in [1, +\infty]$, $\text{proj}_S(v)$ denote the orthogonal projection of v onto a nonempty closed convex set S, and δ denote the Dirac measure. We write $\|\cdot\|$ as shorthand for $\|\cdot\|_2$. Let $\mathcal{P}(S)$ denote the space of probability distributions with support contained in the set $S \subseteq \mathbb{R}^{d_y}$. Given $Q_1, Q_2 \in \mathcal{P}(S)$, let $\Pi(Q_1, Q_2)$ denote the set of joint distributions with marginals Q_1 and Q_2. The p-Wasserstein distance $d_{W,p}(Q_1, Q_2)$ between Q_1 and Q_2 with respect to the ℓ_2-norm[1] is given by

$$d_{W,p}(Q_1, Q_2) := \left(\inf_{\pi \in \Pi(Q_1, Q_2)} \int_{S^2} \|y_1 - y_2\|^p d\pi(y_1, y_2) \right)^{1/p}, \quad \text{if } p \in [1, +\infty),$$
$$d_{W,\infty}(Q_1, Q_2) := \inf_{\pi \in \Pi(Q_1, Q_2)} \pi\text{-ess sup}_{S \times S} \|y_1 - y_2\|,$$

where $\pi\text{-ess sup}_{S\times S}\|y_1 - y_2\| := \inf\{C : \pi(\|y_1 - y_2\| > C) = 0\}$ denotes the essential supremum with respect to the measure π. For any $S \subseteq \mathbb{R}^{d_z}$, let $L^\infty(S)$ denote the Banach space of essentially bounded functions on S equipped with the supremum norm. For sets $A, B \subseteq \mathbb{R}^{d_z}$, let $\mathbb{D}(A, B) := \sup_{v \in A} \text{dist}(v, B)$ denote the deviation of A from B, where $\text{dist}(v, B) := \inf_{w \in B} \|v - w\|$.

The abbreviations 'a.e.', 'a.s.', 'LLN', 'i.i.d.', and 'r.h.s.' are shorthand for 'almost everywhere', 'almost surely', 'law of large numbers', 'independent and identically distributed', and 'right-hand side'. For a

[1] Our results can be extended to Wasserstein distances defined using ℓ_q-norms with $q \neq 2$.

random vector V with probability measure P_V, we write a.e. $v \in V$ to denote P_V-a.e. $v \in V$. The symbols $\xrightarrow{p}$, $\xrightarrow{a.s.}$, and $\xrightarrow{d}$ denote convergence in probability, almost surely, and in distribution with respect to the probability measure generating the joint data on the random variables Y and the random covariates X. For random sequences $\{V_n\}$ and $\{W_n\}$, we write $V_n = o_p(W_n)$ and $V_n = O_p(W_n)$ to convey that $V_n = R_n W_n$ with $\{R_n\}$ converging in probability to zero, or being bounded in probability, respectively. We write $O(1)$ to denote generic constants and $v_n = \Theta(w_n)$ to mean that the sequence $\{v_n\}$ is asymptotically bounded both above and below by the sequence $\{w_n\}$. We ignore measurability-related issues throughout this work (see [40, 42] for consideration of these issues).

2 Preliminaries

2.1 Framework

We assume throughout that the random vector Y is related to the random covariates X as $Y = f^*(X)+\varepsilon$, where $f^*(x) := \mathbb{E}[Y \mid X = x]$ is the regression function and the random vector ε is the associated regression error. We also assume that the zero-mean errors ε are independent of the covariates X, and that f^* is known to belong to a class of functions $\mathcal{F}$. The model class $\mathcal{F}$ can be infinite-dimensional and can depend on the sample size n. Let $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, and $\Xi \subseteq \mathbb{R}^{d_y}$ denote the supports of Y, X, and ε, respectively. Additionally, let $P_{Y|X=x}$ denote the conditional distribution of Y given $X = x$ and P_ε denote the distribution of ε. Finally, we assume that the support $\mathcal{Y}$ is nonempty and convex, which ensures that the orthogonal projection onto $\mathcal{Y}$ is unique and Lipschitz continuous. If $\mathcal{Y}$ is not convex (e.g., if it is discrete), one option is to instead project onto its convex hull, $\text{conv}(\mathcal{Y})$, and replace $\mathcal{Y}$ by $\text{conv}(\mathcal{Y})$ in our formulations, assumptions, and results.

Under the above assumptions, problem (SP) is equivalent to

$$v^*(x) := \min_{z \in \mathcal{Z}} \{g(z; x) := \mathbb{E}[c(z, f^*(x) + \varepsilon)]\}, \tag{3}$$

where the expectation is computed with respect to the distribution P_ε of ε. We refer to problem (3) as *the true problem*. We assume throughout that the set $\mathcal{Z}$ is nonempty and compact, $\mathbb{E}[|c(z, f^*(x) + \varepsilon)|] < +\infty$ for each $z \in \mathcal{Z}$ and a.e. $x \in \mathcal{X}$, and the function $g(\cdot; x)$ is lower semicontinuous on $\mathcal{Z}$ for a.e. $x \in \mathcal{X}$. These assumptions ensure that problem (3) is well-defined and the set $S^*(x)$ of optimal solutions to problem (3) is nonempty for a.e. $x \in \mathcal{X}$.

2.2 Review of data-driven SAA formulations

We now summarize the residuals-based SAA formulations considered in [28]. Let $\mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n$ denote the joint observations of (Y, X) and $\{\varepsilon^i\}_{i=1}^n$, with $\varepsilon^i := y^i - f^*(x^i)$, $\forall i \in [n]$, denote the corresponding realizations of the errors. If we know the regression function f^*, then we can construct the following *full-information SAA* (FI-SAA) to problem (3) using the data $\mathcal{D}_n$:

$$\min_{z \in \mathcal{Z}} \Big\{ g_n^*(z; x) := \frac{1}{n} \sum_{i=1}^n c(z, f^*(x) + \varepsilon^i) \Big\}. \tag{4}$$

Because f^* is unknown, we first estimate it by $\hat{f}_n$ using a regression method on the data $\mathcal{D}_n$. We then use $\hat{f}_n$ and its residuals on the training data $\hat{\varepsilon}_n^i := y^i - \hat{f}_n(x^i)$, $i \in [n]$, to construct the following *empirical residuals-based SAA*[2] (ER-SAA) to problem (3):

$$\hat{v}_n^{ER}(x) := \min_{z \in \mathcal{Z}} \Big\{ \hat{g}_n^{ER}(z; x) := \frac{1}{n} \sum_{i=1}^n c\big(z, \text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{\varepsilon}_n^i)\big) \Big\}. \tag{5}$$

When the sample size n is small relative to the complexity of the regression method, the empirical residuals $\{\hat{\varepsilon}_n^i\}_{i=1}^n$ may be optimistically biased and provide a poor estimate of the samples $\{\varepsilon^i\}_{i=1}^n$ of ε.

[2] In contrast with [28], we project the points $(\hat{f}_n(x) + \hat{\varepsilon}_n^i)$, $i \in [n]$, onto the support $\mathcal{Y}$ in this work. This projection step ensures that the Wasserstein and sample robust optimization-based DRO formulations considered in Section 3 are tractable under suitable assumptions on the true problem (3). We stick with this modification of the ER-SAA formulation (5) throughout for uniformity.

This motivated our construction in [28] of two alternative SAA formulations that instead use leave-one-out residuals to construct scenarios of Y given $X = x$. We called these formulations Jackknife-based SAA (J-SAA) and Jackknife+-based SAA (J+-SAA).

Let $P_n^*(x)$ denote the *true empirical distribution* of Y given $X = x$ corresponding to the FI-SAA problem (4) and $\hat{P}_n^{ER}(x)$ denote the *estimated empirical distribution* corresponding to the ER-SAA problem (5), i.e.,

$$P_n^*(x) := \frac{1}{n}\sum_{i=1}^{n} \delta_{f^*(x)+\varepsilon^i}, \qquad \hat{P}_n^{ER}(x) := \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathrm{proj}_{\mathcal{Y}}(\hat{f}_n(x)+\hat{\varepsilon}_n^i)}.$$

A main component of the analysis conducted in this paper is controlling the distance between the estimated empirical distribution $\hat{P}_n^{ER}(x)$ and the true empirical distribution $P_n^*(x)$. By the Lipschitz continuity of orthogonal projections[3], we have for each $x \in \mathcal{X}$

$$\|\mathrm{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{\varepsilon}_n^i) - (f^*(x) + \varepsilon^i)\| \leq \|\tilde{\varepsilon}_n^i(x)\|, \quad \forall i \in [n], \tag{6}$$

where $\tilde{\varepsilon}_n^i(x) := (\hat{f}_n(x) + \hat{\varepsilon}_n^i) - (f^*(x) + \varepsilon^i) = \big(\hat{f}_n(x) - f^*(x)\big) + \big(f^*(x^i) - \hat{f}_n(x^i)\big)$. Note that $\tilde{\varepsilon}_n^i(x)$ equals the sum of the *prediction error* at the new covariate realization $x \in \mathcal{X}$ and the *estimation error* at the training point $x^i \in \mathcal{X}$. We are now ready to present our data-driven DRO formulations.

3 Residuals-based DRO formulations

We consider the following DRO extension of the data-driven SAA formulations reviewed in Section 2.2 to approximate the solution to the true problem (3):

$$\hat{v}_n^{DRO}(x) = \min_{z \in \mathcal{Z}} \sup_{Q \in \hat{\mathcal{P}}_n(x)} \mathbb{E}_{Y \sim Q}\left[c(z, Y)\right], \tag{7}$$

where $\hat{\mathcal{P}}_n(x)$ is a data-driven ambiguity set for the distribution of Y given $X = x$ that is centered at $\hat{P}_n^{ER}(x)$. Let $\hat{z}_n^{DRO}(x)$ denote an optimal solution to problem (7) and $\hat{S}_n^{DRO}(x)$ denote its set of optimal solutions. We assume throughout that for each $x \in \mathcal{X}$, the function $\mathbb{E}_{Y \sim Q}\left[c(\cdot, Y)\right]$ is well-defined and lower semicontinuous on $\mathcal{Z}$ for each $Q \in \hat{\mathcal{P}}_n(x)$. This ensures that for each $x \in \mathcal{X}$, the objective function of the problem (7) is lower semicontinuous on $\mathcal{Z}$ and its optimal solution set $\hat{S}_n^{DRO}(x)$ is nonempty.

Recall that our main motivation for solving the DRO problem (7) is to obtain a solution $\hat{z}_n^{DRO}(x)$ with good out-of-sample performance $g(\hat{z}_n^{DRO}(x); x)$ for small sample sizes n. Since this is hard to formalize, we instead focus on formulations with other desirable guarantees. Given a risk level $\alpha \in (0, 1)$, we wish to construct the ambiguity set $\hat{\mathcal{P}}_n(x)$ such that one or more of the following properties hold for a.e. $x \in \mathcal{X}$ (cf. [7, 19]):

1. **Consistency and asymptotic optimality:** the optimal value $\hat{v}_n^{DRO}(x)$ and solution $\hat{z}_n^{DRO}(x)$ of the residuals-based DRO problem (7) satisfy

$$\hat{v}_n^{DRO}(x) \xrightarrow{p} v^*(x), \;\; \mathrm{dist}(\hat{z}_n^{DRO}(x), S^*(x)) \xrightarrow{p} 0, \;\; g(\hat{z}_n^{DRO}(x); x) \xrightarrow{p} v^*(x).$$

2. **Rate of convergence:** for some constant $r \in (0, 1]$ (ideally close to one), the optimal value $\hat{v}_n^{DRO}(x)$ and solution $\hat{z}_n^{DRO}(x)$ satisfy

$$\left|\hat{v}_n^{DRO}(x) - v^*(x)\right| = O_p\big(n^{-r/2}\big), \quad \left|g(\hat{z}_n^{DRO}(x); x) - v^*(x)\right| = O_p\big(n^{-r/2}\big).$$

3. **Finite sample solution guarantee:** for any $\eta > 0$, there exist positive constants $\Gamma(\eta, x)$ and $\gamma(\eta, x)$ such that $\hat{z}_n^{DRO}(x)$ satisfies

$$\mathbb{P}\left\{\mathrm{dist}(\hat{z}_n^{DRO}(x), S^*(x)) \geq \eta\right\} \leq \Gamma(\eta, x)\exp(-n\gamma(\eta, x)).$$

4. **Finite sample certificate guarantee:** the optimal value $\hat{v}_n^{DRO}(x)$ provides the following certificate on the out-of-sample cost of $\hat{z}_n^{DRO}(x)$:

$$\mathbb{P}\left\{g(\hat{z}_n^{DRO}(x); x) \leq \hat{v}_n^{DRO}(x)\right\} \geq 1 - \alpha.$$

[3] For any $u, v \in \mathbb{R}^{d_y}$, $\|\mathrm{proj}_{\mathcal{Y}}(u) - \mathrm{proj}_{\mathcal{Y}}(v)\| \leq \|u - v\|$.

We would also like the DRO problem (7) to be efficiently solvable in practice.

We call problem (7) with the ambiguity set $\hat{\mathcal{P}}_n(x)$ centered at $\hat{P}_n^{ER}(x)$ the *empirical residuals-based DRO* (ER-DRO) problem. While in this paper we focus our attention on ER-DRO formulations, note that the ambiguity set $\hat{\mathcal{P}}_n(x)$ can also be centered at the estimated empirical distributions corresponding to its Jackknife-based counterparts. The analysis in [28, Appendix EC.1] can be used to extend this paper's results for ER-DRO to its Jackknife-based variants.

In the remainder of this work, we focus on the use of the following data-driven ambiguity sets $\hat{\mathcal{P}}_n(x)$ in the construction of ER-DRO problem (7). Unlike the classical DRO setting [37], we allow the radius of these ambiguity sets $\hat{\mathcal{P}}_n(x)$ to depend not only on the sample size n and the risk level α that, e.g., shows up in the finite sample certificate, but also on the covariate realization $x \in \mathcal{X}$; see $\zeta_n(x)$ and $\mu_n(x)$ below. We often omit the dependence of the radius on α to simplify notation.

1. Wasserstein-based ambiguity sets (cf. [19, 25, 36]): given radius $\zeta_n(x) \geq 0$ and order $p \in [1, +\infty]$, set
$$\hat{\mathcal{P}}_n(x) = \left\{Q \in \mathcal{P}(\mathcal{Y}) : d_{W,p}(Q, \hat{P}_n^{ER}(x)) \leq \zeta_n(x)\right\}.$$
2. Sample robust optimization-based ambiguity sets (cf. [10, 45]): given radius $\mu_n(x) \geq 0$ and parameter $p \in [1, +\infty]$, set[4]
$$\hat{\mathcal{P}}_n(x) = \left\{Q = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{y}^i} : \|\bar{y}^i - \text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{\varepsilon}_n^i)\|_p \leq \mu_n(x), \bar{y}^i \in \mathcal{Y}, \forall i \in [n]\right\}.$$
We focus on ambiguity sets constructed using $p = 2$ to keep the exposition simple, but our analysis also extends to ambiguity sets with $p \neq 2$.
3. Ambiguity sets with the same support as $\hat{P}_n^{ER}(x)$ (cf. [4, 6], for instance): given radius $\zeta_n(x) \geq 0$, set
$$\hat{\mathcal{P}}_n(x) = \left\{Q = \sum_{i=1}^n p_i \delta_{\text{proj}_{\mathcal{Y}}(\hat{f}_n(x)+\hat{\varepsilon}_n^i)} : p \in \mathfrak{P}_n(x; \zeta_n(x))\right\},$$
where $\mathfrak{P}_n(x; \zeta_n(x))$ is a generic ambiguity set for the n-dimensional vector of probabilities p. We focus on sets $\mathfrak{P}_n(x; \zeta_n(x))$ that satisfy for each $x \in \mathcal{X}$
$$\begin{aligned} &p \in \mathbb{R}_+^n \text{ and } \sum_{i=1}^n p_i = 1, \quad \forall p \in \mathfrak{P}_n(x; \zeta_n(x)), \\ &\lim_{\zeta \downarrow 0} \mathfrak{P}_n(x; \zeta) = \mathfrak{P}_n(x; 0) = \left\{\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)\right\}. \end{aligned} \tag{8}$$
The above family of ambiguity sets—that use the same support as $\hat{P}_n^{ER}(x)$—result in tractable ER-DRO formulations (7) under milder assumptions on the true problem (3) compared to Wasserstein and sample robust optimization ambiguity sets, which go beyond the support of $\hat{P}_n^{ER}(x)$.

We now provide two examples of the last category of ambiguity sets. Appendix B includes a third example based on mean-upper semideviations.

Example 1 CVaR-based ambiguity set [38, 40]: given radius $\zeta_n(x) \in [0, 1)$, set
$$\mathfrak{P}_n(x; \zeta_n(x)) := \left\{p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1,\ p_i \leq \frac{1}{n(1-\zeta_n(x))},\ \forall i \in [n]\right\}.$$
Observe that $\zeta_n(x)$ enters the ambiguity set $\mathfrak{P}_n(x; \zeta_n(x))$ through the CVaR risk parameter.

Example 2 Phi-divergence-based ambiguity sets [4, 6]: Let $\phi : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ be a lower semicontinuous, convex phi-divergence function with a unique minimum at 1 and $\phi(1) = 0$. Given radius $\zeta_n(x) \geq 0$, define $\hat{\mathcal{P}}_n(x)$ using
$$\mathfrak{P}_n(x; \zeta_n(x)) := \left\{p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1,\ \frac{1}{n}\sum_{i=1}^n \phi(np_i) \leq \zeta_n(x)\right\}.$$
Particular instances include Kullback Leibler divergence, variation distance, and Hellinger distance-based ambiguity sets.

[4] We use $\mu_n(x)$ to avoid a clash with the notation $\zeta_n(x)$ for the radius of ambiguity sets with the same support as $\hat{P}_n^{ER}(x)$. Having different notation for these two radii will prove useful in our unified analysis of the corresponding ER-DRO problems in Section 5.

In the next section, we investigate the theoretical properties of using Wasserstein ambiguity sets within the ER-DRO problem. In Section 5, we present a unified analysis of the theoretical properties of using both sample robust optimization ambiguity sets and ambiguity sets with the same support as $\hat{P}_n^{ER}(x)$. Hereafter, we often write $\hat{\mathcal{P}}_n(x;\zeta_n(x))$ instead of $\hat{\mathcal{P}}_n(x)$ to make its dependence on the radius $\zeta_n(x)$ explicit. We also write $\zeta_n(\alpha,x)$ instead of $\zeta_n(x)$ when we want to emphasize the dependence of the radius on the risk level α.

4 Wasserstein-based ambiguity sets

We now establish asymptotic optimality, rates of convergence, and finite sample guarantees for ER-DRO formulations defined using p-Wasserstein distance-based ambiguity sets with $p \in [1,+\infty)$. Section 5 presents analysis for ambiguity sets defined using the ∞-Wasserstein distance by exploiting a link with sample robust optimization [10]. Sections 4.1 and 5 of [19] and Section 2.2 of [29] identify conditions under which the resulting ER-DRO formulation (7) is computationally tractable. References [3, 25, 27] also consider solution approaches for when problem (3) is a two-stage stochastic program.

We begin with a light-tail assumption on the distribution P_ε of the errors ε.

Assumption 1 There is a constant $a > p$ such that $\mathbb{E}\left[\exp(\|\varepsilon\|^a)\right] < +\infty$.

Next, we make a finite sample assumption on the regression estimate $\hat{f}_n$.

Assumption 2 The regression estimate $\hat{f}_n$ possesses the following finite sample property: for a.e. $x \in \mathcal{X}$ and any risk level $\alpha \in (0,1)$, there exists a constant $\kappa_{p,n}(\alpha,x) > 0$ such that

$$\mathbb{P}\left\{\|f^*(x) - \hat{f}_n(x)\|^p > \kappa_{p,n}^p(\alpha,x)\right\} \le \alpha, \quad \text{and}$$
$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^p > \kappa_{p,n}^p(\alpha,x)\right\} \le \alpha.$$

Appendix EC.3 of [28] identifies conditions under which parametric regression methods such as OLS and Lasso regression satisfy Assumption 2 for the case $p = 2$ with constants $\kappa_{2,n}^2(\alpha,x) = O(n^{-1}\log(\alpha^{-1}))$. Similar bounds readily hold even for $p \neq 2$, e.g., if the support $\mathcal{X}$ of the covariates is compact. Non-parametric regression methods, on the other hand, typically only satisfy Assumption 2 with constants $\kappa_{p,n}^p(\alpha,x) = O(n^{-1}\log(\alpha^{-1}))^{O(1)/d_x}$ and perform worse with increasing covariate dimension d_x due to the curse of dimensionality. If Assumption 2 holds for $p = 2$, the power mean inequality implies that it also holds for any $p \in [1,2)$ with $\kappa_{p,n}(\alpha,x) = \kappa_{2,n}(\alpha,x)$.

We make the light-tail Assumption 1 on the distribution P_ε of the errors ε to invoke the concentration inequality in Lemma 1 for the true empirical distribution $P_n^*(x)$. Throughout, we assume $p \neq d_y/2$ for a slightly simpler form of this concentration inequality; see [21, Theorem 2] for the case $p = d_y/2$. Lemma 1 also applies to non-i.i.d. data $\mathcal{D}_n$ such as time series data (cf. [16]).

Lemma 1 (Theorem 2 of [21]) *Suppose Assumption 1 holds, $p \neq d_y/2$, and the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. Then, for all $\kappa > 0$, $n \in \mathbb{N}$, and $x \in \mathcal{X}$*

$$\mathbb{P}\left\{d_{W,p}(P_n^*(x), P_{Y|X=x}) \ge \kappa\right\} \le \begin{cases} O(1)\exp(-O(1)n\kappa^{\max\{d_y/p,2\}}) & \text{if } \kappa \le 1 \\ O(1)\exp(-O(1)n\kappa^{a/p}) & \text{if } \kappa > 1 \end{cases}.$$

We require a few intermediate results before we can establish a finite sample certificate guarantee for Wasserstein ER-DRO estimators in Theorem 5 (cf. [29, Theorem 19]). The first result bounds the p-Wasserstein distance between the estimated empirical distribution $\hat{P}_n^{ER}(x)$ and the conditional distribution $P_{Y|X=x}$ of Y given $X = x$.

Lemma 2 *For each $x \in \mathcal{X}$*

$$d_{W,p}(\hat{P}_n^{ER}(x), P_{Y|X=x}) \le \left(\frac{1}{n}\sum_{i=1}^n \|\hat{\varepsilon}_n^i(x)\|\right)^{1/p} + d_{W,p}(P_n^*(x), P_{Y|X=x}).$$

Proof The triangle inequality for the p-Wasserstein distance yields

$$d_{W,p}(\hat{P}_n^{ER}(x), P_{Y|X=x}) \leq d_{W,p}(\hat{P}_n^{ER}(x), P_n^*(x)) + d_{W,p}(P_n^*(x), P_{Y|X=x}).$$

The stated result then follows from the definition of the p-Wasserstein distance and inequality (6) since

$$\begin{aligned} d_{W,p}(\hat{P}_n^{ER}(x), P_n^*(x)) &\leq \left(\frac{1}{n}\sum_{i=1}^n \|\mathrm{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{\varepsilon}_n^i) - (f^*(x) + \varepsilon^i)\|^p\right)^{1/p} \\ &\leq \left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^p\right)^{1/p}. \end{aligned}$$

□

The next result bounds the power mean deviation $\left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^p\right)^{1/p}$.

Lemma 3 *For each $x \in \mathcal{X}$*

$$\left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^p\right)^{1/p} \leq \|f^*(x) - \hat{f}_n(x)\| + \left(\frac{1}{n}\sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^p\right)^{1/p}.$$

Proof We have from the definition of $\tilde{\varepsilon}_n^i(x)$ that

$$\begin{aligned} \left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^p\right)^{1/p} &\leq \left(\frac{1}{n}\sum_{i=1}^n \left(\|f^*(x) - \hat{f}_n(x)\| + \|f^*(x^i) - \hat{f}_n(x^i)\|\right)^p\right)^{1/p} \\ &\leq \|f^*(x) - \hat{f}_n(x)\| + \left(\frac{1}{n}\sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^p\right)^{1/p}, \end{aligned}$$

where the first step follows from the triangle inequality for the ℓ_2-norm, and the second step follows from the triangle inequality for the ℓ_p-norm. □

Finally, we derive a finite sample guarantee for $\left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^p\right)^{1/p}$.

Lemma 4 *Suppose Assumption 2 holds and $\alpha \in (0,1)$. Then for a.e. $x \in \mathcal{X}$,*

$$\mathbb{P}\left\{\left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^p\right)^{1/p} > 2\kappa_{p,n}\left(\frac{\alpha}{4}, x\right)\right\} \leq \frac{\alpha}{2}.$$

Proof We have for a.e. $x \in \mathcal{X}$

$$\begin{aligned} &\mathbb{P}\left\{\left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^p\right)^{1/p} > 2\kappa_{p,n}\left(\frac{\alpha}{4}, x\right)\right\} \\ \leq &\mathbb{P}\left\{\|f^*(x) - \hat{f}_n(x)\| + \left(\frac{1}{n}\sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^p\right)^{1/p} > 2\kappa_{p,n}\left(\frac{\alpha}{4}, x\right)\right\} \\ \leq &\mathbb{P}\left\{\|f^*(x) - \hat{f}_n(x)\| > \kappa_{p,n}\left(\frac{\alpha}{4}, x\right)\right\} + \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^p > \kappa_{p,n}^p\left(\frac{\alpha}{4}, x\right)\right\} \\ \leq &\frac{\alpha}{4} + \frac{\alpha}{4} = \frac{\alpha}{2}, \end{aligned}$$

where the first step follows by Lemma 3, the second step follows from the inequality $\mathbb{P}(V+W > c_1+c_2) \leq \mathbb{P}(V > c_1) + \mathbb{P}(W > c_2)$ for any random variables V, W and constants c_1, c_2, and the last step holds by Assumption 2. □

To establish asymptotic and finite sample guarantees in Theorems 5 to 8, we need to enlarge the radius of the Wasserstein ambiguity set that is used in the absence of covariate information [19, 29]. This enlargement accounts for the error in estimating the regression function f^*. In particular, for a given covariate realization $x \in \mathcal{X}$ and risk level $\alpha \in (0,1)$, we use

$$\zeta_n(\alpha, x) := \kappa_{p,n}^{(1)}(\alpha, x) + \kappa_{p,n}^{(2)}(\alpha) \tag{9}$$

as the radius of the ambiguity set, where $\kappa^{(1)}_{p,n}(\alpha, x) := 2\kappa_{p,n}\left(\frac{\alpha}{4}, x\right)$ and

$$\kappa^{(2)}_{p,n}(\alpha) := \begin{cases} \left(\frac{O(1)\log(O(1)\alpha^{-1})}{n}\right)^{\min\{p/d_y, 1/2\}} & \text{if } n \geq O(1)\log(O(1)\alpha^{-1}) \\ \left(\frac{O(1)\log(O(1)\alpha^{-1})}{n}\right)^{p/a} & \text{if } n < O(1)\log(O(1)\alpha^{-1}) \end{cases}.$$

The constants a and $\kappa_{p,n}$ above are defined in Assumptions 1 and 2. While this choice of ζ_n helps us derive our theoretical guarantees, it involves unknown constants and is typically conservative in practice (see Remark 2). We investigate practical data-driven approaches for choosing the radius ζ_n in Section 6.

Theorem 5 (Finite sample certificate guarantee) *Suppose Assumptions 1 and 2 hold, $\alpha \in (0,1)$ is a given risk level, and the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. Then, for a.e. $x \in \mathcal{X}$, the finite sample certificate guarantee holds for the ER-DRO problem (7) with the radius $\zeta_n(\alpha, x)$ of the ambiguity set $\hat{\mathcal{P}}_n(x; \zeta_n(\alpha, x))$ specified by equation (9).*

Proof Lemma 4 and Lemma 1 imply that

$$\mathbb{P}\left\{\left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}^i_n(x)\|^p\right)^{1/p} > \kappa^{(1)}_{p,n}(\alpha, x)\right\} \leq \frac{\alpha}{2}, \quad \text{for a.e. } x \in \mathcal{X},$$
$$\mathbb{P}\left\{d_{W,p}(P^*_n(x), P_{Y|X=x}) > \kappa^{(2)}_{p,n}(\alpha)\right\} \leq \frac{\alpha}{2}, \quad \forall x \in \mathcal{X}.$$

Consequently, equation (9), Lemma 2, and the same probability inequality used in the proof of Lemma 4 imply that

$$\mathbb{P}\left\{d_{W,p}(\hat{P}^{ER}_n(x), P_{Y|X=x}) > \zeta_n(\alpha, x)\right\} \leq \alpha \text{ for a.e. } x \in \mathcal{X}.$$

The stated result follows from the definition of the ER-DRO problem (7). □

We now make the following assumption along the lines of [19, 25, 29] to show in Theorem 6 that solutions to the ER-DRO problem (7) with radii $\zeta_n(\alpha_n, x)$ are asymptotically optimal for a suitable sequence of risk levels $\{\alpha_n\}$.

Assumption 3 The function $c(\cdot, Y)$ is lower semicontinuous on $\mathcal{Z}$ for each $Y \in \mathcal{Y}$ and the function $c(z, \cdot)$ is continuous on $\mathcal{Y}$ for each $z \in \mathcal{Z}$. Furthermore, there exists a constant $B_{c,p} \geq 0$ such that

$$|c(z, Y)| \leq B_{c,p}(1 + \|Y\|^p), \quad \forall z \in \mathcal{Z},\ Y \in \mathcal{Y}.$$

We also make either of the following assumptions on the function c to establish a rate of convergence of the ER-DRO estimator in Theorem 7.

Assumption 4 For each $z \in \mathcal{Z}$, the function $c(z, \cdot)$ is Lipschitz continuous on $\mathcal{Y}$ with Lipschitz constant $L_1(z)$.

Assumption 5 The Wasserstein order $p \geq 2$. Furthermore, for each $z \in \mathcal{Z}$, the function $c(z, \cdot)$ is differentiable on $\mathcal{Y}$ with $\mathbb{E}\left[\|\nabla c(z, Y)\|^2\right] < +\infty$ and

$$\|\nabla c(z, \bar{y}) - \nabla c(z, y)\| \leq L_2(z)\|\bar{y} - y\|, \quad \forall y, \bar{y} \in \mathcal{Y}.$$

Assumptions 3, 4, and 5 hold for broad classes of stochastic programs, including two-stage stochastic mixed-integer linear programs (MIPs) with continuous recourse [28, Appendix EC.2]. With these assumptions in place, we obtain the following asymptotic results.

Theorem 6 (Consistency and asymptotic optimality) *Suppose Assumptions 1, 2, and 3 hold, the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d., there is a sequence of risk levels $\{\alpha_n\}_{n\in\mathbb{N}} \subset (0,1)$ such that $\sum_n \alpha_n < +\infty$, and $\lim_{n\to\infty} \zeta_n(\alpha_n, x) = 0$ for a.e. $x \in \mathcal{X}$ with the radius ζ_n defined in equation (9). Then, for a.e. $x \in \mathcal{X}$, the optimal value and solution of the ER-DRO problem (7) with ambiguity set $\hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x))$ are consistent and asymptotically optimal.*

Theorem 7 (Rate of convergence) *Suppose the assumptions of Theorem 6 and either Assumption 4 or Assumption 5 hold. Then, for a.e. $x \in \mathcal{X}$, the optimal value and solution of the ER-DRO problem (7) with ambiguity set $\hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x))$ satisfy*

$$\left|\hat{v}^{DRO}_n(x) - v^*(x)\right| = O_p\left(\zeta_n(\alpha_n, x)\right), \quad \left|g(\hat{z}^{DRO}_n(x); x) - v^*(x)\right| = O_p\left(\zeta_n(\alpha_n, x)\right).$$

We relegate the proofs of Theorems 6 and 7 to Appendices A.1 and A.2. The proof of Theorem 6 mirrors the proof of [19, Theorem 3.6]. Similar to the setting without covariate information [19], we can typically choose the sequence of risk levels $\{\alpha_n\}$ in Theorems 6 and 7 to be any summable sequence that converges to zero more slowly than the sequence $\{\exp(-n)\}$ when the errors ε are sub-Gaussian (see the discussion following Assumption 2).

Remark 1 Assumption 5 can be weakened to consider functions c that satisfy

$$\|\nabla c(z,\bar{y}) - \nabla c(z,y)\| \leq L_2(z,y)\|\bar{y}-y\|^{\kappa}, \quad \forall z \in \mathcal{Z},\ y,\bar{y} \in \mathcal{Y},$$

and $\mathbb{E}\left[\|L_2(z,Y)\|^{p/(p-1)}\right] < +\infty$, $\forall z \in \mathcal{Z}$, for some constant $\kappa \in (0,1]$ and orders $p \geq 1+\kappa$, see [24, Proposition 1]. Furthermore, Assumption 5 can also be weakened to consider functions c of the form $c(z,Y) = \max_{j\in[N_c]} c_j(z,Y)$, where $N_c \in \mathbb{N}$ and for each $z \in \mathcal{Z}$, the constituent functions $c_j(z,\cdot)$ are differentiable on $\mathcal{Y}$ and satisfy $\mathbb{E}\big[\max_{j\in[N_c]}\|\nabla c_j(z,Y)\|^2\big] < +\infty$ and

$$\|\nabla c_j(z,\bar{y}) - \nabla c_j(z,y)\| \leq L_{j,2}(z)\|\bar{y}-y\|, \quad \forall y,\bar{y} \in \mathcal{Y},\ j \in [N_c].$$

The above weakening of Assumption 5 makes it applicable to a larger class of stochastic programs. We stick with Assumption 5 for simplicity.

Remark 2 Recall the radius given in (9) consists of two parts. For the part that relates to the Wasserstein ambiguity set without covariate information, because the rate $d_{W,p}(P_n^*(x), P_{Y|X=x}) = O_p(n^{-p/d_y})$ cannot be improved in general (see [29, Example 3]), we usually have $\kappa_{p,n}^{(2)}(\alpha_n)$ converging to zero only at the slow rate $\Theta(n^{-p/d_y})$. Therefore, the convergence rate afforded by Theorem 7 suffers from the curse of dimensionality even when we use parametric regression methods, which typically exhibit better rates of convergence on the part of the radius that relates to the estimation of f^* (cf. [28, Theorem 2]). The analysis in Gao [23] implies that, under certain assumptions, using the larger radius $\zeta_n(\alpha,x) := \max\{\kappa_{p,n}^{(1)}(\alpha,x), \bar{\kappa}_{p,n}^{(2)}(\alpha)\}$ with suitably chosen $\bar{\kappa}_{p,n}^{(2)}(\alpha) = O(n^{-1/2})$ results in estimators with a finite sample certificate-type guarantee (cf. [15]). This larger choice of the radius ζ_n also yields estimators with the conventional $O_p(n^{-1/2})$ rate of convergence when we use parametric regression methods to estimate the function f^*.

We now identify conditions under which the ER-DRO estimators possess a finite sample solution guarantee. In order to achieve this, we first refine Assumption 2 to a more convenient, stronger form in Assumption 6.

Assumption 6 The regression estimate $\hat{f}_n$ possesses the following large deviation properties: for any constant $\kappa > 0$, there exist positive constants $K_{p,f}(\kappa,x)$, $\bar{K}_{p,f}(\kappa)$, $\beta_{p,f}(\kappa,x)$, and $\bar{\beta}_{p,f}(\kappa)$ satisfying

$$\mathbb{P}\left\{\|f^*(x) - \hat{f}_n(x)\|^p > \kappa^p\right\} \leq K_{p,f}(\kappa,x)\exp\left(-n\beta_{p,f}(\kappa,x)\right), \text{ for a.e. } x \in \mathcal{X},$$
$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^p > \kappa^p\right\} \leq \bar{K}_{p,f}(\kappa)\exp\left(-n\bar{\beta}_{p,f}(\kappa)\right).$$

Appendix EC.3 of [28] verifies Assumption 6 for some popular regression setups for $p = 2$; see the discussion after Assumption 2 for cases when $p \neq 2$.

Theorem 8 (Finite sample solution guarantee) *Suppose Assumptions 1, 2, 3, and 6 hold, the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d., and either Assumption 4 or Assumption 5 holds. Then, for a.e. $x \in \mathcal{X}$, there is a risk level $\alpha \in (0,1)$ such that a finite sample solution guarantee holds for the ER-DRO problem (7) when the radius $\zeta_n(\alpha,x)$ of the ambiguity set is specified by equation (9).*

Proof We first show that for any $\kappa > 0$, there exist constants $\tilde{\Gamma}(\kappa,x) > 0$ and $\tilde{\gamma}(\kappa,x) > 0$ such that

$$\mathbb{P}\left\{g(\hat{z}_n^{DRO}(x);x) > v^*(x) + \kappa\right\} \leq \tilde{\Gamma}(\kappa,x)\exp(-n\tilde{\gamma}(\kappa,x)) \tag{10}$$

for a suitable choice of the risk level α. First, note that Theorem 5 implies $g(\hat{z}_n^{DRO}(x);x) \leq \hat{v}_n^{DRO}(x)$ with probability at least $1-\alpha$ when the radius $\zeta_n(\alpha,x)$ is chosen according to equation (9). This together with the inequality $\mathbb{P}\{V+W>0\} \leq \mathbb{P}\{V>0\} + \mathbb{P}\{W>0\}$ with $V = g(\hat{z}_n^{DRO}(x);x) - \hat{v}_n^{DRO}(x)$ and $W = \hat{v}_n^{DRO}(x) - v^*(x) - \kappa$ yields

$$\mathbb{P}\left\{g(\hat{z}_n^{DRO}(x);x) > v^*(x) + \kappa\right\} \leq \alpha + \mathbb{P}\left\{\hat{v}_n^{DRO}(x) > v^*(x) + \kappa\right\}.$$

Suppose Assumption 4 holds. From inequality (15) in Appendix A.2, we have for any $z^*(x) \in S^*(x)$ that

$$\mathbb{P}\left\{\hat{v}_n^{DRO}(x) > v^*(x) + 2L_1(z^*(x))\zeta_n(\alpha, x)\right\} \leq \alpha.$$

Therefore, if we choose $\alpha \in (0,1)$ so that $2L_1(z^*(x))\zeta_n(\alpha, x) \leq \kappa$, we have

$$\mathbb{P}\left\{g(\hat{z}_n^{DRO}(x); x) > v^*(x) + \kappa\right\} \leq 2\alpha.$$

Equation (9) implies that $2L_1(z^*(x))\kappa_{p,n}^{(2)}(\alpha) \leq \kappa/2$ whenever the risk level $\alpha \geq O(1)\exp\big(-O(1)n\big(\frac{\kappa}{4L_1(z^*(x))}\big)^{1/\theta}\big)$ with θ equal to $\min\{p/d_y, 1/2\}$ or p/a. Assumption 6 implies that we can choose the constant $\kappa_{p,n}^{(1)}(\alpha, x)$ in equation (9) such that for a.e. $x \in \mathcal{X}$, $2L_1(z^*(x))\kappa_{p,n}^{(1)}(\alpha, x) \leq \kappa/2$ whenever

$$\alpha \geq 4\max\Big\{K_{p,f}\big(\tfrac{\kappa}{8L_1(z^*(x))}, x\big)\exp\big(-n\beta_{p,f}\big(\tfrac{\kappa}{8L_1(z^*(x))}, x\big)\big), \\ \bar{K}_{p,f}\big(\tfrac{\kappa}{8L_1(z^*(x))}\big)\exp\big(-n\bar{\beta}_{p,f}\big(\tfrac{\kappa}{8L_1(z^*(x))}\big)\big)\Big\}.$$

The above bounds yield a risk level α such that $2L_1(z^*(x))\zeta_n(\alpha, x) \leq \kappa$ holds, which in turn implies inequality (10) holds for suitably defined constants.

Next, suppose instead that Assumption 5 holds. From inequality (16) in Appendix A.2, we have for any $z^*(x) \in S^*(x)$ that

$$\mathbb{P}\big\{\hat{v}_n^{DRO}(x) > v^*(x) + O(1)\zeta_n(\alpha, x) + 4L_2(z^*(x))\zeta_n^2(\alpha, x)\big\} \leq \alpha.$$

Therefore, if we pick $\alpha \in (0,1)$ so that $O(1)\zeta_n(\alpha, x) + 4L_2(z^*(x))\zeta_n^2(\alpha, x) \leq \kappa$, then

$$\mathbb{P}\left\{g(\hat{z}_n^{DRO}(x); x) > v^*(x) + \kappa\right\} \leq 2\alpha.$$

Similar to the analysis above, inequality (10) can be obtained by bounding the smallest value of α using Assumption 6 and Lemma 1 so that $\kappa \geq O(1)\zeta_n(\alpha_n, x) + 4L_2(z^*(x))\zeta_n^2(\alpha_n, x)$.

We now argue that inequality (10) implies the stated result. Suppose for some $\eta > 0$, $x \in \mathcal{X}$, and some sample path, we have $\mathrm{dist}(\hat{z}_n^{DRO}(x), S^*(x)) \geq \eta$. Since $g(\cdot; x)$ is lower semicontinuous on the compact set $\mathcal{Z}$ for a.e. $x \in \mathcal{X}$, [28, Lemma 3] implies $g(\hat{z}_n^{DRO}(x); x) > v^*(x) + \kappa(\eta, x)$ for some constant $\kappa(\eta, x) > 0$ on that path (except for some paths of measure zero). We now bound the probability of this event. The above arguments imply for a.e. $x \in \mathcal{X}$,

$$\mathbb{P}\left\{\mathrm{dist}(\hat{z}_n^{DRO}(x), S^*(x)) \geq \eta\right\} \leq \mathbb{P}\left\{g(\hat{z}_n^{DRO}(x); x) > v^*(x) + \kappa(\eta, x)\right\} \\ \leq \tilde{\Gamma}(\kappa(\eta, x), x)\exp(-n\tilde{\gamma}(\kappa(\eta, x), x)).$$ □

5 Sample robust optimization-based ambiguity sets and ambiguity sets with the same support as $\hat{P}_n^{ER}(x)$

We present a unified analysis of using both sample robust optimization-based ambiguity sets and ambiguity sets with the same support as $\hat{P}_n^{ER}(x)$ within problem (7). Throughout this section, we consider ambiguity sets of the form

$$\hat{\mathcal{P}}_n(x) := \left\{Q = \sum_{i=1}^n p_i \delta_{\bar{y}^i} : p \in \mathfrak{P}_n(x; \zeta_n(x)),\ \bar{y}^i \in \hat{\mathcal{Y}}_n^i(x; \mu_n(x)), \forall i \in [n]\right\},$$
$$\hat{\mathcal{Y}}_n^i(x; \mu_n(x)) := \left\{y \in \mathcal{Y} : \|y - \mathrm{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{\varepsilon}_n^i)\| \leq \mu_n(x)\right\}, \forall i \in [n],$$

where $\mu_n(x)$ and $\zeta_n(x)$ are nonnegative radii, and the ambiguity set $\mathfrak{P}_n(x; \zeta_n(x))$ for the probabilities p satisfies (8). This family of ambiguity sets generalizes both sample robust optimization-based ambiguity sets constructed using the ℓ_2-norm (obtained by setting $\zeta_n(x) = 0$) and ambiguity sets with the same support as $\hat{P}_n^{ER}(x)$ (obtained by setting $\mu_n(x) = 0$). We establish asymptotic optimality, rates of convergence, and finite sample-type guarantees for the corresponding ER-DRO estimators (7).

When $\mu_n(x) = 0$ and problem (3) is a tractable convex program, the resulting ER-DRO problem (7) remains tractable and convex for many choices of the ambiguity set $\mathfrak{P}_n(x; \zeta_n(x))$ such as Examples 1 and 2 (see, e.g., [6]). On the other hand, when $\mu_n(x) > 0$ and problem (3) is a two-stage stochastic linear program, then the ER-DRO problem (7) exhibits a min - max - min structure whose solution is in general NP-hard. References [11, 44] investigate approaches for approximately solving the ER-DRO problem (7) when the true problem (3) is a two-stage stochastic LP and $\zeta_n(x) = 0$.

To facilitate our analysis, denote by $\hat{g}^{ER}_{s,n}$ and $g^*_{s,n}$ the functions

$$\hat{g}^{ER}_{s,n}(z;x) := \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n p_i \sup_{y\in\hat{\mathcal{Y}}^i_n(x;\mu_n(x))} c(z,y),$$

$$g^*_{s,n}(z;x) := \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n p_i c(z, f^*(x)+\varepsilon^i).$$

Note that the function $\hat{g}^{ER}_{s,n}$ is equivalent to the objective function of the ER-DRO problem (7) with the above definition of the ambiguity set $\hat{\mathcal{P}}_n(x)$. Additionally, $g^*_{s,n}$ is equivalent to the objective function of the FI-SAA problem (4) when $\zeta_n(x) = 0$ and condition (8) holds.

We begin by investigating conditions under which the optimal value and set of optimal solutions to the ER-DRO problem (7) converge in probability to the true problem (3). We make the following assumptions in this regard.

Assumption 7 For each $z \in \mathcal{Z}$, the function $c(z,\cdot)$ is Lipschitz continuous on $\mathcal{Y}$ with Lipschitz constant $L(z)$ satisfying $\sup_{z\in\mathcal{Z}} L(z) < +\infty$.

Assumption 8 For a.e. $x \in \mathcal{X}$, the sequence of FI-SAA objectives $\{g^*_n(\cdot;x)\}$ converges in probability to the function $g(\cdot;x)$ uniformly on the set $\mathcal{Z}$.

Assumption 9 The regression estimate $\hat{f}_n$ has the consistency properties

$$\hat{f}_n(x) \xrightarrow{p} f^*(x), \text{ for a.e. } x\in\mathcal{X}, \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 \xrightarrow{p} 0.$$

Assumption 7 is a uniform Lipschitz continuity assumption that strengthens Assumption 4. Appendix EC.2 of [28] verifies that Assumption 7 holds for two-stage stochastic MIPs with continuous recourse. Assumption 8 is a uniform weak LLN assumption, whereas Assumption 9 is a mild consistency assumption that holds for many popular regression setups (cf. Assumptions 3 and 4 of [28]). Assumption 9 is weaker than the finite sample Assumption 2. We require the following additional assumptions.

Assumption 10 The radius $\zeta_n(x)$ of the ambiguity set is chosen such that

$$\sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n \left(p_i - \frac{1}{n}\right)^2 = O(n^{-\rho}), \quad \text{for a.e. } x\in\mathcal{X},$$

for some constant $\rho > 1$.

Assumption 11 The following weak uniform LLN holds for a.e. $x \in \mathcal{X}$:

$$\sup_{z\in\mathcal{Z}} \left| \frac{1}{n}\sum_{i=1}^n \left(c(z, f^*(x)+\varepsilon^i)\right)^2 - \mathbb{E}\left[\left(c(z, f^*(x)+\varepsilon)\right)^2\right]\right| \xrightarrow{p} 0,$$

with $\sup_{z\in\mathcal{Z}} \mathbb{E}\left[\left(c(z, f^*(x)+\varepsilon)\right)^2\right] < +\infty$ for a.e. $x \in \mathcal{X}$.

Assumption 10 requires us to choose the radius $\zeta_n(x)$ so that the ambiguity set $\mathfrak{P}_n(x;\zeta_n(x))$ converges to the singleton $\left(\frac{1}{n},\ldots,\frac{1}{n}\right)$ at a fast enough rate. This is always possible since we assume equation (8) holds. We are interested in cases when Assumption 10 holds with $\rho \in (1,2]$ (see Theorem 11). Appendix B provides conditions when such a convergence rate holds. Theorem 7.48 of [40] presents conditions under which Assumption 11 holds when the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. Assumption 11 can also be equivalently stated as a weak uniform LLN assumption on the sample variance of the sequence $\{c(z, f^*(x)+\varepsilon^i)\}_{i=1}^n$ [17].

Our first result identifies conditions under which the sequence of objective functions $\{\hat{g}^{ER}_{s,n}(\cdot;x)\}$ of the ER-DRO problem (7) converges uniformly to the objective function $g(\cdot;x)$ of the true problem (3) on $\mathcal{Z}$. Theorem 9 of [17] presents an analogous result for a class of phi-divergence-based ambiguity sets in the absence of covariate information.

Proposition 9 *Suppose Assumptions 7 to 11 hold and the radius $\mu_n(x)$ satisfies $\lim_{n\to\infty}\mu_n(x) = 0$ for a.e. $x \in \mathcal{X}$. Then, for a.e. $x \in \mathcal{X}$, the sequence of objectives $\{\hat{g}^{ER}_{s,n}(\cdot;x)\}$ of the ER-DRO problem (7) converges in probability to the objective $g(\cdot;z)$ of the true problem (3) uniformly on the set $\mathcal{Z}$.*

Proof We wish to show $\sup_{z\in\mathcal{Z}} \left|\hat{g}^{ER}_{s,n}(z;x) - g(z;x)\right| \xrightarrow{p} 0$ for a.e. $x \in \mathcal{X}$. By first adding and subtracting $g^*_n(z;x)$, defined in problem (4), and then doing the same with $g^*_{s,n}(z;x)$, we obtain

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} \left|\hat{g}^{ER}_{s,n}(z;x) - g(z;x)\right| \\
&\leq \sup_{z\in\mathcal{Z}} \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n p_i \left| \sup_{y\in\hat{\mathcal{Y}}^i_n(x;\mu_n(x))} c(z,y) - c\left(z, f^*(x)+\varepsilon^i\right) \right| + \\
&\quad \sup_{z\in\mathcal{Z}} \left|g^*_{s,n}(z;x) - g^*_n(z;x)\right| + \sup_{z\in\mathcal{Z}} |g^*_n(z;x) - g(z;x)|. \qquad (11)
\end{aligned}$$

The third term on the r.h.s. of (11) vanishes in the limit in probability under Assumption 8. We show that the first two terms also converge to zero in probability; the stated result then follows from $o_p(1)+o_p(1) = o_p(1)$.

Consider the first term on the r.h.s. of (11). We have for a.e. $x \in \mathcal{X}$

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n p_i \left| \sup_{y\in\hat{\mathcal{Y}}^i_n(x;\mu_n(x))} c(z,y) - c\left(z, f^*(x)+\varepsilon^i\right) \right| \\
&\leq \sup_{z\in\mathcal{Z}} \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n p_i \sup_{y\in\hat{\mathcal{Y}}^i_n(x;\mu_n(x))} L(z)\|y - (f^*(x)+\varepsilon^i)\| \\
&\leq \sup_{z\in\mathcal{Z}} \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n p_i L(z)\left(\mu_n(x) + \|\hat{\varepsilon}^i_n(x)\|\right) \\
&= \sup_{z\in\mathcal{Z}} L(z) \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n p_i \left(\mu_n(x) + \|\hat{\varepsilon}^i_n(x)\|\right) \\
&\leq \sup_{z\in\mathcal{Z}} L(z)\left(\mu_n(x) + \left(\frac{1}{n}\sum_{i=1}^n \left(\|\hat{\varepsilon}^i_n(x)\|\right)^2\right)^{\frac{1}{2}}\right) \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \left(n\sum_{i=1}^n p_i^2\right)^{\frac{1}{2}} \\
&= \sup_{z\in\mathcal{Z}} L(z)\left(\mu_n(x) + \left(\frac{1}{n}\sum_{i=1}^n \left(\|\hat{\varepsilon}^i_n(x)\|\right)^2\right)^{\frac{1}{2}}\right) \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \left(1 + n\sum_{i=1}^n \left(p_i - \frac{1}{n}\right)^2\right)^{\frac{1}{2}} \\
&= O(1)o_p(1)O(1) = o_p(1), \qquad (12)
\end{aligned}$$

where the first step follows from Assumption 7, the second step follows from the definition of the set $\hat{\mathcal{Y}}^i_n(x;\mu_n(x))$, the triangle inequality, and inequality (6), the fourth step follows by applying the Cauchy-Schwarz inequality twice, and the sixth step follows from Assumptions 7, 9 and 10, $\lim_{n\to\infty}\mu_n(x) = 0$, and [28, Lemma 1].

Next, consider the second term on the r.h.s. of (11). We have for a.e. $x \in \mathcal{X}$

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} |g^*_{s,n}(z;x) - g^*_n(z;x)| \\
&= \sup_{z\in\mathcal{Z}} \left| \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n p_i c(z, f^*(x)+\varepsilon^i) - \frac{1}{n}\sum_{i=1}^n c(z, f^*(x)+\varepsilon^i) \right| \\
&= \sup_{z\in\mathcal{Z}} \left| \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n \left(p_i - \frac{1}{n}\right) c(z, f^*(x)+\varepsilon^i) \right| \\
&\leq \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \left(n\sum_{i=1}^n \left(p_i - \frac{1}{n}\right)^2\right)^{\frac{1}{2}} \sup_{z\in\mathcal{Z}} \left(\frac{1}{n}\sum_{i=1}^n \left(c(z, f^*(x)+\varepsilon^i)\right)^2\right)^{\frac{1}{2}} \\
&= o(1)O_p(1) = o_p(1), \qquad (13)
\end{aligned}$$

where the inequality follows by Cauchy-Schwarz, and the last line follows from Assumptions 10 and 11. □

It is clear from the proof above that Assumptions 10 and 11 are not required for sample robust optimization-based DRO, i.e., when the radius $\zeta_n(x) \equiv 0$.

Remark 3 Assumption 7 can be weakened to a local Lipschitz continuity assumption under stronger assumptions on the regression setup. In particular, when $\zeta_n(x) \equiv 0$, the conclusion of Proposition 9 holds if we replace Assumption 7 with [28, Assumption 2]. When $\zeta_n(x) \neq 0$, we need to replace Assumption 7 with strengthened versions of Assumption 9 and [28, Assumption 2] involving fourth degree terms.

Proposition 9 provides the foundation for showing that the ER-DRO estimators are asymptotically optimal. We omit the proof of Theorem 10 since it is identical to the proof of [28, Theorem 1] in light of Proposition 9.

Theorem 10 (Consistency and asymptotic optimality) *Suppose the assumptions of Proposition 9 hold. Then, for a.e. $x \in \mathcal{X}$*

$$\hat{v}_n^{DRO}(x) \xrightarrow{p} v^*(x), \;\; \mathbb{D}\left(\hat{S}_n^{DRO}(x), S^*(x)\right) \xrightarrow{p} 0, \;\; \sup_{z \in \hat{S}_n^{DRO}(x)} g(z;x) \xrightarrow{p} v^*(x).$$

Next, we investigate the rate of convergence of the optimal value of the ER-DRO problem (7) to that of the true problem (3). To enable this, we require the following rate of convergence assumptions on the FI-SAA problem (3) and the regression estimate $\hat{f}_n$ (cf. Assumptions 5 and 6 of [28]).

Assumption 12 The function c in problem (3) and the data $\mathcal{D}_n$ satisfy the following functional central limit theorem for the FI-SAA objective:

$$\sqrt{n}\,(g_n^*(\cdot;x) - g(\cdot;x)) \xrightarrow{d} V(\cdot;x), \quad \text{for a.e. } x \in \mathcal{X},$$

where $g_n^*(\cdot;x)$, $g(\cdot;x)$, and $V(\cdot;x)$ are (random) elements of $L^\infty(\mathcal{Z})$.

Assumption 13 There is a constant[5] $0 < r \leq 1$ such that the regression estimate $\hat{f}_n$ satisfies the following convergence rate criteria for a.e. $x \in \mathcal{X}$:

$$\|f^*(x) - \hat{f}_n(x)\|^2 = O_p(n^{-r}), \;\; \frac{1}{n}\sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 = O_p(n^{-r}).$$

Assumption 13 strengthens Assumption 9. It typically holds with $r = 1$ for parametric regression methods such as OLS and Lasso regression under mild assumptions. On the other hand, nonparametric regression methods such as kernel regression and random forests usually satisfy Assumption 13 only with $r = O(1)/d_x$ due to the curse of dimensionality.

Our next result establishes a convergence rate for the ER-DRO problem (7).

Theorem 11 (Rate of convergence) *Suppose Assumptions 7, 11, 12, and 13 hold. In addition, suppose Assumption 10 holds with $\rho = 1 + r$ and the radius $\mu_n(x)$ satisfies $\mu_n(x) = O(n^{-r/2})$ for a.e. $x \in \mathcal{X}$, where the constant r is defined in Assumption 13. Then, for a.e. $x \in \mathcal{X}$, the solution of the ER-DRO problem (7) satisfies*

$$\left|\hat{v}_n^{DRO}(x) - v^*(x)\right| = O_p(n^{-r/2}), \;\; \left|g(\hat{z}_n^{DRO}(x);x) - v^*(x)\right| = O_p(n^{-r/2}).$$

Proof Assumptions 7, 10, and 13, $\mu_n = O(n^{-r/2})$, and the inequality chain (12) imply that the first term on the r.h.s. of inequality (11) satisfies

$$\sup_{z \in \mathcal{Z}} \left|\hat{g}_{s,n}^{ER}(z;x) - g_{s,n}^*(z;x)\right| = O_p(n^{-r/2}).$$

Assumptions 10 and 11 and the inequality chain (13) imply that the second term on the r.h.s. of inequality (11) satisfies

$$\sup_{z \in \mathcal{Z}} \left|g_{s,n}^*(z;x) - g_n^*(z;x)\right| = O_p(n^{-r/2}).$$

Finally, Assumption 12 implies $\sup_{z \in \mathcal{Z}} |g_n^*(z;x) - g(z;x)| = O_p(n^{-1/2})$. Putting the above three inequalities together into inequality (11), we obtain

$$\sup_{z \in \mathcal{Z}} \left|\hat{g}_{s,n}^{ER}(z;x) - g(z;x)\right| = O_p(n^{-r/2}), \quad \text{for a.e. } x \in \mathcal{X}.$$

[5] The constant r is independent of n, but could depend on the covariate dimension d_x.

This implies that for a.e. $x \in \mathcal{X}$ and any $\alpha > 0$, there exists $M_\alpha > 0$ such that

$$\mathbb{P}\left\{\sup_{z\in\mathcal{Z}}\left|\hat{g}^{ER}_{s,n}(z;x) - g(z;x)\right| > M_\alpha n^{-r/2}\right\} < \alpha.$$

Consequently, we have for a.e. $x \in \mathcal{X}$

$$\begin{aligned}
\mathbb{P}\left\{\hat{v}^{DRO}_n(x) > v^*(x) + M_\alpha n^{-\frac{r}{2}}\right\} &\le \mathbb{P}\left\{\hat{g}^{ER}_{s,n}(z^*(x);x) > v^*(x) + M_\alpha n^{-\frac{r}{2}}\right\} \\
&\le \mathbb{P}\left\{\left|\hat{g}^{ER}_{s,n}(z^*(x);x) - v^*(x)\right| > M_\alpha n^{-\frac{r}{2}}\right\}, \\
\mathbb{P}\left\{v^*(x) > \hat{v}^{DRO}_n(x) + M_\alpha n^{-\frac{r}{2}}\right\} &\le \mathbb{P}\left\{g(\hat{z}^{DRO}_n(x);x) > \hat{v}^{DRO}_n(x) + M_\alpha n^{-\frac{r}{2}}\right\} \\
&\le \mathbb{P}\left\{\left|\hat{v}^{DRO}_n(x) - g(\hat{z}^{DRO}_n(x);x)\right| > M_\alpha n^{-\frac{r}{2}}\right\}.
\end{aligned}$$

Therefore, $|\hat{v}^{DRO}_n(x) - v^*(x)| = \left|g(\hat{z}^{DRO}_n(x);x) - v^*(x)\right| = O_p(n^{-r/2})$. □

Finally, we make the following assumptions to establish a finite sample certificate-type guarantee for sample robust optimization-based ER-DRO, i.e., when the radius $\zeta_n(x) \equiv 0$. To achieve this, we utilize a connection between sample robust optimization-based ambiguity sets and ambiguity sets defined using the ∞-Wasserstein distance. In particular, Theorem 5 of [10] implies that the sample robust optimization-based ER-DRO problem is equivalent to the ∞-Wasserstein distance-based ER-DRO problem (7) with ambiguity set $\hat{\mathcal{P}}_n(x) := \left\{Q \in \mathcal{P}(\mathcal{Y}) : d_{W,\infty}(Q, \hat{P}^{ER}_n(x)) \le \mu_n(x)\right\}$.

Assumption 14 The regression estimate $\hat{f}_n$ possesses the following finite sample property: for any risk level $\alpha \in (0,1)$, there exists a positive constant $\kappa_n(\alpha, x)$ such that $\mathbb{P}\{\sup_{x\in\mathcal{X}}\|f^*(x) - \hat{f}_n(x)\| > \kappa_n(\alpha,x)\} \le \alpha$.

Assumption 15 For a.e. $x \in \mathcal{X}$, the conditional distribution $P_{Y|X=x}$ has a density $\Lambda_Y(\cdot;x) : \bar{\mathcal{Y}} \to [0,+\infty)$, where $\bar{\mathcal{Y}} \subset \mathcal{Y}$ is an open, connected and bounded set with a Lipschitz boundary. Furthermore, for each $y \in \bar{\mathcal{Y}}$ and a.e. $x \in \mathcal{X}$, the density satisfies $1/\lambda(x) \le \Lambda_Y(y;x) \le \lambda(x)$, for some $\lambda(x) \ge 1$.

Assumption 14 strengthens Assumption 2. Appendix EC.3 of [28] verifies that Assumption 14 holds for some parametric and nonparametric regression methods such as OLS, Lasso, and kNN regression when the support $\mathcal{X}$ of the covariates is compact. Trillos and Slepčev [41] consider cases when Assumption 15 holds. This assumption yields the following concentration of measure result for the true empirical distribution $P^*_n(x)$. Note that Lemma 12 applies to settings with non-i.i.d. data $\mathcal{D}_n$ such as time series data.

Lemma 12 (Theorem 1.1 of [41]) *Suppose Assumption 15 holds and the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. Then, for any constant $\beta > 2$ and a.e. $x \in \mathcal{X}$*

$$\mathbb{P}\left\{d_{W,\infty}(P^*_n(x), P_{Y|X=x}) \ge O(1)\frac{\log(n)}{n^{1/d_y}}\right\} \le O(n^{-\beta/2}),$$

where the $O(1)$ term depends only on β, $\bar{\mathcal{Y}}$, and $\lambda(x)$ in Assumption 15.

The next result is the analogue of Lemma 2 for the ∞-Wasserstein distance.

Lemma 13 *For each $x \in \mathcal{X}$*

$$d_{W,\infty}(\hat{P}^{ER}_n(x), P_{Y|X=x}) \le 2\sup_{x\in\mathcal{X}}\|f^*(x) - \hat{f}_n(x)\| + d_{W,\infty}(P^*_n(x), P_{Y|X=x}).$$

Proof The triangle inequality for the ∞-Wasserstein distance yields

$$d_{W,\infty}(\hat{P}^{ER}_n(x), P_{Y|X=x}) \le d_{W,\infty}(\hat{P}^{ER}_n(x), P^*_n(x)) + d_{W,\infty}(P^*_n(x), P_{Y|X=x}).$$

The result then follows from (6) and the definition of $d_{W,\infty}$, which yield

$$\begin{aligned}
d_{W,\infty}(\hat{P}^{ER}_n(x), P^*_n(x)) &\le \sup_{i\in[n]}\|\mathrm{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{\varepsilon}^i_n) - (f^*(x) + \varepsilon^i)\| \\
&\le \sup_{i\in[n]}\|(\hat{f}_n(x) + \hat{\varepsilon}^i_n) - (f^*(x) + \varepsilon^i)\| \\
&\le 2\sup_{x\in\mathcal{X}}\|f^*(x) - \hat{f}_n(x)\|.
\end{aligned}$$

□

For a given realization $x \in \mathcal{X}$ and risk level $\alpha \in (0,1)$, we hereafter use

$$\zeta_n(\alpha, x) := 0, \quad \mu_n(\alpha, x) := \kappa^{(1)}_{\infty,n}(\alpha, x) + \kappa^{(2)}_{\infty,n}(x) \tag{14}$$

as the radii for the sample robust optimization-based ambiguity set, where

$$\kappa^{(1)}_{\infty,n}(\alpha, x) := 2\kappa_n(\alpha, x), \quad \kappa^{(2)}_{\infty,n}(x) := O(1)n^{-\theta/d_y},$$

the constant κ_n is defined in Assumption 14 and the constant $\theta < 1$. Similar to the specification of the Wasserstein DRO radius in (9), the sample robust optimization radius μ_n equals the sum of two contributions—the first accounts for the error in estimating f^*, and the second corresponds to the radius used in the absence of covariate information [11]. While the above choice of μ_n helps us derive our theoretical guarantees, it involves unknown constants and is typically conservative in practice (cf. Remark 2). We investigate practical approaches for choosing the radius μ_n in Section 6.

Theorem 14 (Finite sample certificate-type guarantee) *Suppose Assumptions 14 and 15 hold, the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d., there exists a sequence of risk levels $\{\alpha_n\}_{n\in\mathbb{N}} \subset (0,1)$ such that $\sum_n \alpha_n < +\infty$, and for a.e. $x \in \mathcal{X}$, $\lim_{n\to\infty} \mu_n(\alpha_n, x) = 0$ with μ_n defined in equation (14). Then, for a.e. $x \in \mathcal{X}$, there exists $N(x) \in \mathbb{N}$ such that the solution of the ER-DRO problem (7) with radii $\zeta_n(\alpha_n, x)$ and $\mu_n(\alpha_n, x)$ specified by equation (14) a.s. satisfies*

$$g(\hat{z}_n^{DRO}(x); x) \le \hat{v}_n^{DRO}(x), \quad \forall n \ge N(x).$$

Proof Our proof follows the outline of the proof of [11, Theorem 1].

Lemma 13, the union bound, and Assumption 14, yield for a.e. $x \in \mathcal{X}$:

$$\begin{aligned}&\mathbb{P}\big\{d_{W,\infty}(\hat{P}_n^{ER}(x), P_{Y|X=x}) > \mu_n(\alpha_n, x)\big\}\\ &\le \alpha_n + \mathbb{P}\big\{d_{W,\infty}(P_n^*(x), P_{Y|X=x}) > \kappa^{(2)}_{\infty,n}(x)\big\}.\end{aligned}$$

Consider $\beta = 4$ in Lemma 12. Because $\kappa^{(2)}_{\infty,n}(x) \ge O(1)\log(n)/n^{1/d_y}$ for n large enough, we have from Lemma 12 that for a.e. $x \in \mathcal{X}$ and n large enough:

$$\mathbb{P}\big\{d_{W,\infty}(\hat{P}_n^{ER}(x), P_{Y|X=x}) > \mu_n(\alpha_n, x)\big\} \le \alpha_n + O(n^{-2}).$$

Therefore, we have $\sum_{n=1}^{\infty} \mathbb{P}\big\{d_{W,\infty}(\hat{P}_n^{ER}(x), P_{Y|X=x}) > \mu_n(\alpha_n, x)\big\} < +\infty$. The Borel-Cantelli lemma then implies that for a.e. $x \in \mathcal{X}$, there a.s. exists $N(x) \in \mathbb{N}$ such that for $n \ge N(x)$, $d_{W,\infty}(\hat{P}_n^{ER}(x), P_{Y|X=x}) \le \mu_n(\alpha_n, x)$.

Recall that our sample robust optimization-based ER-DRO problem is equivalent to the ∞-Wasserstein distance-based ER-DRO problem with ambiguity set $\hat{\mathcal{P}}_n(x) := \big\{Q \in \mathcal{P}(\mathcal{Y}) : d_{W,\infty}(Q, \hat{P}_n^{ER}(x)) \le \mu_n(\alpha_n, x)\big\}$ [10, Theorem 5]. The stated result then follows by the definition of the ∞-Wasserstein distance-based ER-DRO problem (7). □

Hereafter, we revert to the shortened notation $\zeta_n(x)$ and also use it to denote the radius of sample robust optimization ambiguity sets for simplicity.

6 Specifying the radius of the ambiguity set

Determining the optimal radius $\zeta_n(x)$ of the ambiguity sets in Section 3 using the theory in Sections 4 and 5 is hard for two reasons: (i) the theory usually involves unknown constants, and (ii) even if these constants are known or estimated, this specification of $\zeta_n(x)$ is typically conservative in practice (see Remark 2). Therefore, we propose data-driven approaches that use cross-validation (CV) to specify $\zeta_n(x)$ for the ER-DRO problem (7) with the goal of minimizing the out-of-sample cost $g(\hat{z}_n^{DRO}(x); x)$ of the resulting ER-DRO solution $\hat{z}_n^{DRO}(x)$. Once we choose $\zeta_n(x)$, we re-solve the ER-DRO problem (7) with the ambiguity set of radius $\zeta_n(x)$ centered at the empirical distribution $\hat{P}_n^{ER}(x)$ to determine the optimal value $\hat{v}_n^{DRO}(x)$ and a solution $\hat{z}_n^{DRO}(x)$.

We outline two approaches, Algorithms 1 and 2, for choosing the radius $\zeta_n(x)$ independently of the covariate realization $x \in \mathcal{X}$. Algorithm 1 ignores covariate information altogether, whereas Algorithm 2 uses all of the data $\mathcal{D}_n$, including covariates, but does not use the new covariate realization $x \in \mathcal{X}$ for specifying the radius. Algorithm 3 in Appendix C presents an alternative that also uses the realization $x \in \mathcal{X}$ to choose $\zeta_n(x)$. Both classes of algorithms have advantages. Algorithms 1 and 2 are less data and

computation intensive and can be readily used in applications where the DRO problem (7) is repeatedly solved for different covariate realizations. Allowing $\zeta_n(x)$ to depend on the realization $x \in \mathcal{X}$, on the other hand, could yield estimators with better out-of-sample performance, which might justify the added computational cost of Algorithm 3.

Algorithm 1 chooses a covariate-independent radius ζ_n for the ambiguity set $\hat{\mathcal{P}}_n(x)$ using K-fold CV on a DRO extension of a naive SAA problem that does not use covariate information (cf. [19, Section 7.2.2]). This algorithm does not require estimation of the regression function f^*. The parameter ζ_n determined using Algorithm 1 necessarily converges to zero as the sample size n is increased. This may result in suboptimal estimators $\hat{z}_n^{DRO}(x)$ when the prediction model is misspecified, in which case it may be beneficial to use a positive value of ζ_n even for large values of n (cf. Figure 6 in Appendix C). Algorithm 2 determines a covariate-independent radius ζ_n using K-fold CV on ER-DRO problems. Note that the objective in line 12 of Algorithm 2 for choosing the radius ζ_n is similar to the objective in line 8 of Algorithm 1.

Algorithm 1 Specifying a covariate-independent radius ζ_n using a naive SAA-based DRO problem

1: **Input**: data $\mathcal{D}_n$, set of candidate radii Δ, and number of folds K.
2: Partition $[n]$ into K subsets $S_1, \ldots, S_K$ of (roughly) equal size at random.
3: **for** $k = 1, \ldots, K$ **do**
4: **for** $\zeta \in \Delta$ **do**
5: Solve the following DRO problem to get a solution $\hat{z}_{-k}^{DRO}(\zeta)$:

$$\min_{z \in \mathcal{Z}} \sup_{Q \in \hat{\mathcal{P}}_{-k}} \mathbb{E}_{Y \sim Q}\left[c(z, Y)\right],$$

where the ambigiuity set $\hat{\mathcal{P}}_{-k}$ with radius ζ is centered at the empirical distribution $\tilde{P}_{-k} := \dfrac{1}{n - |S_k|} \displaystyle\sum_{i \in [n] \setminus S_k} \delta_{y^i}$.

6: **end for**
7: **end for**
8: **Output**: Radius $\zeta_n \in \arg\min_{\zeta \in \Delta} \dfrac{1}{K} \displaystyle\sum_{k \in [K]} \dfrac{1}{|S_k|} \sum_{i \in S_k} c(\hat{z}_{-k}^{DRO}(\zeta), y^i)$ of the ambiguity set $\hat{\mathcal{P}}_n(x)$ for the ER-DRO problem (7).

Algorithm 2 Specifying a covariate-independent radius ζ_n using the ER-DRO problem

1: **Input**: data $\mathcal{D}_n$, set of candidate radii Δ, number of folds K, and number of covariate realizations sampled during each fold $T \leq \lfloor \frac{n}{K} \rfloor$.
2: Partition $[n]$ into subsets $S_1, \ldots, S_K$ of (roughly) equal size at random. Let $\mathcal{D}_{-k} := \mathcal{D}_n \setminus \{(y^i, x^i)\}_{i \in S_k}$.
3: **for** $k = 1, \ldots, K$ **do**
4: Pick without replacement a random subset $\bar{\mathcal{X}}$ of $\{x^i\}_{i \in S_k}$ of size T.
5: **for** $\bar{x} \in \bar{\mathcal{X}}$ **do**
6: **for** $\zeta \in \Delta$ **do**
7: Fit a regression model $\hat{f}_{-k}$ using the data $\mathcal{D}_{-k}$ and compute its in-sample residuals $\{\hat{\varepsilon}_{-k}^i\}_{i \notin S_k} := \{y^i - \hat{f}_{-k}(x^i)\}_{i \notin S_k}$.
8: Solve the ER-DRO problem below at covariate $\bar{x}$ to get solution $\hat{z}_{-k}^{DRO}(\bar{x}, \zeta)$

$$\min_{z \in \mathcal{Z}} \sup_{Q \in \hat{\mathcal{P}}_{-k}(\bar{x})} \mathbb{E}_{Y \sim Q}\left[c(z, Y)\right],$$

where the ambigiuity set $\hat{\mathcal{P}}_{-k}(\bar{x})$ with radius ζ is centered at the estimated empirical distribution $\hat{P}_{-k}^{ER}(\bar{x}) := \dfrac{1}{n - |S_k|} \displaystyle\sum_{i \notin S_k} \delta_{\hat{f}_{-k}(\bar{x}) + \hat{\varepsilon}_{-k}^i}$.

9: **end for**
10: **end for**
11: **end for**
12: **Output**: Radius $\zeta_n \in \arg\min_{\zeta \in \Delta} \dfrac{1}{T} \displaystyle\sum_{\bar{x} \in \bar{\mathcal{X}}} \dfrac{1}{K} \sum_{k \in [K]} \dfrac{1}{|S_k|} \sum_{i \in S_k} c(\hat{z}_{-k}^{DRO}(\bar{x}, \zeta), y^i)$ for the ambiguity set $\hat{\mathcal{P}}_n(x)$ for the ER-DRO problem (7).

7 Computational experiments

We consider instances of the following mean-risk portfolio optimization model adapted from [19]:

$$\min_{z\in\mathcal{Z}} \mathbb{E}\left[-Y^{\mathrm{T}}z\right] + \rho\,\mathrm{CVaR}_\beta(-Y^{\mathrm{T}}z),$$

where $\mathcal{Z} := \left\{z \in \mathbb{R}^{d_z}_+ : \sum_i z_i = 1\right\}$, ρ and β are given parameters, and

$$\mathrm{CVaR}_\beta(-Y^{\mathrm{T}}z) := \min_{\tau\in\mathbb{R}} \mathbb{E}\left[\tau + \frac{1}{1-\beta}\max\{0, -Y^{\mathrm{T}}z - \tau\}\right].$$

For each $i \in [d_z]$, the decision variable z_i denotes the fraction of capital invested in asset i and the random variable Y_i denotes the net return of asset i. The parameters $\rho \geq 0$ and $\beta \in (0,1)$ specify the decision-maker's risk aversion level, with CVaR_β (roughly) averaging over the $100(1-\beta)\%$ worst return outcomes under the distribution of Y. Following [19], we use $\beta = 0.8$, $\rho = 10$, and $d_y = d_z = 10$.

Similar to [28], we assume that the returns Y can be modeled as

$$Y_j = \nu^*_j + \sum_{l\in\mathcal{L}^*} \mu^*_{jl}(X_l)^\theta + \bar{\varepsilon}_j + \omega, \quad \forall j \in [d_y],$$

where X_l, $l \in \mathcal{L}$ are covariates, $\theta \in \{0.5, 1, 2\}$ is a fixed parameter that determines the model class, $\bar{\varepsilon}_j \sim \mathcal{N}(0, 0.02j)$ and $\omega \sim \mathcal{N}(0, 0.02)$ are additive errors, ν^* and μ^* are model parameters, and $\mathcal{L}^* \subseteq \mathcal{L}$ contains the indices of a subset of covariates with predictive power (note that $\mathcal{L}^*$ does not depend on the index $j \in [d_y]$). Throughout, we assume that $|\mathcal{L}^*| = 3$, i.e., the returns truly depend only on three covariates. We simulate i.i.d. data $\mathcal{D}_n$ with

$$\nu^*_j = 0.01j, \quad \mu^*_{j1} = 0.025j + \xi_{j1}, \quad \mu^*_{j2} = 0.015j + \xi_{j2}, \quad \mu^*_{j3} = 0.01j + \xi_{j3},$$

for each $j \in [d_y]$, where ξ_{j1}, ξ_{j2}, and ξ_{j3} are i.i.d. samples from the uniform distribution $U(-0.005j, 0.005j)$. We draw covariate samples $\{x^i\}_{i=1}^n$ from a multivariate *folded-normal/half-normal* distribution with the underlying normal distribution having zero mean and covariance matrix equal to a random correlation matrix generated using the *vine method* of [32].

Given joint data $\mathcal{D}_n$ on the random returns and random covariates, we estimate the coefficients of the linear model

$$Y_j = \nu_j + \sum_{l\in\mathcal{L}} \mu_{jl}X_l + \eta_j, \quad \forall j \in [d_y],$$

where η_j are zero-mean errors, using OLS or Lasso regression and use this prediction model within our residuals-based formulations. We use this linear prediction model even when the degree $\theta \neq 1$, in which case it is misspecified. Note that OLS regression estimates $d_x + 1$ parameters for each $j \in [d_y]$.

We compare the ER-SAA formulation (5) (denoted by `E`) with ER-DRO formulations that use the 1-Wasserstein-based ambiguity set defined using the ℓ_1-norm (denoted by `W`), the sample robust optimization-based ambiguity set constructed using the ℓ_1-norm (denoted by `S`), and the ambiguity set with the same support as $\hat{P}^{ER}_n(x)$ defined using the Hellinger distance (denoted by `H`, see Example 2 in Section 3). Different from the setup in Section 3, we use the ℓ_1-norm to define the 1-Wasserstein and sample robust optimization-based ambiguity sets so that the resulting ER-DRO problems can be expressed as LPs [19]. Formulation `H` can be expressed as a conic quadratic program [6].

We vary the dimension d_x of the covariates, the sample size n, and the degree θ in our computational experiments. We use Algorithms 1 and 2 to specify the radii ζ_n of the above ambiguity sets for the ER-DRO problem (7) with $K = 5$ folds in both algorithms and $T = \min\{50, \lfloor \frac{n}{5} \rfloor\}$ in Algorithm 2. We investigate the performance of Algorithm 3 in Appendix C. For all ER-DRO formulations, following [19], we choose the radius ζ_n from the set of 28 candidate points $\{b \times 10^e : b \in \{0, 1, \ldots, 9\},\ e \in \{-1, -2, -3\}\}$ instead of $\mathbb{R}_+$.

Solutions obtained from the different approaches are compared by estimating a normalized version of the upper bound of a 99% confidence interval (UCB) on their optimality gaps using the multiple replication procedure (MRP) [33]; see Algorithm 1 in [28] for a detailed description of the MRP in our context. We use 5000 i.i.d. samples from the conditional distribution of Y given $X = x$ to compute these UCBs. Because the data-driven solutions depend on the realization of $\mathcal{D}_n$, we perform 50 data replications per test instance, sample 20 different covariate realizations $x \in \mathcal{X}$, and report our results in the form of box plots of these $50 \times 20 = 1000$ UCBs. The boxes denote the 25^{th}, 50^{th}, and 75^{th} percentiles

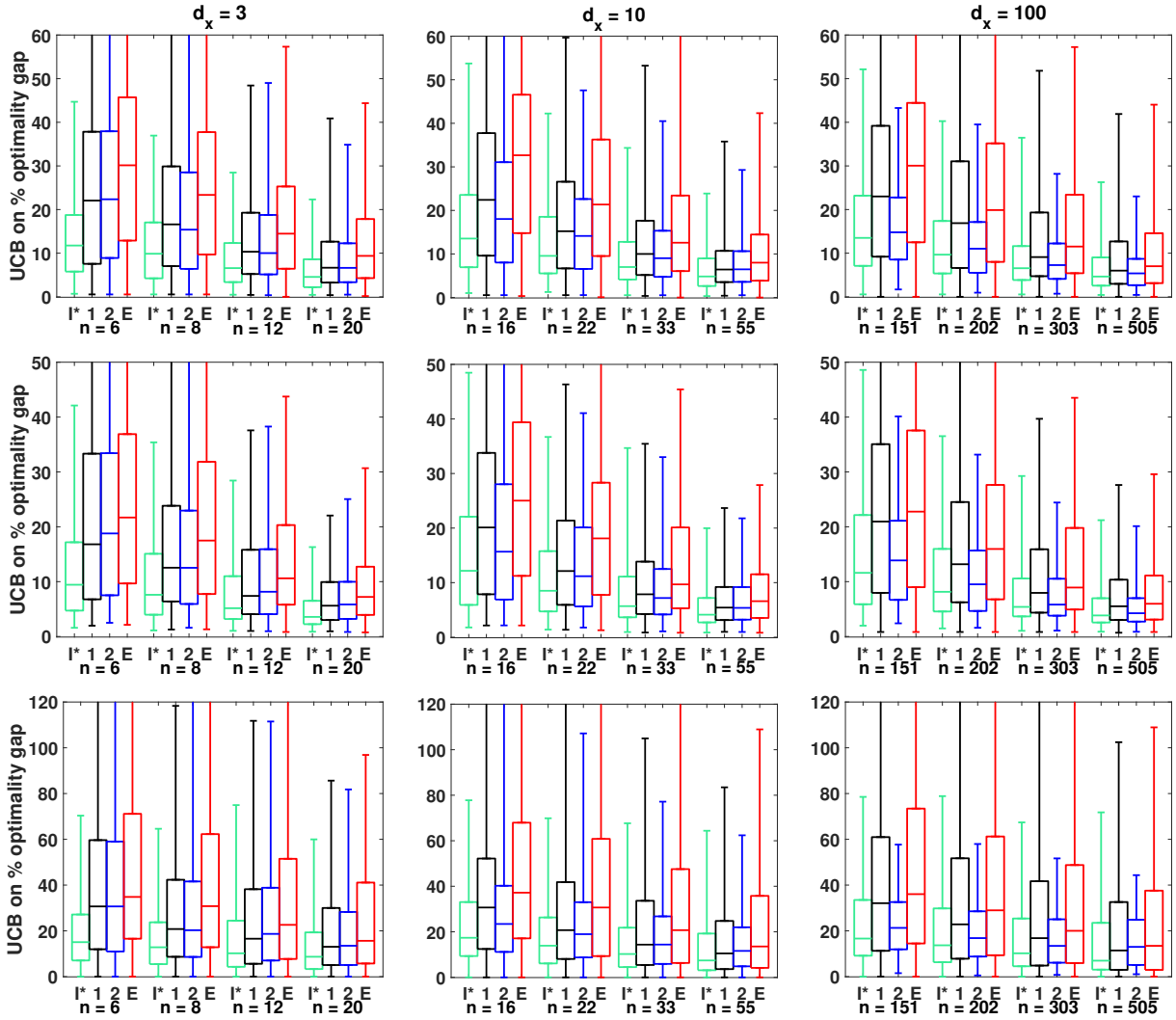

Fig. 1: **(Covariate-independent tuning of the Wasserstein radius)** Comparison of the E+OLS approach (E) with the optimal covariate-independent tuning (I*) of the W+OLS radius and the covariate-independent tuning of the W+OLS radius using Algorithm 1 (1) and Algorithm 2 (2). Top row: $\theta = 1$. Middle row: $\theta = 0.5$. Bottom row: $\theta = 2$. Left column: $d_x = 3$. Middle column: $d_x = 10$. Right column: $d_x = 100$.

of the 99% UCBs, and the whiskers denote the 2nd and 98th percentiles of the 99% UCBs over the 1000 instances.

We compare Algorithms 1 and 2 with an "optimal covariate-independent" specification of ζ_n. This optimal covariate-independent radius is determined by choosing ζ_n such that the medians of the 99% UCBs over the 20 different covariate realizations are minimized. Determining this optimal covariate-independent radius ζ_n is impractical because it requires 5000 i.i.d. samples from the conditional distribution of Y given $X = x$ (which a decision-maker does not have). We consider them only to benchmark the performance of Algorithms 1 and 2.

Source code and data for the test instances will be made available at https://github.com/rohitkannan/DD-DRO. Our codes are written in Julia 0.6.4 [13], use Gurobi 8.1.0 to solve LPs and conic quadratic programs through the JuMP 0.18.5 interface [18], and use `glmnet` 0.3.0 [22] for Lasso regression. All computational tests were conducted through the UW-Madison Center for High Throughput Computing (CHTC) software `HTCondor` (http://chtc.cs.wisc.edu/).

Covariate-independent tuning of the radius. Figure 1 compares the performance of the E+OLS formulation with the W+OLS formulation when the radius ζ_n of the ambiguity set is determined using Algorithms 1 and 2 and optimal covariate-independent tuning. We vary the model degree θ, the covariate dimension among $d_x \in \{3, 10, 100\}$, and the sample size among $n \in \{1.5(d_x + 1), 2(d_x + 1), 3(d_x +$

$1), 5(d_x + 1)\}$ in these experiments. As noted in Section 1, we focus on the small sample size regime. In this regime, the W+OLS formulations perform better than the E+OLS formulation across all cases. The radius specified by Algorithm 2 exhibits better performance than the radius specified using Algorithm 1, with the difference between the performance of Algorithm 1 and 2 accentuated for larger covariate dimensions. The difference between the performance of Algorithm 2 and the optimal covariate-independent tuning of the radius ζ_n reduces with increasing sample size and covariate dimension. Finally, as expected, the benefits of the ER-DRO formulations diminish with increasing sample size.

Comparison of the different DRO formulations. Figure 2 compares the performance of the E+OLS formulation with the W+OLS, S+OLS, and H+OLS formulations over the same range of parameter values as in Figure 1. The radius ζ_n of the ambiguity sets of all three ER-DRO formulations are specified using Algorithm 2. The performance of the S+OLS formulation is similar to that of the W+OLS formulation, whereas the H+OLS formulation hardly performs better than the E+OLS formulation with only a slight improvement for larger covariate dimensions. Recall that the Wasserstein (W) and sample robust optimization (S) ambiguity sets allow distributions with support different from $\hat{P}_n^{ER}(x)$, whereas Hellinger (H) ambiguity set only considers distributions with the same support as $\hat{P}_n^{ER}(x)$. Because the data $\mathcal{D}_n$ comes from a continuous distribution and $\hat{P}_n^{ER}(x)$ may be a crude estimate of $P_n^*(x)$ for small n, this highlights the advantage of DRO formulations that go beyond the estimated empirical distribution $\hat{P}_n^{ER}(x)$. From this point on, we do not include any additional results for the S formulations because they are similar to those of the W formulations. We also do not consider the H formulations any further because they do not perform much better than the ER-SAA formulation.

Impact of the prediction step. Figure 3 compares the performance of the E+Lasso approach with the W+Lasso approach when ζ_n is specified using Algorithm 2. We consider $d_x = 100$, vary the model degree θ, and vary the sample size among $n \in \{0.5(d_x + 1), 0.8(d_x + 1), 1.2(d_x + 1), 1.5(d_x + 1)\}$ in these experiments. We consider smaller sample sizes and larger covariate dimensions because the Lasso is most effective in this regime. These experiments also illustrate the modularity of our residuals-based formulations. The W+Lasso formulation outperforms the E+Lasso formulation for small n. Note that the y-axis limits are different for the different values of θ. Once again, the benefit of the ER-DRO formulation diminishes with increasing sample size.

Wasserstein-DRO certificates. Figure 4 compares normalized versions[6] of the optimal objective value $\hat{v}_n^{ER}(x)$ of the E+OLS formulation with the optimal objective value $\hat{v}_n^{DRO}(x)$ of the W+OLS formulations when the radius ζ_n is specified by Algorithm 2. We consider $d_x = 100$, vary the model degree θ, and vary the sample size among $n \in \{1.5(d_x+1), 2(d_x+1), 3(d_x+1), 5(d_x+1)\}$ in these experiments. We omit the results for smaller covariate dimensions for brevity. First, we see that the ER-SAA solutions are optimistically biased and the bias reduces with increasing sample size (cf. [10, 19, 33]). Second, the mean and the median of the UCBs for the ER-DRO solutions are closer to zero, which implies the ER-DRO formulations reduce the bias of the ER-SAA formulation. This is expected since we chose the radius ζ_n with the goal of reducing the out-of-sample costs of the ER-DRO estimators. Note again that the y-axis limits are different for the different values of θ.

8 Conclusion and future work

We propose a flexible data-driven DRO framework for incorporating covariate information in stochastic optimization when we only have limited concurrent observations of random variables and covariates. We study formulations that build a Wasserstein ambiguity set or an ambiguity set with only discrete distributions on top of a data-driven SAA formulation. Our approach seamlessly generalizes existing DRO formulations that do not use covariate information without sacrificing tractability or favorable theoretical guarantees. We explore data-driven approaches for sizing our ambiguity sets. Numerical experiments illustrate that our residuals-based Wasserstein and sample robust optimization DRO formulations can outperform the ER-SAA formulation in the limited data regime. As with classical DRO formulations, the benefit of our residuals-based DRO formulations diminishes with increasing sample size. This suggests that the added computational cost of the ER-DRO formulations (over the ER-SAA formulation) may not be justified for large sample sizes. Also note that the ER-SAA formulation remains tractable under

[6] We plot $100\left(\frac{\hat{v}_n^{ER}(x) - v^*(x)}{v^*(x)}\right)$ for the ER-SAA formulation and $100\left(\frac{\hat{v}_n^{DRO}(x) - v^*(x)}{v^*(x)}\right)$ for the ER-DRO formulation.

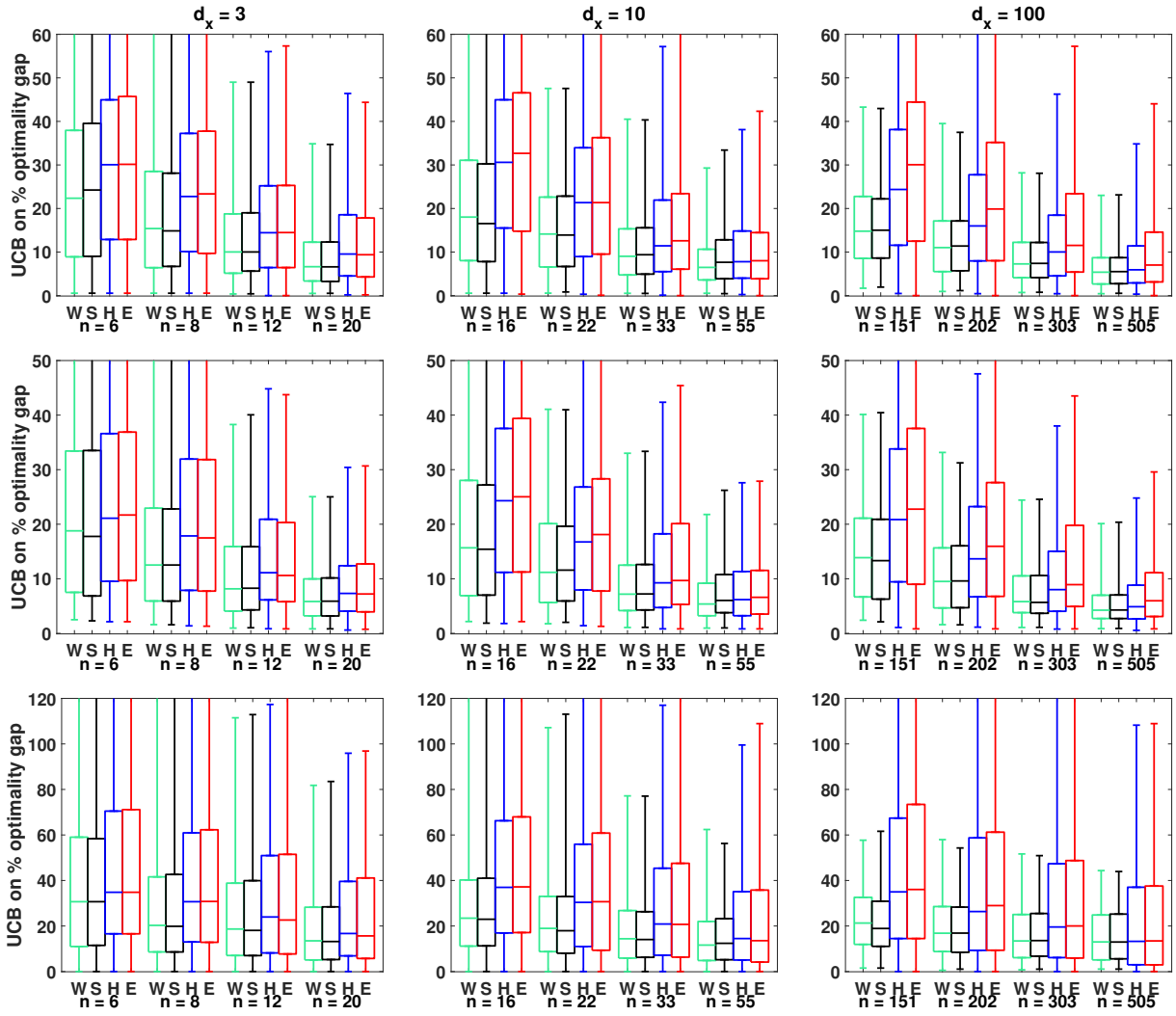

Fig. 2: **(Comparison of the different ER-DRO formulations)** Comparison of the E+OLS approach (E) with the covariate-independent tuning of the W+OLS radius (W), the S+OLS radius (S), and the H+OLS radius (H), all tuned using Algorithm 2. Top row: $\theta = 1$. Middle row: $\theta = 0.5$. Bottom row: $\theta = 2$. Left column: $d_x = 3$. Middle column: $d_x = 10$. Right column: $d_x = 100$.

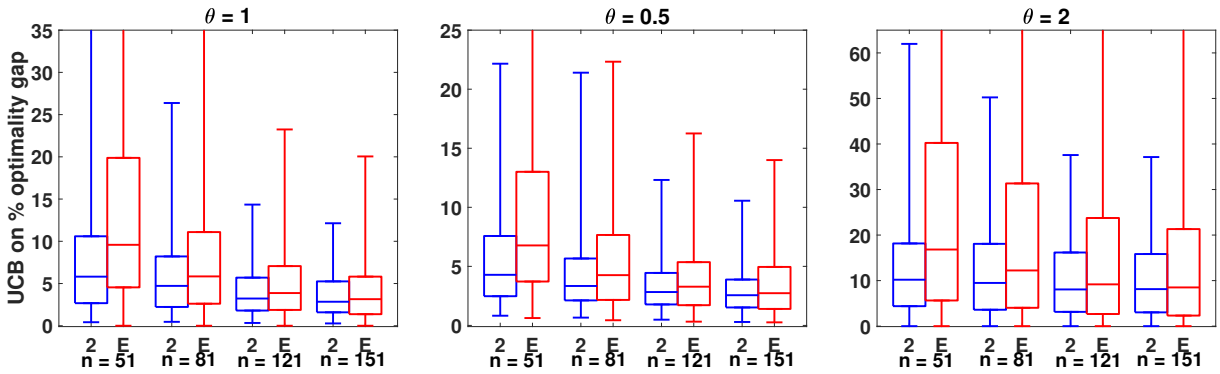

Fig. 3: **(Wasserstein-DRO with the Lasso)** Comparison of the E+Lasso approach (E) with the covariate-independent tuning of the W+Lasso radius using Algorithm 2 (2) for $d_x = 100$. Left: $\theta = 1$. Middle: $\theta = 0.5$. Right: $\theta = 2$.

milder assumptions on the true problem (3) compared to the Wasserstein and sample robust optimization-based ER-DRO formulations (which generally result in NP-hard formulations for two-stage stochastic

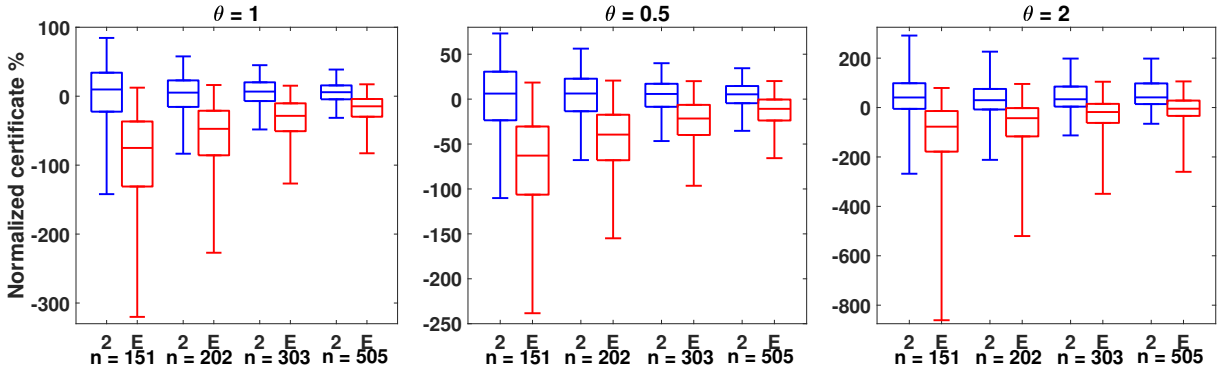

Fig. 4: **(Wasserstein-DRO certificate)** Comparison of the E+OLS approach (E) with the covariate-independent tuning of the W+OLS Wasserstein radius using Algorithm 2 (2) for $d_x = 100$. Left: $\theta = 1$. Middle: $\theta = 0.5$. Right: $\theta = 2$.

programs). However, there is a growing literature on techniques for approximately solving these DRO formulations [29, 37].

Designing residuals-based SAA and DRO formulations that weaken the independence assumption between the covariates X and the errors ε and analyzing their theoretical properties are interesting avenues for future work. Extensions of the ER-SAA and ER-DRO formulations for the multi-stage stochastic programming setting (cf. [9]), for the case when decisions affect the realizations of the random variables (cf. [8]), and for problems with stochastic constraints (cf. [34]) also merit further investigation.

Acknowledgments

This research was performed using the computing resources of the UW-Madison CHTC in the Department of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation. R.K. thanks Nam Ho-Nguyen for helpful discussions.

References

1. Ban, G.Y., Gallien, J., Mersereau, A.J.: Dynamic procurement of new products with covariate information: The residual tree method. Manufacturing & Service Operations Management **21**(4), 798–815 (2019)
2. Ban, G.Y., Rudin, C.: The big data newsvendor: Practical insights from machine learning. Operations Research **67**(1), 90–108 (2018)
3. Bansal, M., Huang, K.L., Mehrotra, S.: Decomposition algorithms for two-stage distributionally robust mixed binary programs. SIAM Journal on Optimization **28**(3), 2360–2383 (2018)
4. Bayraksan, G., Love, D.K.: Data-driven stochastic programming using phi-divergences. In: The Operations Research Revolution, pp. 1–19. INFORMS TutORials in Operations Research (2015)
5. Bazier-Matte, T., Delage, E.: Generalization bounds for regularized portfolio selection with market side information. INFOR: Information Systems and Operational Research pp. 1–28 (2020)
6. Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. Management Science **59**(2), 341–357 (2013)
7. Bertsimas, D., Gupta, V., Kallus, N.: Robust sample average approximation. Mathematical Programming **171**(1-2), 217–282 (2018)
8. Bertsimas, D., Kallus, N.: From predictive to prescriptive analytics. Management Science **66**(3), 1025–1044 (2020)
9. Bertsimas, D., McCord, C., Sturt, B.: Dynamic optimization with side information. arXiv preprint arXiv:1907.07307 pp. 1–37 (2019)
10. Bertsimas, D., Shtern, S., Sturt, B.: A data-driven approach for multi-stage linear optimization. Optimization Online. URL: http://www.optimization-online.org/DB_FILE/2018/11/6907.pdf (2018)
11. Bertsimas, D., Shtern, S., Sturt, B.: Two-stage sample robust optimization. arXiv preprint arXiv:1907.07142 (2019)
12. Bertsimas, D., Van Parys, B.: Bootstrap robust prescriptive analytics. arXiv preprint arXiv:1711.09974 pp. 1–24 (2017)
13. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.: Julia: a fresh approach to numerical computing. SIAM Review **59**(1), 65–98 (2017)
14. Blanchet, J., Kang, Y., Murthy, K.: Robust Wasserstein profile inference and applications to machine learning. Journal of Applied Probability **56**(3), 830–857 (2019)
15. Blanchet, J., Murthy, K., Si, N.: Confidence regions in Wasserstein distributionally robust estimation. arXiv preprint arXiv:1906.01614 (2019)
16. Dou, X., Anitescu, M.: Distributionally robust optimization with correlated data from vector autoregressive processes. Operations Research Letters **47**(4), 294–299 (2019)

17. Duchi, J., Glynn, P., Namkoong, H.: Statistics of robust optimization: A generalized empirical likelihood approach. arXiv preprint arXiv:1610.03425 (2016)
18. Dunning, I., Huchette, J., Lubin, M.: JuMP: A modeling language for mathematical optimization. SIAM Review **59**(2), 295–320 (2017)
19. Esfahani, P.M., Kuhn, D.: Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming **171**(1-2), 115–166 (2018)
20. Esteban-Pérez, A., Morales, J.M.: Distributionally robust stochastic programs with side information based on trimmings. arXiv preprint arXiv:2009.10592 (2020)
21. Fournier, N., Guillin, A.: On the rate of convergence in Wasserstein distance of the empirical measure. Probability Theory and Related Fields **162**(3-4), 707–738 (2015)
22. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**(1), 1–22 (2010)
23. Gao, R.: Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. Optimization Online. URL: http://www.optimization-online.org/DB_HTML/2020/09/8012.html (2020)
24. Gao, R., Chen, X., Kleywegt, A.J.: Wasserstein distributional robustness and regularization in statistical learning. arXiv preprint arXiv:1712.06050 (2017)
25. Gao, R., Kleywegt, A.J.: Distributionally robust stochastic optimization with Wasserstein distance. arXiv preprint arXiv:1604.02199 (2016)
26. Hanasusanto, G.A., Kuhn, D.: Robust data-driven dynamic programming. In: Advances in Neural Information Processing Systems, pp. 827–835 (2013)
27. Hanasusanto, G.A., Kuhn, D.: Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. Operations Research **66**(3), 849–869 (2018)
28. Kannan, R., Bayraksan, G., Luedtke, J.: Data-driven sample average approximation with covariate information. Optimization Online. URL: http://www.optimization-online.org/DB_HTML/2020/07/7932.html (2020)
29. Kuhn, D., Esfahani, P.M., Nguyen, V.A., Shafieezadeh-Abadeh, S.: Wasserstein distributionally robust optimization: Theory and applications in machine learning. In: Operations Research & Management Science in the Age of Analytics, pp. 130–166. INFORMS (2019)
30. Lam, H.: Robust sensitivity analysis for stochastic systems. Mathematics of Operations Research **41**(4), 1248–1275 (2016)
31. Lam, H.: Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. Operations Research **67**(4), 1090–1105 (2019)
32. Lewandowski, D., Kurowicka, D., Joe, H.: Generating random correlation matrices based on vines and extended onion method. Journal of multivariate analysis **100**(9), 1989–2001 (2009)
33. Mak, W.K., Morton, D.P., Wood, R.K.: Monte Carlo bounding techniques for determining solution quality in stochastic programs. Operations Research Letters **24**(1-2), 47–56 (1999)
34. Homem-de Mello, T., Bayraksan, G.: Monte Carlo sampling-based methods for stochastic optimization. Surveys in Operations Research and Management Science **19**(1), 56–85 (2014)
35. Nguyen, V.A., Zhang, F., Blanchet, J., Delage, E., Ye, Y.: Distributionally robust local non-parametric conditional estimation. arXiv preprint arXiv:2010.05373 (2020)
36. Pflug, G., Wozabal, D.: Ambiguity in portfolio selection. Quantitative Finance **7**(4), 435–442 (2007)
37. Rahimian, H., Mehrotra, S.: Distributionally robust optimization: A review. arXiv preprint arXiv:1908.05659 (2019)
38. Rockafellar, R.T., Uryasev, S.: Optimization of conditional value-at-risk. Journal of risk **2**, 21–42 (2000)
39. Sen, S., Deng, Y.: Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming. Optimization Online. URL: http://www.optimization-online.org/DB_FILE/2017/03/5904.pdf (2017)
40. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on stochastic programming: modeling and theory. SIAM (2009)
41. Trillos, N.G., Slepčev, D.: On the rate of convergence of empirical measures in ∞-transportation distance. Canadian Journal of Mathematics **67**(6), 1358–1383 (2015)
42. van der Vaart, A.W., Wellner, J.A.: Weak convergence and empirical processes: with applications to statistics. Springer (1996)
43. Villani, C.: Optimal transport: old and new, vol. 338. Springer Science & Business Media (2008)
44. Xie, W.: Tractable reformulations of two-stage distributionally robust linear programs over the type-∞ Wasserstein ball. Operations Research Letters **48**(4), 513–523 (2020)
45. Xu, H., Caramanis, C., Mannor, S.: A distributional interpretation of robust optimization. Mathematics of Operations Research **37**(1), 95–110 (2012)

A Omitted Proofs

A.1 Proof of Theorem 6

From Theorem 5, we have

$$\mathbb{P}\big\{d_{W,p}(\hat{P}_n^{ER}(x), P_{Y|X=x}) > \zeta_n(\alpha_n, x)\big\} \leq \alpha_n, \quad \text{for a.e. } x \in \mathcal{X}.$$

Adapting the arguments in [19, Lemma 3.7], we a.s. have $\lim_{n\to\infty} d_{W,p}(P_{Y|X=x}, Q_n(x)) = 0$ for a.e. $x \in \mathcal{X}$ for any distribution $Q_n(x) \in \hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x))$. Theorem 6.9 of [43] then implies that $Q_n(x)$ converges weakly to $P_{Y|X=x}$ in the space of distributions with finite pth moments for a.e. $x \in \mathcal{X}$.

Theorem 5 also implies that for a.e. $x \in \mathcal{X}$:

$$\mathbb{P}\big\{v^*(x) \leq g(\hat{z}_n^{DRO}(x); x) \leq \hat{v}_n^{DRO}(x)\big\} \geq 1 - \alpha_n, \quad \forall n \in \mathbb{N}.$$

Since $\sum_n \alpha_n < +\infty$, the Borel-Cantelli lemma a.s. implies that for all n large enough:

$$v^*(x) \leq g(\hat{z}_n^{DRO}(x); x) \leq \hat{v}_n^{DRO}(x), \quad \text{for a.e. } x \in \mathcal{X}.$$

Therefore, to establish $\lim_{n\to\infty} \hat{v}_n^{DRO}(x) = v^*(x) = \lim_{n\to\infty} g(\hat{z}_n^{DRO}(x); x)$ for a.e. $x \in \mathcal{X}$, it suffices to show that $\limsup_{n\to\infty} \hat{v}_n^{DRO}(x) \leq v^*(x)$ for a.e. $x \in \mathcal{X}$.

Fix $\eta > 0$. For a.e. $x \in \mathcal{X}$, let $z^*(x) \in S^*(x)$ be an optimal solution to the true problem (3), and $Q_n^*(x) \in \hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x))$ be such that

$$\sup_{Q \in \hat{\mathcal{P}}_n(x;\zeta_n(\alpha_n,x))} \mathbb{E}_{Y\sim Q}\left[c(z^*(x), Y)\right] \leq \mathbb{E}_{Y\sim Q_n^*(x)}\left[c(z^*(x), Y)\right] + \eta.$$

We have for a.e. $x \in \mathcal{X}$:

$$\begin{aligned}
\limsup_{n\to\infty} \hat{v}_n^{DRO}(x) &\leq \limsup_{n\to\infty} \sup_{Q \in \hat{\mathcal{P}}_n(x;\zeta_n(\alpha_n,x))} \mathbb{E}_{Y\sim Q}\left[c(z^*(x), Y)\right] \\
&\leq \limsup_{n\to\infty} \mathbb{E}_{Y\sim Q_n^*(x)}\left[c(z^*(x), Y)\right] + \eta \\
&= g(z^*(x); x) + \eta = v^*(x) + \eta,
\end{aligned}$$

where the penultimate equality follows from Definition 6.8 and Theorem 6.9 of [43] by virtue of Assumption 3 and the fact that $Q_n^*(x)$ converges weakly to $P_{Y|X=x}$. Since $\eta > 0$ was arbitrary, we conclude that $\limsup_{n\to\infty} \hat{v}_n^{DRO}(x) \leq v^*(x)$ for a.e. $x \in \mathcal{X}$.

Finally, we show that any accumulation point of $\{\hat{z}_n^{DRO}(x)\}$ is almost surely an element of $S^*(x)$ for a.e. $x \in \mathcal{X}$, and argue that this implies $\text{dist}(\hat{z}_n^{DRO}(x), S^*(x)) \xrightarrow{a.s.} 0$ for a.e. $x \in \mathcal{X}$. Note that we a.s. have

$$\liminf_{n\to\infty} g(\hat{z}_n^{DRO}(x); x) \leq \lim_{n\to\infty} \hat{v}_n^{DRO}(x) = v^*(x), \quad \text{for a.e. } x \in \mathcal{X}.$$

Let $\bar{z}(x)$ be an accumulation point of $\hat{z}_n^{DRO}(x)$ for a.e. $x \in \mathcal{X}$. Assume by moving to a subsequence if necessary that $\lim_{n\to\infty} \hat{z}_n^{DRO}(x) = \bar{z}(x)$. We have for a.e. $x \in \mathcal{X}$:

$$v^*(x) \leq g(\bar{z}(x); x) \leq \mathbb{E}\left[\liminf_{n\to\infty} c(\hat{z}_n^{DRO}(x), f^*(x) + \varepsilon)\right] \leq \liminf_{n\to\infty} g(\hat{z}_n^{DRO}(x); x) \leq v^*(x),$$

where the second inequality follows from the lower semicontinuity of $c(\cdot, Y)$ on $\mathcal{Z}$ for each $Y \in \mathcal{Y}$ and the third inequality follows from Fatou's lemma by virtue of Assumption 3. Consequently, we a.s. have that $\bar{z}(x) \in S^*(x)$.

Suppose by contradiction that $\text{dist}(\hat{z}_n^{DRO}(x), S^*(x))$ does not a.s. converge to zero for a.e. $x \in \mathcal{X}$. Then, there exists $\bar{\mathcal{X}} \subseteq \mathcal{X}$ with $P_X(\bar{\mathcal{X}}) > 0$ such that for each $x \in \bar{\mathcal{X}}$, $\text{dist}(\hat{z}_n^{DRO}(x), S^*(x))$ does not a.s. converge to zero. Since $\mathcal{Z}$ is compact, any sequence of estimators $\{\hat{z}_n^{DRO}(x)\}$ has a convergent subsequence for each $x \in \bar{\mathcal{X}}$. Therefore, whenever $\text{dist}(\hat{z}_n^{DRO}(x), S^*(x))$ does not converge to zero for some $x \in \bar{\mathcal{X}}$ and a realization of the data $\mathcal{D}_n$, there exists an accumulation point of the sequence $\{\hat{z}_n^{DRO}(x)\}$ that is not a solution to problem (3). This contradicts the fact that every accumulation point of $\{\hat{z}_n^{DRO}(x)\}$ is almost surely a solution to problem (3) for a.e. $x \in \mathcal{X}$. □

A.2 Proof of Theorem 7

From the proof of Theorem 6, we a.s. have for all n large enough that

$$v^*(x) \leq g(\hat{z}_n^{DRO}(x); x) \leq \hat{v}_n^{DRO}(x), \quad \text{for a.e. } x \in \mathcal{X}.$$

Adapting the arguments in [19, Lemma 3.7], we a.s. have for any distribution $Q_n(x) \in \hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x))$ and n large enough that $d_{W,p}(P_{Y|X=x}, Q_n(x)) \leq 2\zeta_n(\alpha_n, x)$ for a.e. $x \in \mathcal{X}$.

Let $z^*(x) \in S^*(x)$ be an optimal solution to the true problem (3) for a.e. $x \in \mathcal{X}$. Suppose Assumption 4 holds. Define

$$\bar{\mathcal{P}}_{1,n}(x; \zeta_n(\alpha_n, x)) := \big\{Q \in \mathcal{P}(\mathcal{Y}) : d_{W,1}(Q, P_{Y|X=x}) \leq 2\zeta_n(\alpha_n, x)\big\}.$$

Using the fact that $\hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x)) \subseteq \bar{\mathcal{P}}_{1,n}(x; \zeta_n(\alpha_n, x))$ for all orders $p \in [1, +\infty)$, we a.s. have for n large enough and for a.e. $x \in \mathcal{X}$ that

$$\hat{v}_n^{DRO}(x) \leq \sup_{Q \in \bar{\mathcal{P}}_{1,n}(x;\zeta_n(\alpha_n,x))} \mathbb{E}_{Y\sim Q}\left[c(z^*(x), Y)\right] \leq g(z^*(x); x) + 2L_1(z^*(x))\zeta_n(\alpha_n, x), \tag{15}$$

where the second inequality follows from Assumption 4 and the Kantorovich-Rubinstein theorem (cf. [29, Theorem 5]). The desired result follows.

Suppose instead that Assumption 5 holds and $p \in [2, +\infty)$. Define

$$\bar{\mathcal{P}}_{2,n}(x; \zeta_n(\alpha_n, x)) := \{Q \in \mathcal{P}(\mathcal{Y}) : d_{W,2}(Q, P_{Y|X=x}) \leq 2\zeta_n(\alpha_n, x)\}.$$

Since $\hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x)) \subseteq \bar{\mathcal{P}}_{2,n}(x; \zeta_n(\alpha_n, x))$ for all orders $p \in [2, +\infty)$, we a.s. have for n large enough and a.e. $x \in \mathcal{X}$ that

$$\begin{aligned} \hat{v}_n^{DRO}(x) &\leq \sup_{Q \in \bar{\mathcal{P}}_{2,n}(x; \zeta_n(\alpha_n, x))} \mathbb{E}_{Y \sim Q}\left[c(z^*(x), Y)\right] \\ &\leq g(z^*(x); x) + 2\left(\mathbb{E}\left[\|\nabla c(z^*(x), Y)\|^2\right]\right)^{1/2} \zeta_n(\alpha_n, x) + 4L_2(z^*(x))\zeta_n^2(\alpha_n, x), \end{aligned} \tag{16}$$

where the second inequality follows from Assumption 5 and [23, Lemma 2] (see also [24]). The desired result then readily follows. □

B Ambiguity sets satisfying Assumption 10

Lemma 13 of [17] (cf. [6, 30, 31]) shows that for phi-divergence ambiguity sets $\mathfrak{P}_n(x; \zeta_n(x))$ constructed using a twice continuously differentiable and strictly convex divergence function ϕ with $\phi'(1) = 0$ (these conditions are satisfied by most of the divergence functions listed in [6, Table 2]), we have

$$\sup_{p \in \mathfrak{P}_n(x; \zeta_n(x))} \sum_{i=1}^n \left(p_i - \frac{1}{n}\right)^2 = \Theta\left(\frac{\zeta_n(x)}{n}\right).$$

Consequently, Assumption 10 holds for such phi-divergence-based ambiguity sets $\mathfrak{P}_n(x; \zeta_n(x))$ whenever the radius $\zeta_n(x) = O(n^{1-\rho})$. Clearly, this bound on $\zeta_n(x)$ is *sharp* in the sense that Assumption 10 does not hold if $\zeta_n(x)$ grows faster than $n^{1-\rho}$ asymptotically. Lemma 15 determines sharp bounds on the radius $\zeta_n(x)$ for some other families of ambiguity sets to satisfy Assumption 10. First, we introduce a third example of the ambiguity set $\mathfrak{P}_n(x; \zeta_n(x))$ to add to Examples 1 and 2 in Section 3.

Example 3 Mean-upper-semideviation-based ambiguity sets [40]: given order $a \in [1, +\infty)$ and radius $\zeta_n(x) \geq 0$, let $b := a/(a-1)$ and define $\hat{\mathcal{P}}_n(x)$ using

$$\begin{aligned} \mathfrak{P}_n(x; \zeta_n(x)) := \Bigg\{ p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1 \text{ and } \exists\, q \in \mathbb{R}_+^n \text{ such that } \|q\|_b \leq \zeta_n(x), \\ p_i = \frac{1}{n}\left[1 + q_i - \frac{1}{n}\sum_{j=1}^n q_j\right], \forall i \in [n] \Bigg\}. \end{aligned}$$

Lemma 15 *The following ambiguity sets satisfy Assumption 10 with constant $\rho \in (1, 2]$:*

(a) CVaR-based ambiguity sets (see Example 1) with radius $\zeta_n(x) = O(n^{1-\rho})$,
(b) Variation distance-based ambiguity sets (see Example 2) with radius $\zeta_n(x) = O(n^{-\rho/2})$,
(c) Mean-upper-semideviation-based ambiguity sets of order $a \in [1, +\infty)$ (see Example 3) with radius

$$\zeta_n(x) = \begin{cases} O(n^{1-\rho/2}) & \text{if } a \geq 2 \\ O(n^{3/2-1/a-\rho/2}) & \text{if } a < 2 \end{cases}.$$

Furthermore, these bounds are sharp in the sense described above.

Proof (a) Assume that the radius $\zeta_n(x) < 0.5$. We begin by noting that there exists an optimal solution to the problem $\sup_{p \in \mathfrak{P}_n(x; \zeta_n(x))} \sum_{i=1}^n \left(p_i - \frac{1}{n}\right)^2$ that is an extreme point of the polytope $\mathfrak{P}_n(x; \zeta_n(x))$. Furthermore, every extreme point of $\mathfrak{P}_n(x; \zeta_n(x))$ satisfies at least $n-1$ of the set of $2n$ inequalities $\left\{p_i \geq 0,\ i \in [n],\ p_i \leq \frac{1}{n(1-\zeta_n(x))},\ i \in [n]\right\}$, with equality. This implies that there exists an optimal solution at which at least $n-1$ of the p_is either take the value zero, or take the value $\frac{1}{n(1-\zeta_n(x))}$. At this solution, $n-1$ of the terms $\left(p_i - \frac{1}{n}\right)^2$ are either $\frac{1}{n^2}$ or $\frac{1}{n^2}\left(\frac{\zeta_n(x)}{1-\zeta_n(x)}\right)^2$ (with $\frac{1}{n^2}$ larger since $\zeta_n(x) < 0.5$ by assumption).

Suppose $M \in \{0, \ldots, n-1\}$ of the inequalities $p_i \geq 0$, $i \in [n]$, are satisfied with equality at such an optimal solution. Since $\sum_{i=1}^n p_i = 1$ and $p_i \leq \frac{1}{n(1-\zeta_n(x))}$, $\forall i \in [n]$, we require $(n-M)\frac{1}{n(1-\zeta_n(x))} \geq 1$, which implies $M \leq n\zeta_n(x)$. Consequently, $M \leq n\zeta_n(x) < n/2$ of the inequalities $p_i \geq 0$, $i \in [n]$, are satisfied with equality and at least $(n-1-M) \geq n(1-\zeta_n(x)) - 1 > n/2 - 1$ of the inequalities $p_i \leq \frac{1}{n(1-\zeta_n(x))}$, $i \in [n]$, are satisfied with equality. Therefore, whenever $\zeta_n(x) < 0.5$, we have:

$$\begin{aligned} \sum_{i=1}^n \left(p_i - \frac{1}{n}\right)^2 &\leq (n\zeta_n(x) + 1)\frac{1}{n^2} + n(1-\zeta_n(x))\frac{1}{n^2}\left(\frac{\zeta_n(x)}{1-\zeta_n(x)}\right)^2 \\ &= \frac{1}{n^2} + \frac{1}{n}\left(\frac{\zeta_n(x)}{1-\zeta_n(x)}\right). \end{aligned}$$

Because the above analysis is constructive, it can be immediately used to deduce that the bound on $\zeta_n(x)$ is sharp.

(b) VD-based ambiguity sets: the stated result follows from the fact that

$$\sum_{i=1}^{n}\left(p_i-\frac{1}{n}\right)^2 \le \left(\sum_{i=1}^{n}\left|p_i-\frac{1}{n}\right|\right)^2 \le \zeta_n^2(x), \quad \forall p\in\mathfrak{P}_n(x;\zeta_n(x)),\ x\in\mathcal{X}.$$

To see that the above bound is sharp, assume without loss of generality that $n\ge 2$ and $\zeta_n(x)\le 1$. Then, because

$$\left(\frac{1}{n}+\frac{\zeta_n(x)}{2}, \underbrace{\frac{1}{n}-\frac{\zeta_n(x)}{2n-2},\ldots,\frac{1}{n}-\frac{\zeta_n(x)}{2n-2}}_{n-1 \text{ terms}}\right)\in\mathfrak{P}_n(x;\zeta_n(x)),$$

we have

$$\sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))}\sum_{i=1}^{n}\left(p_i-\frac{1}{n}\right)^2 \ge \frac{\zeta_n^2(x)}{4}+\frac{\zeta_n^2(x)}{4(n-1)}.$$ □

(c) mean-upper-semideviation-based ambiguity sets of order $a\in[1,+\infty)$: Let $\bar{q}:=\frac{1}{n}\sum_{i=1}^{n}q_i$. We have:

$$\sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))}\sum_{i=1}^{n}\left(p_i-\frac{1}{n}\right)^2 \le \sup_{q\in\mathfrak{Q}_n(x;\zeta_n(x))}\frac{1}{n^2}\sum_{i=1}^{n}(q_i-\bar{q})^2,$$

where $\mathfrak{Q}_n(x;\zeta_n(x)) := \left\{q\in\mathbb{R}_+^n : \|q\|_b\le\zeta_n(x)\right\}$. Note that for each $q\in\mathfrak{Q}_n(x;\zeta_n(x))$, we have $|\bar{q}|\le n^{-1}\|q\|_1\le n^{-1/b}\|q\|_b$, which in turn implies

$$\|q-\bar{q}\mathbf{1}\|_b\le\|q\|_b+|\bar{q}|\|\mathbf{1}\|_b=\|q\|_b+|\bar{q}|n^{1/b}\le\|q\|_b+\|q\|_b\le 2\zeta_n(x),$$

where $\mathbf{1}$ is a vector of ones of appropriate dimension. Additionally, note that

$$\sum_{i=1}^{n}(q_i-\bar{q})^2=\|q-\bar{q}\mathbf{1}\|^2\le\begin{cases}\|q-\bar{q}\mathbf{1}\|_b^2 & \text{if } b\le 2\\ n^{1-2/b}\|q-\bar{q}\mathbf{1}\|_b^2 & \text{if } b>2\end{cases}.$$

The desired result then follows from

$$\begin{aligned}\sup_{q\in\mathfrak{Q}_n(x;\zeta_n(x))}\frac{1}{n^2}\sum_{i=1}^{n}(q_i-\bar{q})^2 &\le \sup_{\{q:\|q-\bar{q}\mathbf{1}\|_b\le 2\zeta_n(x)\}}\frac{1}{n^2}\|q-\bar{q}\mathbf{1}\|^2\\ &\le\begin{cases}\frac{4}{n^2}\zeta_n^2(x) & \text{if } b\le 2\\ \frac{4}{n^{1+2/b}}\zeta_n^2(x) & \text{if } b>2\end{cases}.\end{aligned}$$

We now show that the above bounds are sharp.

Consider first the case when $b\le 2$ and assume without loss of generality that $\zeta_n(x)=O(\sqrt{n})$. Note that $p_i=\frac{1}{n}\left[1+q_i-\frac{1}{n}\sum_{j=1}^{n}q_j\right]$, $i\in[n]$, with $q_1=\zeta_n(x)$ and $q_i=0$, $\forall i\ge 2$, is an element of $\mathfrak{P}_n(x;\zeta_n(x))$. Therefore

$$\sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))}\sum_{i=1}^{n}\left(p_i-\frac{1}{n}\right)^2=\Theta\left(\frac{\zeta_n^2(x)}{n^2}\right).$$

Next, suppose instead that $b>2$ and assume without loss of generality that $\zeta_n(x)=O(n^{1/b})$. Note that $p_i=\frac{1}{n}\left[1+q_i-\frac{1}{n}\sum_{j=1}^{n}q_j\right]$ with $q_i=\begin{cases}\left(\frac{2}{n}\right)^{1/b}\zeta_n(x) & \text{if } i\equiv 0 \pmod 2\\ 0 & \text{if } i\equiv 1 \pmod 2\end{cases}$, $i\in[n]$, is an element of $\mathfrak{P}_n(x;\zeta_n(x))$. Therefore

$$\sup_{p\in\mathfrak{P}_n(x;\zeta_{1,n}(x))}\sum_{i=1}^{n}\left(p_i-\frac{1}{n}\right)^2=\Theta\left(\frac{\zeta_n^2(x)}{n^{1+2/b}}\right).$$ □

C Additional computational results

We first introduce Algorithm 3 that determines a covariate-dependent radius $\zeta_n(x)$ using K-fold CV on the ER-DRO problem (7). For each fold, this algorithm estimates the regression function f^* twice: once using the data omitted in the fold for setting up the ER-DRO problem (7), and once using the data in the fold for estimating the out-of-sample costs of the constructed DRO solutions. Clearly, there is a trade-off between the number of data samples used to construct each estimate of f^*. Because we are particularly interested in the limited data regime, we propose to use a sparse estimation technique (such as the Lasso) for the second estimation step (i.e., for line 7 of Algorithm 3).

For the numerical experiments in this section, we use Lasso regression in line 7 of Algorithm 3 and 5-fold CV (i.e., $K=5$). Similar to Algorithms 1 and 2, we choose the radius $\zeta_n(x)$ from the set of 28 candidate points $\{b\times 10^e : b\in\{0,1,\ldots,9\},\ e\in\{-1,-2,-3\}\}$ instead of $\mathbb{R}_+$. We benchmark Algorithm 3 against the "optimal covariate dependent" specification of $\zeta_n(x)$ that is determined by choosing $\zeta_n(x)$ such that the 99% UCBs are minimized. We again stress that determining this optimal radius $\zeta_n(x)$ is impractical because it requires 5000 i.i.d. samples from the conditional distribution of Y given $X=x$, which a decision-maker does not have.

Covariate-dependent tuning of the radius. Figure 5 compares the performance of the E+OLS formulation with the W+OLS formulations when the radius $\zeta_n(x)$ of the ambiguity set is determined using Algorithms 2 and 3 and optimal

Algorithm 3 Specifying a covariate-dependent radius $\zeta_n(x)$ using the ER-DRO problem

1: **Input**: data $\mathcal{D}_n$, set of candidate radii Δ, number of folds K, and new covariate realization $x \in \mathcal{X}$.
2: Partition $[n]$ into subsets $S_1, \ldots, S_K$ of (roughly) equal size at random. Let $\mathcal{D}_{-k} := \mathcal{D}_n \backslash \{(y^i, x^i)\}_{i \in S_k}$.
3: **for** $k = 1, \ldots, K$ **do**
4: **for** $\zeta \in \Delta$ **do**
5: Fit a regression model $\hat{f}_{-k}$ using the data $\mathcal{D}_{-k}$ and compute its in-sample residuals $\{\hat{\varepsilon}^i_{-k}\}_{i \notin S_k} := \{y^i - \hat{f}_{-k}(x^i)\}_{i \notin S_k}$.
6: Solve the ER-DRO problem below at covariate x to obtain solution $\hat{z}^{DRO}_{-k}(x, \zeta)$

$$\min_{z \in \mathcal{Z}} \sup_{Q \in \hat{\mathcal{P}}_{-k}(x)} \mathbb{E}_{Y \sim Q}\left[c(z, Y)\right],$$

where the ambigiuity set $\hat{\mathcal{P}}_{-k}(x)$ with radius ζ is centered at the estimated empirical distribution $\hat{P}^{ER}_{-k}(x) := \dfrac{1}{n - |S_k|} \displaystyle\sum_{i \notin S_k} \delta_{\hat{f}_{-k}(x) + \hat{\varepsilon}^i_{-k}}$.

7: Fit a regression model $\hat{f}_k$ using the data $\{(y^i, x^i)\}_{i \in S_k}$ and compute its in-sample residuals $\{\hat{\varepsilon}^i_k\}_{i \in S_k} := \{y^i - \hat{f}_k(x^i)\}_{i \in S_k}$.
8: **end for**
9: **end for**
10: **Output**: Radius $\zeta_n(x) \in \arg\min_{\zeta \in \Delta} \dfrac{1}{K} \displaystyle\sum_{k \in [K]} \dfrac{1}{|S_k|} \sum_{i \in S_k} c(\hat{z}^{DRO}_{-k}(x, \zeta), \hat{f}_k(x) + \hat{\varepsilon}^i_k)$ for the ambiguity set $\hat{\mathcal{P}}_n(x)$ for the ER-DRO problem (7).

covariate-dependent tuning. We vary the model degree θ, the covariate dimension among $d_x \in \{3, 10, 100\}$, and the sample size among $n \in \{8(d_x+1), 10(d_x+1), 12(d_x+1), 15(d_x+1)\}$ for $d_x = 3$, among $n \in \{3(d_x+1), 4(d_x+1), 5(d_x+1), 10(d_x+1)\}$ for $d_x = 10$, and among $n \in \{1.5(d_x+1), 2(d_x+1), 3(d_x+1), 5(d_x+1)\}$ for $d_x = 100$ in these experiments[7]. The ER-DRO formulations perform better than the ER-SAA+OLS approach across all the cases. The radius specified by Algorithm 2 performs slightly better than the radius specified using Algorithm 3 for most cases and more so at smaller sample sizes. The difference between the performance of Algorithm 3 and the ideal covariate-dependent tuning of the radius persists even with increasing covariate dimension and increasing sample size. These results indicate that Algorithm 3 requires more data to obtain a good estimate of the optimal covariate-dependent radius $\zeta_n(x)$.

Comparison of the radii specified by Algorithms 1, 2, and 3. Figure 6 compares the radii specified by Algorithms 1, 2, and 3 with the optimal covariate-dependent radius and optimal covariate-independent radius for the W+OLS formulation. We consider $d_x = 100$, vary the model degree θ, and vary the sample size among $n \in \{1.5(d_x+1), 2(d_x+1), 3(d_x+1), 5(d_x+1)\}$ in these experiments. First, note that the radius specified by Algorithm 1 shrinks very quickly to zero for all three values of θ. Consequently, we note from Figure 1 that the resulting ER-DRO estimators do not perform much better than the corresponding ER-SAA estimators. Second, we see that the covariate-independent specifications of the radii result in more narrow distributions compared to the covariate-dependent specifications. This may be because the covariate-independent specifications of the radius attempt to choose a single value of $\zeta_n(x)$ for all possible covariate realizations $x \in \mathcal{X}$, whereas the covariate-dependent specifications can choose a different value of $\zeta_n(x)$ depending on the realization $x \in \mathcal{X}$. Third, the distribution of the radii determined using Algorithm 3 converges to the distribution of the optimal covariate-dependent radii as the sample size increases. Similarly, the distribution of the radii determined using Algorithm 2 converges to the distribution of the optimal covariate-independent radii as n increases (except for the case when $\theta = 2$ because the optimal solution on line 12 of Algorithm 2 is not unique). Finally, as noted in Section 6, it may be advantageous to use a positive radius for the ambiguity set when the prediction model is misspecified (e.g., using OLS regression even when $\theta \neq 1$—the true dependence of Y on X is nonlinear in this case). This is corroborated by the plots for $\theta = 2$, where the distribution of the optimal covariate-dependent radius is far from the zero distribution even for large sample sizes n (note that the y-axis limits for $\theta = 2$ are different from those for $\theta = 1$ and $\theta = 0.5$).

Comparison with the Jackknife-based formulations. Figure 7 compares the performance of the ER-SAA+OLS and J-SAA+OLS approaches with the W+OLS formulations when the radius $\zeta_n(x)$ is specified using Algorithms 2 and 3. We consider $d_x = 100$, vary the model degree θ, and vary the sample size among $n \in \{1.3(d_x+1), 1.5(d_x+1), 2(d_x+1), 3(d_x+1)\}$ in these experiments. As observed in [28], the J-SAA+OLS formulation performs better than the E+OLS formulation in the small sample size regime. Figure 7 shows that the W+OLS formulations outperform the J-SAA+OLS formulation. This is to be expected because the ER-DRO formulations account for both the errors in the approximation of f^* by $\hat{f}_n$ *and* in the approximation of $P_{Y|X=x}$ by $P^*_n(x)$, whereas the J-SAA+OLS formulation only addresses the bias in the residuals obtained from OLS regression (i.e., even if $\hat{f}_n$ is an accurate estimate of f^*, the J-SAA+OLS formulation does not account for the fact that $P^*_n(x)$ may be a crude approximation of $P_{Y|X=x}$). We omit the results for the J+-SAA+OLS formulation because they are similar to those for the J-SAA+OLS formulation.

[7] We use these sample sizes since Algorithm 3 with 5-fold CV requires at least 30 samples. This is because line 7 of Algorithm 3 needs at least 6 points for Lasso regression with CV.

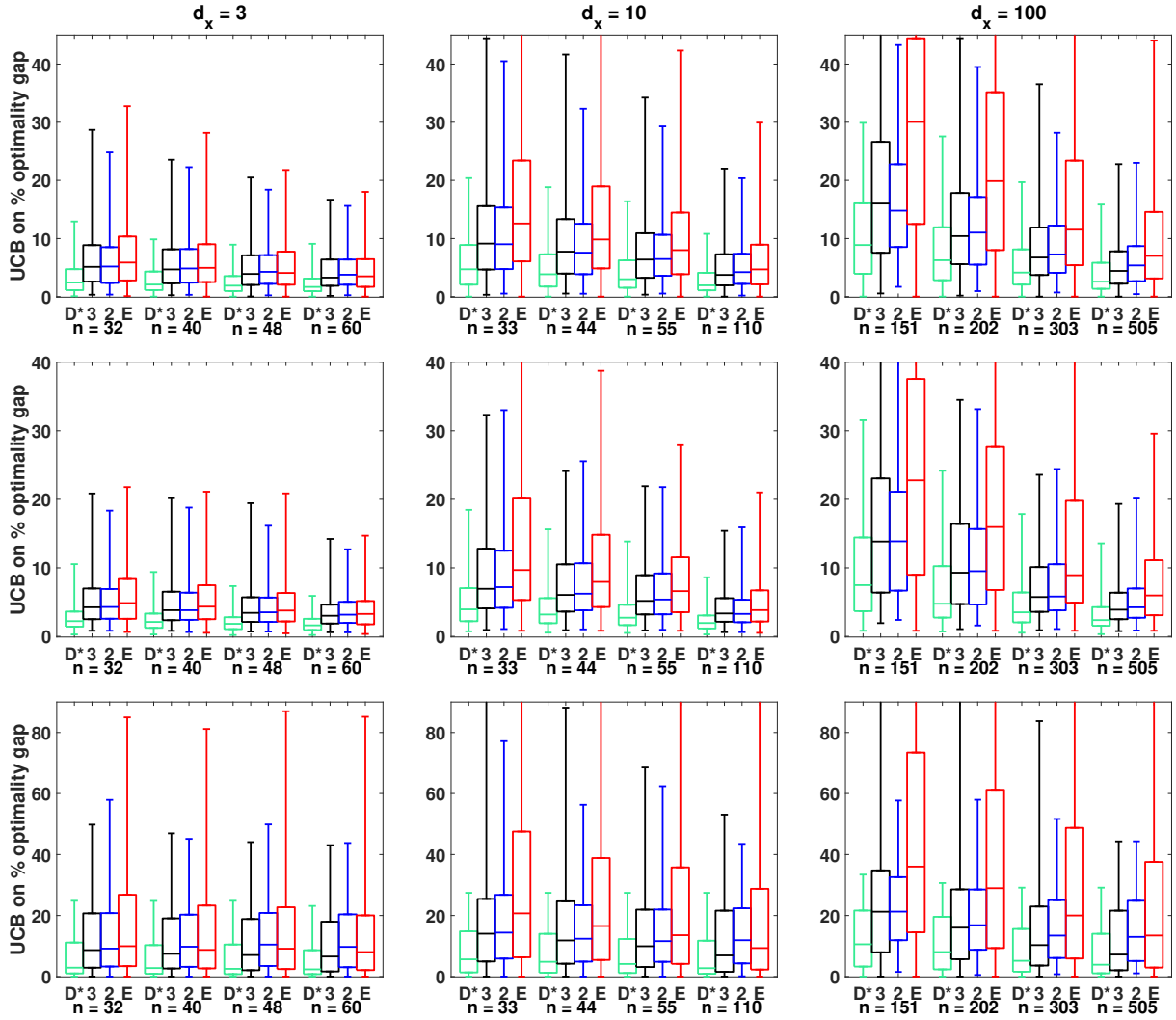

Fig. 5: **(Covariate-dependent tuning of the Wasserstein radius)** Comparison of the E+OLS approach (E) with the optimal covariate-dependent tuning (D*) of the W+OLS radius, the covariate-dependent tuning of the W+OLS radius using Algorithm 3 (3), and the covariate-independent tuning of the W+OLS radius using Algorithm 2 (2). Top row: $\theta = 1$. Middle row: $\theta = 0.5$. Bottom row: $\theta = 2$. Left column: $d_x = 3$. Middle column: $d_x = 10$. Right column: $d_x = 100$.

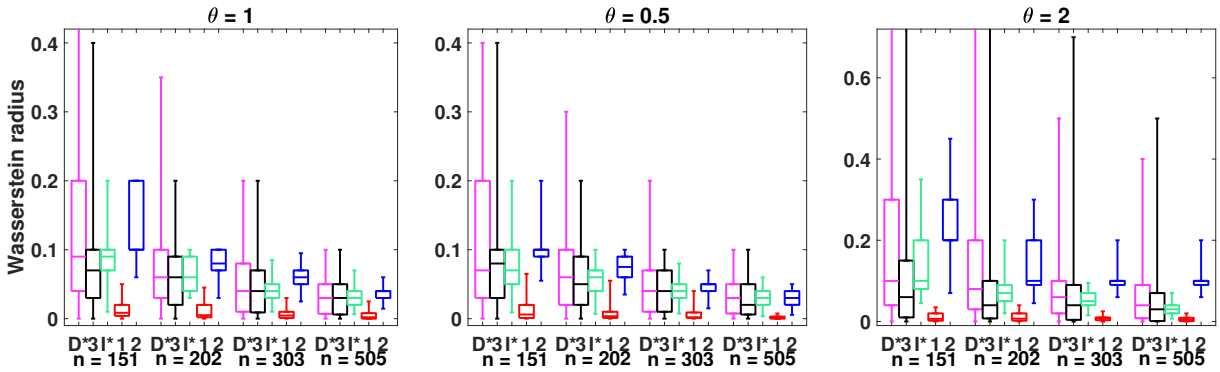

Fig. 6: **(Comparison of the radii specified by Algorithms 1, 2, and 3)** Comparison of the optimal covariate-dependent tuning (D*) of the W+OLS radius, the optimal covariate-independent tuning (I*) of the W+OLS radius, the covariate-dependent tuning of the W+OLS radius using Algorithm 3 (3), and the covariate-independent tuning of the W+OLS radius using Algorithm 1 (1) and Algorithm 2 (2) for $d_x = 100$. Left: $\theta = 1$. Middle: $\theta = 0.5$. Right: $\theta = 2$.

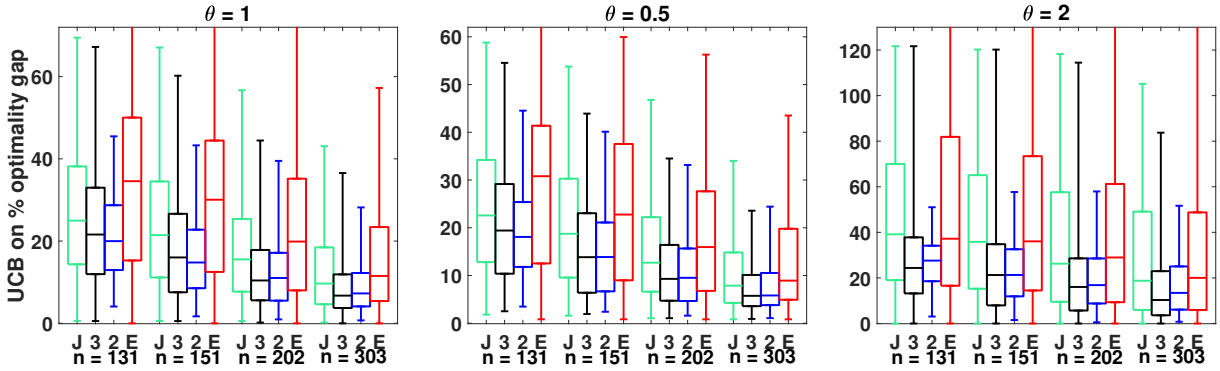

Fig. 7: **(Comparison of Wasserstein-DRO with J-SAA)** Comparison of the E+OLS (E) and J+OLS (J) approaches with the covariate-dependent tuning of the W+OLS radius using Algorithm 3 (3) and the covariate-independent tuning of the W+OLS radius using Algorithm 2 (2) for $d_x = 100$. Left: $\theta = 1$. Middle: $\theta = 0.5$. Right: $\theta = 2$.