Residuals-based distributionally robust optimization with covariate information

Rohit Kannan[1], Güzin Bayraksan[2], and James R. Luedtke[3]

[1]Center for Nonlinear Studies (T-CNLS) and Applied Mathematics & Plasma Physics (T-5), Los Alamos National Laboratory, Los Alamos, NM, USA. E-mail: rohitk@alum.mit.edu
[2]Department of Integrated Systems Engineering, The Ohio State University, Columbus, OH, USA. E-mail: bayraksan.1@osu.edu
[3]Department of Industrial & Systems Engineering and Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA. E-mail: jim.luedtke@wisc.edu

Version 2 (this document): May 3, 2022

Version 1: December 2, 2020

Abstract

We consider data-driven approaches that integrate a machine learning prediction model within distributionally robust optimization (DRO) given limited joint observations of uncertain parameters and covariates. Our framework is flexible in the sense that it can accommodate a variety of regression setups and DRO ambiguity sets. We investigate asymptotic and finite sample properties of solutions obtained using Wasserstein, sample robust optimization, and phi-divergence-based ambiguity sets within our DRO formulations, and explore 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.

Key words: Data-driven stochastic programming, distributionally robust optimization, Wasserstein distance, phi-divergences, covariates, machine learning, convergence rate, large deviations

1 Introduction

Stochastic programming [48] 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 set of feasible decisions, Y denotes a random vector of model parameters with support $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, and $c : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ is the 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) [33, 48], 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 [34], 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 Section 2.2). Despite its favorable theoretical guarantees [33, 34, 48], a limitation of the SAA approach is that its solutions may exhibit disappointing out-of-sample performance in the small sample size regime [9, 23].

Distributionally robust optimization (DRO) [43] is a 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 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 [15, 36, 43]).

We introduce a DRO framework for decision-making under uncertainty in the presence of covariate information and study its theoretical properties. We first consider the setup in [1, 7, 47] 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*

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

Applications of this framework include shipment planning under demand uncertainty [7, 10], where product demands can be predicted using past demands, location, and web search results before making production and inventory decisions, and portfolio optimization under market uncertainty [20], 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 regression model within a DRO framework in a bid to construct estimators for (SP) with better out-of-sample performance. Our data-driven DRO formulations are built around the residuals-based SAA formulations that we studied in [34]. We 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. The analysis in Sections 4 and 5 builds on our previous analysis of residuals-based SAA formulations in [34], but makes several new contributions, including analysis of the average rates of convergence of our DRO formulations and finite sample solution guarantees of Wasserstein DRO solutions. We also investigate data-driven methods for choosing the radii of these ambiguity sets in the presence of covariate observations in Section 6 without access to samples from the conditional distribution of Y given $X = x$. Numerical experiments in Section 7 demonstrate the potential benefits of our modular data-driven DRO framework 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 [1] and Bertsimas and Kallus [7] study policy-based empirical risk minimization and nonparametric regression-based reweighted SAA approaches for solving (SP). Bertsimas and Kallus [7] establish asymptotic optimality of their data-driven decisions, whereas Ban and Rudin [1] 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. [2] and Sen and Deng [47] use parametric regression methods along with their empirical residuals to generate scenarios of the random variables given covariate information. Ban et al. [2] prove asymptotic optimality of their decisions for their particular application. Kannan et al. [34] 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. [34] also 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 [31] 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. [10] consider a multi-stage DRO extension of the approach in [7] using the sample robust optimization method of [11]. They demonstrate asymptotic optimality of their decisions and develop an approximate solution method using linear decision rules. Bertsimas and Van Parys [8] propose a notion of 'bootstrap robustness'. They define DRO extensions of the Nadaraya-Watson and k-nearest neighbors formulations in [7] using ambiguity sets based on discrepancy measures and study their theoretical properties.

Blanchet et al. [14] and Nguyen et al. [41] 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. Boskos et al. [17–19] explore the construction of Wasserstein ambiguity sets for noisy observations of dynamically evolving random variables with *known dynamics* within a control setting. They also consider the case when the random variables may only be estimated using noisy observations of outputs of a linear time-varying system. Similar to our Wasserstein ambiguity sets in Section 4, they propose to enlarge the radius of the ambiguity set to account for errors in the estimates of the random variables due to measurement noise. Dou and Anitescu [20] 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 [24] construct a Wasserstein DRO extension of (SP) by linking trimmings of probability distributions with the partial mass transportation problem. They show that their approach naturally produces DRO extensions of formulations based on some nonparametric regression techniques. 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 regression model within a DRO framework. Our work is similar in spirit to [20, 24], but we consider more general formulations (SP), including two-stage stochastic programs, generic regression 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 [20] is that we consider an ambiguity set for the residuals of the regression model, but do not consider one for its coefficients for the sake of generality. We investigate the theoretical properties of our residuals-based DRO formulations in Sections 4 and 5. The case study in Section 7 demonstrates the modularity benefit 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 residuals-based SAA framework in [34]. Our DRO framework is flexible in the sense that it can accommodate side information effectively using a variety of regression setups and ambiguity sets. It also seamlessly extends existing DRO formulations that do not utilize covariate information.

2. We study asymptotic optimality, pointwise and average 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. An important difference compared to traditional DRO *without* covariate information is that we cannot assume access to samples from the conditional distribution of Y given

$X = x$. Consequently, radius selection strategies used in traditional DRO (that do not use covariate information) may no longer yield the best radius for our ambiguity sets. We empirically demonstrate that our new radius selection strategies yield good-quality decisions that work better for our setting than the strategies in the traditional DRO literature.

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. These experiments also provide insight into the relative performance of different DRO formulations and illustrate the benefit of our framework's modularity.

Notation. Let $[n] := \{1, \ldots, n\}$, $\|\cdot\|_p$ 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

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

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. Let (S, Σ, μ) be a measure space. Given $q \in [1, +\infty]$, we write $\|F\|_{L^q}$ to denote the L^q-norm of a measurable function $F : S \to \mathbb{R}^{d_F}$, i.e., $\|F\|_{L^q} := \left(\int_S \|F\|^q d\mu\right)^{1/q}$. For any $S \subseteq \mathbb{R}^{d_z}$, let $C(S)$ denote the Banach space of real-valued continuous 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 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 Y and 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 assume that all functions, sets and selections are measurable (see [48, 50] for detailed consideration of these issues).

2 Preliminaries

2.1 Framework

We assume throughout that the random vector Y (commonly referred to as "dependent variables") is related to the random covariates X (commonly referred to as "independent variables") 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[2] of the covariates X, and that f^* is known to belong to a class of functions[3] $\mathcal{F}$. The model class $\mathcal{F}$ can be infinite-dimensional and can depend on the

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

[2] We investigate extensions of our framework that can adapt to heteroscedasticity in [35], where we assume that $Y = f^*(X) + Q^*(X)\varepsilon$ with X and ε independent (here, $Q^*(X)$ denotes the covariate-dependent covariance matrix of the errors).

[3] See Remark 1 at the end of this section for a discussion of the case when $f^* \notin \mathcal{F}$.

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_X and P_ε denote the distribution of X and 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

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

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, $\mathbb{E}\left[|c(z, f^*(x) + \varepsilon)|\right] < +\infty$ 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 [34]. 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 $\varepsilon^i := y^i - f^*(x^i)$, $i \in [n]$, 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* (ER-SAA) to problem (3):

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

In contrast with [34], 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 may be helpful when the ER-SAA scenarios $\{\hat{f}_n(x) + \hat{\varepsilon}_n^i\}_{i \in [n]}$ lie outside the support $\mathcal{Y}$ even though the "true" FI-SAA scenarios $\{f^*(x) + \varepsilon^i\}_{i \in [n]}$ are elements of $\mathcal{Y}$. It also 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). This is because the ambiguity set of these DRO formulations then becomes a ball around the ER-SAA distribution defined below, with respect to a suitable Wasserstein metric (cf. Section 4 of [23]). We stick with this modification of the ER-SAA formulation (5) throughout for uniformity.

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. This motivated our construction in [34] of two alternative SAA formulations that instead use leave-one-out (jackknife) residuals to construct scenarios of Y given $X = x$.

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.,

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

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 $P_n^*(x)$. To enable this, note that the Lipschitz continuity of orthogonal projections[4] implies that for each $x \in \mathcal{X}$

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

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 $x^i \in \mathcal{X}$.

Remark 1. Although we assume that the regression function f^* belongs to the model class $\mathcal{F}$ to establish our theoretical guarantees, our data-driven approximations of (3) are well defined even when $f^* \notin \mathcal{F}$, i.e., when the regression model is misspecified. In this setting, under mild assumptions the regression estimates $\hat{f}_n$ converge to the best approximation to f^* in the model class $\mathcal{F}$, denoted by $\bar{f}$, and residuals-based SAA and DRO formulations then yield solutions converging to the optimal solution of the problem

$$\min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^{n} c(z, \bar{f}(x) + \bar{\varepsilon}^i),$$

where $\bar{\varepsilon}^i := f^*(x^i) - \bar{f}(x^i) + \varepsilon^i$. Therefore, we can replace f^* by $\bar{f}$ and $\{\varepsilon^i\}$ by $\{\bar{\varepsilon}^i\}$ in our assumptions and results to characterize the asymptotic and finite sample properties of our data-driven approximations in this case.

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 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 the objective function of problem (7) is real-valued and lower semicontinuous on $\mathcal{Z}$ for each $x \in \mathcal{X}$. This ensures that its optimal solution set $\hat{S}_n^{DRO}(x)$ is nonempty for each $x \in \mathcal{X}$.

We seek to derive DRO formulations (7) that obtain a solution $\hat{z}_n^{DRO}(x)$ with good out-of-sample performance $g(\hat{z}_n^{DRO}(x); x)$ for relatively small sample sizes n. To support our investigation of such formulations, we consider different desirable properties they may have. 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. [9, 23]):

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

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

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[5]

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

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

[5] In special cases (e.g., smooth unconstrained problems, see [30, Section 5]), it may be possible to establish the sharper convergence rate $\left|g(\hat{z}_n^{DRO}(x); x) - v^*(x)\right| = O_p\big(n^{-r}\big)$.

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

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

We would also like the solution $\hat{z}_n^{DRO}(x)$ to possess the following guarantee:

4. **Finite sample solution guarantee:** for a.e. $x \in \mathcal{X}$ and any $\eta > 0$, there exist positive constants $\Gamma(\eta, x)$ and $\gamma(\eta, x)$ such that the solution $\hat{z}_n^{DRO}(x)$ of the DRO problem (7) with a suitable specification of the radius of the ambiguity set $\hat{\mathcal{P}}_n(x)$ satisfies

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

Finally, we would also like problem (7) to be efficiently solvable in practice. Although our asymptotic guarantees are stated in terms of convergence in probability, they can be naturally extended to consider almost sure convergence under stronger assumptions on problem (3) and the regression estimate $\hat{f}_n$.

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 introduced in [34]. The analysis in [34, 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 [43], 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. [23, 28, 42]): 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. [11, 53]): given radius $\mu_n(x) \geq 0$ and parameter $p \in [1, +\infty]$, set[6]

$$\hat{\mathcal{P}}_n(x) = \left\{Q = \frac{1}{n}\sum_{i=1}^n \delta_{\bar{y}^i} : \|\bar{y}^i - \text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{\varepsilon}_n^i)\|_p \leq \mu_n(x), \bar{y}^i \in \mathcal{Y}, \forall i \in [n]\right\}.$$

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 $x \in \mathcal{X}$

$$\begin{aligned} &p \in \mathbb{R}_+^n \text{ and } \sum_{i=1}^n p_i = 1, \quad \forall p \in \mathfrak{P}_n(x; \zeta_n(x)), \\ &\lim_{\zeta \downarrow 0} \mathfrak{P}_n(x; \zeta) = \mathfrak{P}_n(x; 0) = \left\{\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)\right\}. \end{aligned} \tag{8}$$

[6] 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.

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 [45, 48]: 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.

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 [11]. Sections 4.1 and 5 of [23] and Section 2.2 of [36] identify conditions under which the resulting ER-DRO formulation (7) is computationally tractable. References [3, 28, 32] also consider solution approaches for the setting where 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 $\mathbb{E}\left[\exp(\|\varepsilon\|^a)\right] < +\infty$.

Assumption 1 (cf. [23, Assumption 3.3]) may not hold for sub-exponential distributions. Additionally, when $p \geq 2$, it requires the tails of ε to decay at a faster rate than Gaussian tails. However, sub-Gaussian errors (see Definition 1 below) can be handled using $p \in [1, 2)$. Our first result identifies sufficient conditions under which sub-Gaussian errors satisfy Assumption 1 for $p \in [1, 2)$.

Definition 1. A random vector $V \in \mathbb{R}^{d_v}$ is said to be sub-Gaussian with variance proxy σ^2 if $\mathbb{E}[V] = 0$ and

$$\mathbb{E}\left[\exp(su^{\mathrm{T}}V)\right] \leq \exp(0.5\sigma^2 s^2),\ \ \forall s \in \mathbb{R} \text{ and } u \in \mathbb{R}^{d_v} \text{ s.t. } \|u\| = 1.$$

Definition 1 implies that the class of sub-Gaussian random vectors includes zero-mean Gaussian random vectors.

Proposition 1. Suppose ε is a sub-Gaussian random vector with independent components and variance proxy σ^2. Then $\mathbb{E}\left[\exp(\|\varepsilon\|^a)\right] < +\infty$, $\forall a \in (1, 2)$.

Proof. See Appendix A.1. □

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

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

We use same the constant $\kappa_{p,n}(\alpha, x)$ for both the prediction error at the new covariate realization $x \in \mathcal{X}$ and the power-mean estimation error on the training data points $\{x^i\}_{i=1}^n$ to keep the notation simple even though the latter does not depend on x. Appendix EC.3 of [34] identifies conditions under which parametric regression methods such as ordinary least squares (OLS) and Lasso regression satisfy Assumption 2 for the case $p = 2$ with constants $\kappa_{2,n}^2(\alpha, x) = O(n^{-1}\log(\alpha^{-1}))$; we omit the dependence on x here for simplicity. Nonparametric regression methods, on the other hand, typically only satisfy Assumption 2 with constants $\kappa_{p,n}^p(\alpha, x) = O(n^{-1}\log(\alpha^{-1}))^{O(1)/d_x}$. Similar bounds readily hold for $p \neq 2$, e.g., if the support $\mathcal{X}$ of the covariates is compact. 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 2 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 [25, Theorem 2] for the case $p = d_y/2$. Lemma 2 also applies to non-i.i.d. data $\mathcal{D}_n$ such as time series data (cf. [20]).

Lemma 2. [Theorem 2 of [25]] 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 $x \in \mathcal{X}$

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

The $O(1)$ constants in Lemma 2 only depend on a, d_y, and $\mathbb{E}\left[\exp(\|\varepsilon\|^a)\right]$ (see [25, Theorem 2]). We require a few intermediate results before we can establish a finite sample certificate guarantee for Wasserstein ER-DRO estimators in Theorem 7 (cf. [36, 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 3. For each $x \in \mathcal{X}$

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

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

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

□

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 4. 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. □

We also require the following simple inequality.

Lemma 5. Let V and W be random variables and $c_1, c_2 \in \mathbb{R}$. Then

$$\mathbb{P}(V + W > c_1 + c_2) \leq \mathbb{P}(V > c_1) + \mathbb{P}(W > c_2).$$

We are now ready to derive a finite sample guarantee for $\left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^p\right)^{1/p}$.

Lemma 6. 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. $x \in \mathcal{X}$

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

where the first step follows by Lemma 4, the second step follows from Lemma 5, and the last step holds by Assumption 2. □

To establish our asymptotic and finite sample guarantees, we enlarge the radius of the Wasserstein ambiguity set that is used in the absence of covariate information [23, 36]. 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 $\kappa_{p,n}^{(1)}(\alpha, x) := 2\kappa_{p,n}\left(\frac{\alpha}{4}, x\right)$ and

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

The constants a and $\kappa_{p,n}$ above are defined in Assumptions 1 and 2. The term $\kappa_{p,n}^{(2)}(\alpha)$ is obtained by setting the r.h.s. of the inequality in Lemma 2 to $\alpha/2$. While this choice of ζ_n helps us derive our theoretical guarantees, it involves unknown constants and is often conservative in practice (see Remark 3). We investigate practical data-driven approaches for choosing ζ_n in Section 6.

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

Proof. Lemma 6 and Lemma 2 imply that

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

Consequently, equation (9), Lemma 3 and Lemma 5 imply that

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

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

We now make the following assumption along the lines of [23, 28, 36] to show in Theorem 9 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 10.

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 is $p \geq 2$. 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 [34, Appendix EC.2].

Assumption 6. The sequence of risk levels $\{\alpha_n\} \subset (0,1)$ satisfies $\sum_n \alpha_n < \infty$ 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 (9).

We have the following useful result.

Lemma 8. Suppose Assumptions 1, 2, and 6 hold and the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. Then, a.s. for n large enough

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

Furthermore, let $z^*(x) \in S^*(x)$ be an optimal solution to problem (3). Then

A. If Assumption 4 holds, we a.s. have for a.e. $x \in \mathcal{X}$ and n large enough

$$\hat{v}_n^{DRO}(x) \le v^*(x) + 2L_1(z^*(x))\zeta_n(\alpha_n, x).$$

B. If Assumption 5 holds, we a.s. have for a.e. $x \in \mathcal{X}$ and n large enough

$$\hat{v}_n^{DRO}(x) \le v^*(x) + 2\big(\mathbb{E}\left[\|\nabla c(z^*(x), Y)\|^2\right]\big)^{1/2}\zeta_n(\alpha_n, x) + 4L_2(z^*(x))\zeta_n^2(\alpha_n, x).$$

Proof. See Appendix A.2. □

We now state our asymptotic guarantees for Wasserstein ER-DRO.

Theorem 9. [Consistency and asymptotic optimality] Suppose Assumptions 1, 2, 3, and 6 hold and the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. 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, i.e.,

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

Theorem 10. [Rate of convergence] Suppose the assumptions of Theorem 9 and either Assumption 4 or Assumption 5 hold. Then, for a.e. $x \in \mathcal{X}$, the ER-DRO problem (7) with ambiguity set $\hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x))$ satisfies

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

Proofs of Theorems 9 and 10 are in Appendices A.3 and A.4. The proof of Theorem 9 mirrors that of [23, Theorem 3.6] (it shows that the conclusions in fact hold almost surely). Similar to the setting without covariate information [23], we can typically choose the sequence of risk levels $\{\alpha_n\}$ in Assumption 6 to be any sequence converging at a slower rate than $\{\exp(-n)\}$ when the errors ε are sub-Gaussian (see the discussion following Assumption 2).

Remark 2. Assumption 5 can be weakened to consider functions c that satisfy

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

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 $p \ge 1 + \kappa$, see [29, 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

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

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

Remark 3. 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 [36, 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 10 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. [34, Theorem 2]). The analysis in Gao [27] and Blanchet et al. [16] implies that, under certain assumptions, using the smaller radius $\zeta_n(\alpha, x) := \max\{\kappa^{(1)}_{p,n}(\alpha, x), \bar{\kappa}^{(2)}_{p,n}(\alpha)\}$ with suitably chosen $\bar{\kappa}^{(2)}_{p,n}(\alpha) = O(n^{-1/2})$ results in estimators with a finite sample certificate-type guarantee and sharper convergence rates. This smaller 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^*. Consequently, if sharper finite sample guarantees such as those in [16, 27] apply, then Theorem 10 can be readily adapted to derive sharper convergence rates. However, the assumptions in [16, 27] may exclude some formulations of interest, such as two-stage stochastic programs (see, e.g., [16, Assumption (A3)] and [27, Assumption 2]), or may be difficult to verify in general (see [27, Section 5]).

Next, we identify conditions under which the optimal objective value of the Wasserstein ER-DRO problem (7) converges to $v^*(x)$ at a suitable rate with respect to the L^q-norm on $\mathcal{X}$ for $q \in [1, \infty]$. We make the following stronger form of Assumption 2 for simplicity.

Assumption 7. 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 $\kappa_n(\alpha)$ such that $\mathbb{P}\big\{\sup_{x\in\mathcal{X}} \|f^*(x) - \hat{f}_n(x)\| > \kappa_n(\alpha)\big\} \le \alpha$.

Appendix EC.3 of [34] verifies that Assumption 7 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. When Assumption 7 holds, we write $\zeta_n(\alpha)$ instead of $\zeta_n(\alpha, x)$ for the radius specified by (9).

Theorem 11. [Mean convergence rate] Suppose Assumptions 1, 3, 6, and 7 hold, and $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. Let $q \in [1, +\infty]$ and suppose either Assumption 4 holds with $\|L_1(z^*(\cdot))\|_{L^q} < +\infty$, or Assumption 5 holds with $\|\big(\mathbb{E}_Y\big[\|\nabla c(z^*(\cdot), Y)\|^2\big]\big)^{1/2}\|_{L^q} < +\infty$ and $\|L_2(z^*(\cdot))\|_{L^q} < +\infty$. Then, the ER-DRO problem (7) with ambiguity set $\hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x))$ satisfies

$$\big\|\hat{v}_n^{DRO}(X) - v^*(X)\big\|_{L^q} = O_p\big(\zeta_n(\alpha_n)\big),$$
$$\big\|g(\hat{z}_n^{DRO}(X); X) - v^*(X)\big\|_{L^q} = O_p\big(\zeta_n(\alpha_n)\big).$$

Proof. Following the proof of Theorem 7, we have

$$\mathbb{P}\big\{d_{W,p}(\hat{P}_n^{ER}(x), P_{Y|X=x}) > \zeta_n(\alpha_n),\ \ \forall x \in \mathcal{X}\big\} \le \alpha_n.$$

Note that we are able to make the above assertion *jointly* over all $x \in \mathcal{X}$ because the radius $\zeta_n(\alpha_n)$ is independent of x by Assumption 7. Following the proof of Lemma 8, we then a.s. have for all n large enough:

$$v^*(x) \le g(\hat{z}_n^{DRO}(x); x) \le \hat{v}_n^{DRO}(x), \quad \forall x \in \mathcal{X}.$$

Suppose Assumption 4 holds. Let $z^*(x) \in S^*(x)$. From part A of Lemma 8, the above inequalities a.s. imply for all n large enough:

$$\hat{v}_n^{DRO}(x) - v^*(x) \le 2L_1(z^*(x))\zeta_n(\alpha_n), \quad \forall x \in \mathcal{X}.$$

Consequently, when Assumption 4 holds

$$\big\|\hat{v}_n^{DRO}(X) - v^*(X)\big\|_{L^q} = \|L_1(z^*(\cdot))\|_{L^q}\, O_p(\zeta_n(\alpha_n)).$$

Suppose instead that Assumption 5 holds. From part B of Lemma 8, the above inequalities a.s. imply for all n large enough and each $x \in \mathcal{X}$:

$$\hat{v}_n^{DRO}(x) - v^*(x) \le 2\big(\mathbb{E}_Y\big[\|\nabla c(z^*(x), Y)\|^2\big]\big)^{1/2}\zeta_n(\alpha_n) + 4L_2(z^*(x))\zeta_n^2(\alpha_n).$$

Consequently, when Assumption 5 holds

$$\big\|\hat{v}_n^{DRO}(X) - v^*(X)\big\|_{L^q} = \|\big(\mathbb{E}_Y\big[\|\nabla c(z^*(\cdot), Y)\|^2\big]\big)^{1/2}\|_{L^q}\, O_p(\zeta_n(\alpha_n)).$$

□

Appendix EC.2 of [34] presents conditions under which some of the new assumptions of Theorem 11 hold. We now identify conditions under which the ER-DRO estimators possess a finite sample solution guarantee. We first refine Assumption 2 to another more convenient, stronger form in Assumption 8.

Assumption 8. 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 [34] verifies Assumption 8 for some popular regression setups for $p = 2$; see the discussion after Assumption 2 for $p \neq 2$. The following result will prove useful in deriving our finite sample solution guarantee.

Lemma 12. Suppose Assumptions 1, 2, 3, and 8 hold, the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d., and either Assumption 4 or Assumption 5 holds. Then, for a.e. $x \in \mathcal{X}$ and any $\kappa > 0$, there exist positive constants $\tilde{\Gamma}(\kappa, x)$ and $\tilde{\gamma}(\kappa, x)$ such that the solution of the ER-DRO problem (7) with risk level $\alpha = \tilde{\Gamma}(\kappa, x)\exp(-n\tilde{\gamma}(\kappa, x))$, radius $\zeta_n(\alpha, x)$ specified by (9), and ambiguity set $\hat{\mathcal{P}}_n(x; \zeta_n(\alpha, x))$ satisfies

$$\mathbb{P}\left\{g(\hat{z}_n^{DRO}(x); x) > v^*(x) + \kappa\right\} \leq 2\tilde{\Gamma}(\kappa, x)\exp(-n\tilde{\gamma}(\kappa, x)). \tag{10}$$

Proof. See Appendix A.5. □

Theorem 13. [Finite sample solution guarantee] Suppose Assumptions 1, 2, 3, and 8 hold, the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d., and either Assumption 4 or Assumption 5 holds. Then, for a.e. $x \in \mathcal{X}$ and any $\eta > 0$, there exist positive constants $\Gamma(\eta, x)$ and $\gamma(\eta, x)$ such that the solution of the ER-DRO problem (7) with risk level $\alpha = \Gamma(\eta, x)\exp(-n\gamma(\eta, x))$, radius $\zeta_n(\alpha, x)$ determined using equation (9), and ambiguity set $\hat{\mathcal{P}}_n(x; \zeta_n(\alpha, x))$ satisfies

$$\mathbb{P}\left\{\text{dist}(\hat{z}_n^{DRO}(x), S^*(x)) \geq \eta\right\} \leq 2\Gamma(\eta, x)\exp(-n\gamma(\eta, x)).$$

Proof. From Lemma 12, we have for any $\kappa > 0$, there exist $\tilde{\Gamma}(\kappa, x) > 0$ and $\tilde{\gamma}(\kappa, x) > 0$ such that inequality (10) holds with $\alpha = \tilde{\Gamma}(\kappa, x)\exp(-n\tilde{\gamma}(\kappa, x))$. We now argue that inequality (10) implies the stated result.

Suppose we have $\text{dist}(\hat{z}_n^{DRO}(x), S^*(x)) \geq \eta$ for some $\eta > 0$, $x \in \mathcal{X}$, and sample path. Since $g(\cdot; x)$ is lower semicontinuous on the compact set $\mathcal{Z}$ for a.e. $x \in \mathcal{X}$, [34, 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 paths of measure zero). We now bound the probability of this event. The above arguments imply for a.e. $x \in \mathcal{X}$

$$\begin{aligned}\mathbb{P}\left\{\text{dist}(\hat{z}_n^{DRO}(x), S^*(x)) \geq \eta\right\} &\leq \mathbb{P}\left\{g(\hat{z}_n^{DRO}(x); x) > v^*(x) + \kappa(\eta, x)\right\} \\ &\leq 2\tilde{\Gamma}(\kappa(\eta, x), x)\exp(-n\tilde{\gamma}(\kappa(\eta, x), x)).\end{aligned}$$

Therefore, the desired result holds with constants $\Gamma(\eta, x) = \tilde{\Gamma}(\kappa(\eta, x), x)$ and $\gamma(\eta, x) = \tilde{\gamma}(\kappa(\eta, x), x)$. □

Theorem 13 is similar to the finite sample guarantee in [34, Theorem 3] for solutions to the ER-SAA problem. However, unlike [34, Theorem 3], the dependence of the convergence rate on the parameter η in Theorem 13 suffers from the curse of dimensionality even if we use parametric regression methods to estimate f^* (cf. Remark 3). Inequality (10) shows that the out-of-sample cost of the Wasserstein ER-DRO estimators possesses a finite sample guarantee similar to the guarantee in the solution space. This convergence rate estimate also suffers from the curse of dimensionality with respect to the parameter κ (we do not know if faster rates of convergence can be derived).

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

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

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

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 [12, 52] 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 $\hat{g}_{s,n}^{ER}$ and $g_{s,n}^*$ the functions

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

Note that the function $\hat{g}_{s,n}^{ER}$ 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 9. 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 10. 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 11. 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 9 is a uniform Lipschitz continuity assumption that strengthens Assumption 4. Appendix EC.2 of [34] verifies that Assumption 9 holds for two-stage stochastic MIPs with continuous recourse. Assumption 10 is a uniform weak LLN assumption, whereas Assumption 11 is a mild consistency assumption that holds for many popular regression setups (cf. Assumptions 3 and 4 of [34]). Assumption 11 is weaker than Assumption 2. We require the following additional assumptions for ambiguity sets with $\zeta_n(x) > 0$.

Assumption 12. The radius $\zeta_n(x)$ of the ambiguity set is chosen such that

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

for some constant $\rho > 1$.

Assumption 13. The following weak uniform LLN holds for a.e. $x \in \mathcal{X}$:

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

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

Assumption 12 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 $\big\{\big(\frac{1}{n},\ldots,\frac{1}{n}\big)\big\}$ at a fast enough rate. This is always possible since we assume equation (8) holds. We are interested in cases when Assumption 12 holds with $\rho \in (1,2]$ (see Theorem 17). Lemma 13 of [21] (cf. [6, 37, 38]) 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

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

Consequently, Assumption 12 holds for such phi-divergence-based ambiguity sets $\mathfrak{P}_n(x;\zeta_n(x))$ whenever the radius $\zeta_n(x) = O(n^{1-\rho})$. This bound on $\zeta_n(x)$ is *sharp* in the sense that Assumption 12 does not hold if $\zeta_n(x)$ grows faster than $n^{1-\rho}$ asymptotically. Appendix B presents some other examples of ambiguity sets for which Assumption 12 holds.

Theorem 7.48 of [48] presents conditions under which both Assumptions 10 and 13 hold when the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. Note that Assumption 13 can also be equivalently stated as a weak uniform LLN assumption on the sample variance of the sequence $\{c(z, f^*(x)+\varepsilon^i)\}_{i=1}^n$ when $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. [21].

The following result will be useful in deriving asymptotic guarantees for the ER-DRO formulations studied in this section.

Lemma 14. Suppose Assumption 9 holds. We have for each $x \in \mathcal{X}$

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} \big|\hat{g}^{ER}_{s,n}(z;x) - g(z;x)\big| \\
&\le \sup_{z\in\mathcal{Z}} L(z)\Big(\mu_n(x) + \Big(\frac{1}{n}\sum_{i=1}^n \big(\|\hat{\varepsilon}^i_n(x)\|\big)^2\Big)^{\frac{1}{2}}\Big) \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \Big(1 + n\sum_{i=1}^n \Big(p_i - \frac{1}{n}\Big)^2\Big)^{\frac{1}{2}} \\
&\quad + \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \Big(n\sum_{i=1}^n \Big(p_i - \frac{1}{n}\Big)^2\Big)^{\frac{1}{2}} \sup_{z\in\mathcal{Z}} \Big(\frac{1}{n}\sum_{i=1}^n \big(c(z, f^*(x)+\varepsilon^i)\big)^2\Big)^{\frac{1}{2}} \\
&\quad + \sup_{z\in\mathcal{Z}} |g^*_n(z;x) - g(z;x)|. \qquad (11)
\end{aligned}$$

Proof. See Appendix A.6. □

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 [21] presents an analogous result for a class of phi-divergence-based ambiguity sets in the absence of covariate information.

Proposition 15. Suppose Assumptions 9 to 13 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 $g(\cdot; x)$ of the true problem (3) uniformly on the set $\mathcal{Z}$.

Proof. We wish to show that

$$\sup_{z\in\mathcal{Z}} \left|\hat{g}^{ER}_{s,n}(z; x) - g(z; x)\right| \xrightarrow{p} 0, \quad \text{for a.e. } x \in \mathcal{X}.$$

We bound this term from above using Lemma 14.

The third term on the r.h.s. of (11) vanishes in the limit in probability for a.e. $x \in \mathcal{X}$ under Assumption 10. We show that the first two terms also converge to zero in probability; the result then follows by $o_p(1)+o_p(1) = o_p(1)$.

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

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} L(z)\Big(\mu_n(x) + \Big(\frac{1}{n}\sum_{i=1}^{n}\big(\|\tilde{\varepsilon}^i_n(x)\|\big)^2\Big)^{\frac{1}{2}}\Big) \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))}\Big(1 + n\sum_{i=1}^{n}\Big(p_i - \frac{1}{n}\Big)^2\Big)^{\frac{1}{2}} \\
&= O(1)o_p(1)O(1) = o_p(1),
\end{aligned}$$

on account of Assumptions 9, 11 and 12, $\lim_{n\to\infty} \mu_n(x) = 0$, and [34, Lemma 1].

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

$$\begin{aligned}
&\sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))}\Big(n\sum_{i=1}^{n}\Big(p_i - \frac{1}{n}\Big)^2\Big)^{\frac{1}{2}} \sup_{z\in\mathcal{Z}}\Big(\frac{1}{n}\sum_{i=1}^{n}\big(c(z, f^*(x) + \varepsilon^i)\big)^2\Big)^{\frac{1}{2}} \\
&= o(1)O_p(1) = o_p(1),
\end{aligned}$$

on account of Assumptions 12 and 13. □

It can be seen from the proof that Assumptions 12 and 13 are not required for sample robust optimization-based DRO, i.e., when the radius $\zeta_n(x) \equiv 0$.

Remark 4. Assumption 9 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 15 holds if we replace Assumption 9 with [34, Assumption 2]. When $\zeta_n(x) \neq 0$, we need to replace Assumption 9 with strengthened versions of Assumption 11 and [34, Assumption 2] involving fourth degree terms.

Proposition 15 provides the foundation for showing that the ER-DRO estimators are asymptotically optimal. We omit the proof of Theorem 16 since it is identical to the proof of [34, Theorem 1] in light of Proposition 15.

Theorem 16. [Consistency and asymptotic optimality] Suppose the assumptions of Proposition 15 hold. Then, for a.e. $x \in \mathcal{X}$

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

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 [34]).

Assumption 14. 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 $C(\mathcal{Z})$.

Assumption 15. There is a constant[7] $0 < r \leq 1$ 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 15 strengthens Assumption 11. 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 15 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). The choice $\rho = 1 + r$ in Assumption 12 ensures that the resulting ER-DRO estimators enjoy the same rate of convergence as the ER-SAA estimators in [34].

Theorem 17. [Rate of convergence] Suppose Assumptions 9, 13, 14, and 15 hold. In addition, suppose Assumption 12 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 15. 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. We bound $\sup_{z\in\mathcal{Z}}\left|\hat{g}_{s,n}^{ER}(z;x) - g(z;x)\right|$ from above using Lemma 14.

Assumptions 9, 12, and 15 and $\mu_n = O(n^{-r/2})$ imply that the first term on the r.h.s. of inequality (11) satisfies for a.e. $x \in \mathcal{X}$

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

Assumptions 12 and 13 imply that the second term on the r.h.s. of inequality (11) satisfies for a.e. $x \in \mathcal{X}$

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

Finally, Assumption 14 implies $\sqrt{n}\sup_{z\in\mathcal{Z}}|g_n^*(z;x) - g(z;x)| = O_p(1)$ for a.e. $x \in \mathcal{X}$, which in turn 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

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

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

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

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

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

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

Our next result analyzes the rate of convergence of the ER-DRO objective with respect to the L^q-norm. We require the following refined assumptions.

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

Assumption 16. The radius $\zeta_n(x)$ of the ambiguity set is chosen such that

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

for some constant $\rho > 1$.

Assumption 17. There is a constant $0 < r \leq 1$ such that the regression estimate $\hat{f}_n$ satisfies the following convergence rate criteria:

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

Assumption 18. The function c in problem (3) and the data $\mathcal{D}_n$ satisfy:

$$\left\| \sup_{z\in\mathcal{Z}} \left(\frac{1}{n}\sum_{i=1}^n \left(c(z, f^*(X) + \varepsilon^i)\right)^2 \right)^{\frac{1}{2}} \right\|_{L^q} = O_p(1),$$
$$\left\| \sup_{z\in\mathcal{Z}} |g_n^*(z;X) - g(z;X)| \right\|_{L^q} = O_p(n^{-1/2}).$$

Asssumption 16 requires Assumption 12 to hold uniformly over the covariates $x \in \mathcal{X}$. It reduces to Assumption 12 when the ambiguity set $\mathfrak{P}_n(x;\zeta_n(x))$ is chosen to be independent of $x \in \mathcal{X}$. Assumption 17 requires the estimation error to converge to zero on average over the covariates. Appendix EC.3. of [34] identifies conditions under which parametric regression methods satisfy Assumption 17 under LLN and moment assumptions on the covariate distribution. Assumption 18 holds, for example, when the corresponding uniform LLNs (with respect to the decisions $z \in \mathcal{Z}$) hold uniformly over the covariates (cf. Theorem 7.48 of [48]).

Theorem 18. [Mean convergence rate] Suppose Assumptions 9, 17, and 18 hold. Let $q \in [1, +\infty]$. Suppose Assumption 16 holds with $\rho = 1+r$, with constant r defined in Assumption 17, and the radius $\mu_n(x)$ satisfies $\|\mu_n(X)\|_{L^q} = O(n^{-r/2})$. Then, the solution of the ER-DRO problem (7) satisfies

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

Proof. See Appendix A.7. □

Finally, we make the following assumption 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 [11] implies that the sample robust optimization-based ER-DRO problem is equivalent to the ∞-Wasserstein distance-based ER-DRO problem (7) with ambiguity set $\hat{\mathcal{P}}_n(x) := \left\{Q \in \mathcal{P}(\mathcal{Y}) : d_{W,\infty}(Q, \hat{P}_n^{ER}(x)) \leq \mu_n(x)\right\}$.

Assumption 19. 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 $1/\lambda(x) \leq \Lambda_Y(y;x) \leq \lambda(x)$, for some $\lambda(x) \geq 1$.

Trillos and Slepčev [49] consider cases when Assumption 19 holds. This assumption yields the following concentration of measure result for the true empirical distribution $P_n^*(x)$. Note that Lemma 19 also applies to settings with non-i.i.d. data $\mathcal{D}_n$ such as time series data.

Lemma 19. [Theorem 1.1 of [49]] Suppose Assumption 19 holds and the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. Then, for any constant $\beta > 2$ and a.e. $x \in \mathcal{X}$

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

where the $O(1)$ term depends only on β, $\bar{\mathcal{Y}}$, and $\lambda(x)$ in Assumption 19.

The next result is the analogue of Lemma 3 for the ∞-Wasserstein distance.

Lemma 20. For each $x \in \mathcal{X}$

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

Proof. The triangle inequality for the ∞-Wasserstein distance yields

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

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

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

□

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

$$\zeta_n(\alpha, x) := 0, \quad \mu_n(\alpha, x) := \kappa_{\infty,n}^{(1)}(\alpha) + \kappa_{\infty,n}^{(2)}(x) \tag{12}$$

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

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

the constant κ_n is defined in Assumption 7 and the constant $0 < \theta < 1$ may be chosen arbitrarily close to one. The term $\kappa_{\infty,n}^{(2)}(x)$ above is chosen so that it is greater than the $O(1)\log(n)/n^{1/d_y}$ term in Lemma 19 for $\beta = 4$ and n large enough. 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 [12]. 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 3). We investigate practical approaches for choosing the radius μ_n in Section 6.

Theorem 21. [Finite sample certificate-type guarantee] Suppose Assumptions 7 and 19 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 (12). 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 (12) a.s. satisfies

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

Proof. Our proof follows the outline of the proof of [12, Theorem 1].

Lemma 20, the probability inequality used in the proof of Lemma 6, and Assumption 7 yield for a.e. $x \in \mathcal{X}$

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

Consider $\beta = 4$ in Lemma 19. Because $\kappa_{\infty,n}^{(2)}(x) \geq O(1)\log(n)/n^{1/d_y}$ for n large enough, we have from Lemma 19 that for a.e. $x \in \mathcal{X}$ and n large enough

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

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 $n \geq N(x)$, $d_{W,\infty}(\hat{P}_n^{ER}(x), P_{Y|X=x}) \leq \mu_n(\alpha_n, x)$.

Recall that our sample robust optimization-based ER-DRO problem is equivalent to the ∞-Wasserstein distance-based ER-DRO problem with ambiguity set $\hat{\mathcal{P}}_n(x) := \big\{Q \in \mathcal{P}(\mathcal{Y}) : d_{W,\infty}(Q, \hat{P}_n^{ER}(x)) \leq \mu_n(\alpha_n, x)\big\}$ [11, 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.

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).

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 3). 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 $\hat{z}_n^{DRO}(x)$.

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

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

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

            where the ambigiuity set $\hat{\mathcal{P}}_{-k}(\bar{x})$ with radius ζ is centered at the estimated empirical distribution $\hat{P}^{ER}_{-k}(\bar{x}) := \dfrac{1}{n - |S_k|} \displaystyle\sum_{i \notin S_k} \delta_{\hat{f}_{-k}(\bar{x}) + \hat{\varepsilon}^i_{-k}}$.
        end for
    end for
end for
Output: Radius $\zeta_n \in \arg\min_{\zeta \in \Delta} \dfrac{1}{T} \displaystyle\sum_{\bar{x} \in \bar{\mathcal{X}}} \frac{1}{K} \sum_{k \in [K]} \frac{1}{|S_k|} \sum_{i \in S_k} c(\hat{z}^{DRO}_{-k}(\bar{x}, \zeta), y^i)$ for the ambiguity set $\hat{\mathcal{P}}_n(x)$ for the ER-DRO problem (7).

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

Input: data $\mathcal{D}_n$, set of candidate radii Δ, number of folds K, and new covariate realization $x \in \mathcal{X}$.
Partition $[n]$ into subsets $S_1, \ldots, S_K$ of (roughly) equal size at random. Let $\mathcal{D}_{-k} := \mathcal{D}_n \backslash \{(y^i, x^i)\}_{i \in S_k}$.
for $k = 1, \ldots, K$ **do**
    for $\zeta \in \Delta$ **do**
        Fit a regression model $\hat{f}_{-k}$ using the data $\mathcal{D}_{-k}$ and compute its in-sample residuals $\{\hat{\varepsilon}^i_{-k}\}_{i \notin S_k} := \{y^i - \hat{f}_{-k}(x^i)\}_{i \notin S_k}$.
        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 $\hat{P}^{ER}_{-k}(x) := \dfrac{1}{n - |S_k|} \displaystyle\sum_{i \notin S_k} \delta_{\hat{f}_{-k}(x) + \hat{\varepsilon}^i_{-k}}$.
        Fit a regression model $\hat{f}_k$ using the data $\{(y^i, x^i)\}_{i \in S_k}$ and compute its in-sample residuals $\{\hat{\varepsilon}^i_k\}_{i \in S_k} := \{y^i - \hat{f}_k(x^i)\}_{i \in S_k}$.
    end for
end for
Output: Radius $\zeta_n(x) \in \arg\min_{\zeta \in \Delta} \dfrac{1}{K} \displaystyle\sum_{k \in [K]} \frac{1}{|S_k|} \sum_{i \in S_k} c(\hat{z}^{DRO}_{-k}(x, \zeta), \hat{f}_k(x) + \hat{\varepsilon}^i_k)$ for the ambiguity set $\hat{\mathcal{P}}_n(x)$ for the ER-DRO problem (7).

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 presents an alternative that also uses the realization $x \in \mathcal{X}$ to choose $\zeta_n(x)$. 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. [23, Section 7.2.2]). This algorithm does not require estimation of the regression function f^*. The radius ζ_n determined using Algorithm 1 necessarily converges to zero as the sample size n increases. This may result in suboptimal estimators $\hat{z}_n^{DRO}(x)$ when the prediction model is misspecified (cf. Remark 1 in Section 2), in which case it may be beneficial to use a positive value of ζ_n even for large values of n (cf. Figure 7 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 3 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. The motivation for estimating the function f^* a second time is to approximate the following radius selection problem that uses f^* only to construct $|S_k|$ i.i.d. samples from the conditional distribution $P_{Y|X=x}$ for evaluating the quality of the K-fold CV-based ER-DRO solutions $\hat{z}_{-k}^{DRO}(x,\zeta)$:

$$\zeta_n^*(x) \in \arg\min_{\zeta \in \Delta} \frac{1}{K} \sum_{k \in [K]} \frac{1}{|S_k|} \sum_{i \in S_k} c(\hat{z}_{-k}^{DRO}(x,\zeta), f^*(x) + \varepsilon_k^i).$$

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).

7 Computational experiments

We consider instances of the following mean-risk portfolio optimization model adapted from [23]:

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

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

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

We can rewrite this model as the following single-stage stochastic program:

$$\min_{z \in \mathcal{Z}, \tau \in \mathbb{R}} \mathbb{E}\left[-Y^{\mathrm{T}} z + \rho\tau + \frac{\rho}{1-\beta} \max\{0, -Y^{\mathrm{T}} z - \tau\}\right].$$

The variable τ can be bounded under mild conditions on the distribution of Y (see [46, Theorem 10]). For each $j \in [d_z]$, the decision variable z_j denotes the fraction of capital invested in asset j and the random variable Y_j denotes the net return of asset j. 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 [23], we use $\beta = 0.8$, $\rho = 10$, and $d_y = d_z = 10$.

Similar to [34], we assume that the returns Y satisfy the relationship

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

where X_l, $l \in \mathcal{L}$, are covariates, $\theta \in \{0.5, 1, 2\}$ is a fixed parameter that determines the model class, $\bar{\varepsilon}_j \sim \mathcal{N}(0, 0.025j)$ and $\omega \sim \mathcal{N}(0, 0.02)$ are additive errors whose variances are chosen to match the case study in [23, Section 7.2], ν^* and μ_θ^* are model parameters, and $\mathcal{L}^* \subseteq \mathcal{L}$ contains the indices of the covariates with predictive power ($\mathcal{L}^*$ does not depend on the index $j \in [d_y]$). Note that $|\mathcal{L}| = d_x$. 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 [39]. 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

$$\begin{aligned} \nu_j^* &= 0.005j, & \mu_{\theta,j2}^* &= \big(0.0075j\big) s_\theta \xi_{j2}, \\ \mu_{\theta,j3}^* &= \big(0.005j\big) s_\theta \xi_{j3}, & \mu_{\theta,j1}^* &= 0.025j s_\theta - \mu_{\theta,j2}^* - \mu_{\theta,j3}^* \end{aligned}$$

for each $j \in [d_y]$, where ξ_{j2} and ξ_{j3} are i.i.d. samples from the uniform distribution $U(0.8, 1.2)$ and the scaling factor s_θ is (approximately) 1.25, 1.22, and 1 when the exponent θ is equal to 1, 0.5, and 2, respectively. The above coefficients are chosen such that $\mathbb{E}_{X,\bar{\varepsilon},\omega}[Y_j \mid X] = 0.03j$, $\forall j \in [d_y]$, which mirrors the setup in [23, Section 7.2] (the scaling factor s_θ offsets the differences in the term $\mathbb{E}_X[(X_l)^\theta]$ for $\theta \in \{1, 0.5, 2\}$). Once the coefficients ν^* and μ_θ^* are generated, they are considered fixed for different replications of the data $\mathcal{D}_n$.

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, Lasso, or Ridge regression and use this model within our residuals-based formulations. We use this linear model even when the degree $\theta \neq 1$, in which case it is misspecified. Note that OLS, Lasso, and Ridge regression estimate $d_x + 1$ 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 $\hat{P}_n^{ER}(x)$ 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 [23]. 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, 2, and 3 to specify the radii $\zeta_n(x)$ of the above ambiguity sets for the ER-DRO problem (7) with $K = 5$ folds in all three cases and $T = \min\{50, \lfloor \frac{n}{5} \rfloor\}$ in Algorithm 2. We use Lasso regression in line 7 of Algorithm 3 with 5-fold CV. For all ER-DRO formulations, following [23], 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 $\mathbb{R}_+$.

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 out-of-sample optimality gaps using the multiple replication procedure (MRP) [40] (see Algorithm 4 in Appendix C for details). We use 20,000 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 of the 99% UCBs, and the whiskers denote the 5^{th} and 95^{th} percentiles of the 99% UCBs over the 1000 instances.

Source code and data are available at https://github.com/rohitkannan/ER-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

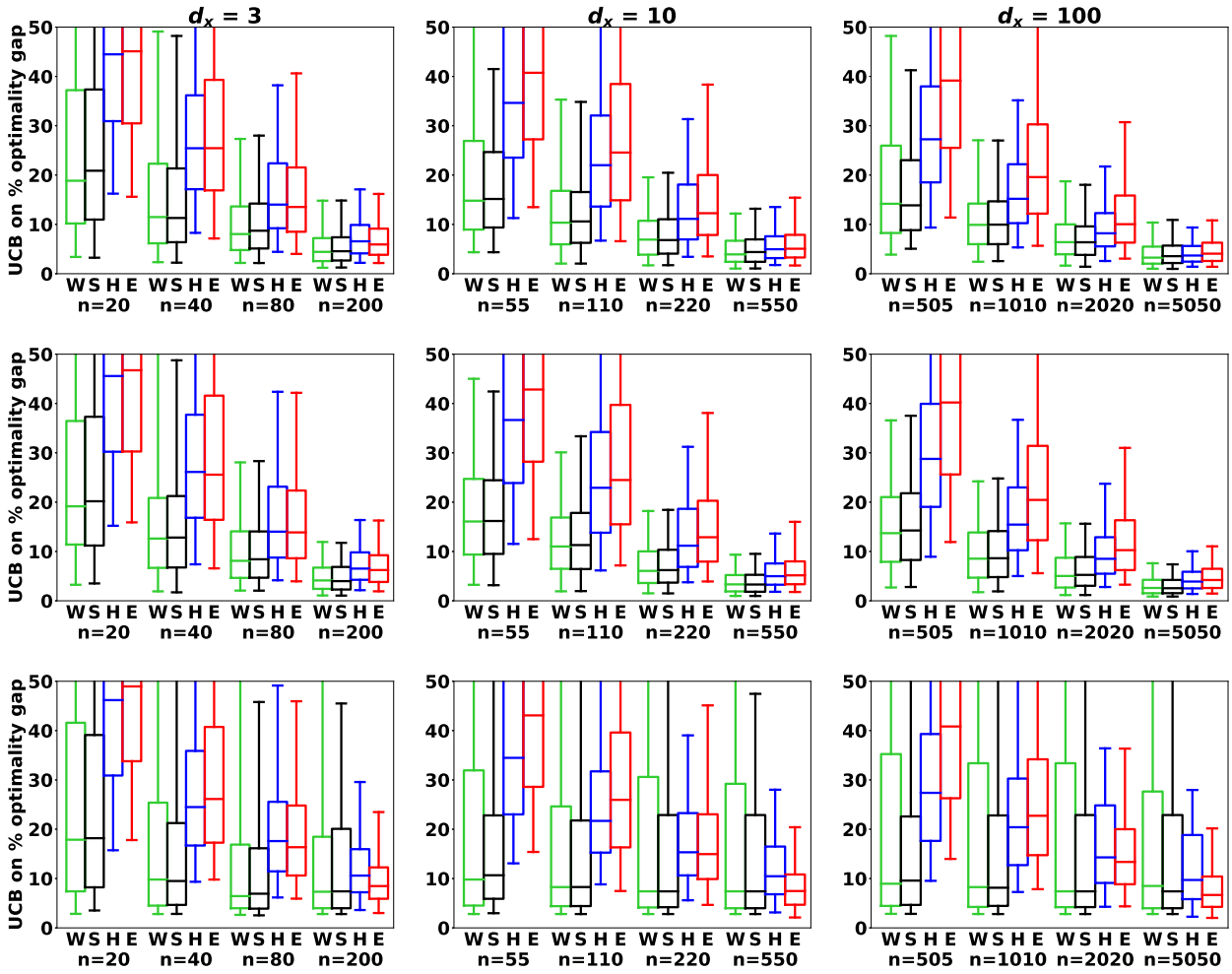

Figure 1: **(Comparison of the different ER-DRO formulations)** Comparison of the E+OLS approach (E) with the covariate-independent tuning of the W+OLS radius (W), the S+OLS radius (S), and the H+OLS radius (H), all tuned using Algorithm 2. Top row: $\theta = 1$. Middle row: $\theta = 0.5$. Bottom row: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

interface [22], and use `glmnet` 0.3.0 [26] for Lasso and Ridge regression. All computational tests were conducted through the UW-Madison Center for High Throughput Computing (CHTC) software `HTCondor` (`http://chtc.cs.wisc.edu/`).

Comparison of the different ER-DRO formulations. Figure 1 compares the performance of the E+OLS formulation with the W+OLS, S+OLS, and H+OLS formulations when the radius ζ_n of the ambiguity sets of all three ER-DRO formulations are specified using Algorithm 2. We vary the model degree θ, the covariate dimension among $d_x \in \{3, 10, 100\}$, and the sample size among $n \in \{5(d_x+1), 10(d_x+1), 20(d_x+1), 50(d_x+1)\}$ in these experiments. Note that OLS regression estimates $d_x + 1$ parameters for each $j \in \mathcal{J}$ even though the true model only contains $|\mathcal{L}^*| + 1 = 4$ nonzero parameters for each j. The performance of the S+OLS formulation is similar to that of the W+OLS formulation with the S+OLS formulation performing slightly better when $\theta = 2$. The H+OLS formulation does not significantly improve over the E+OLS formulation for smaller covariate dimensions but provides an intermediate level of improvement relative to the W+OLS and S+OLS approaches for larger covariate dimensions. Recall that the Wasserstein (W) and sample robust

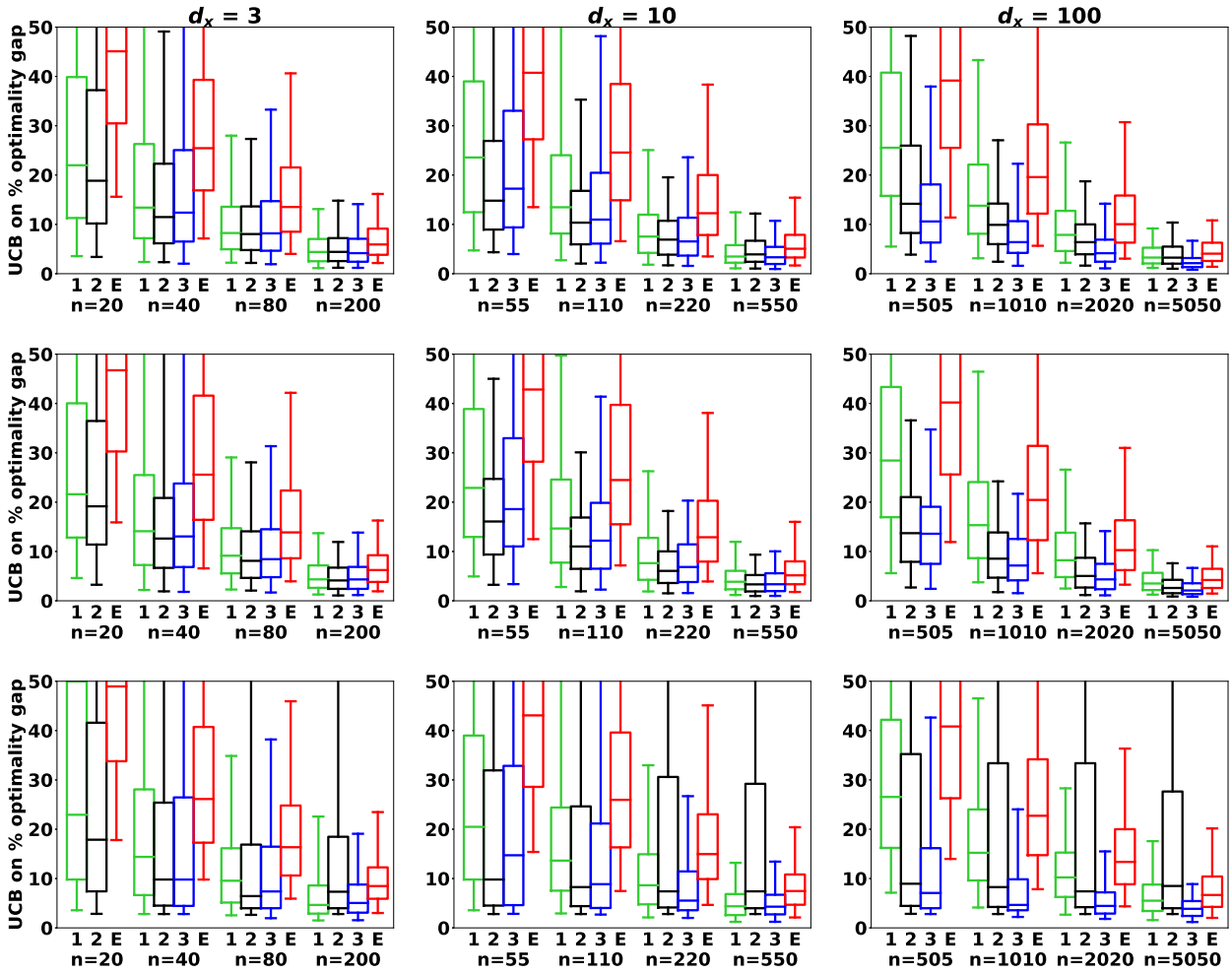

Figure 2: **(Wasserstein-DRO with OLS regression)** Comparison of the E+OLS approach (E) with the tuning of the W+OLS radius using Algorithms 1 (1), 2 (2), and 3 (3). Top row: $\theta = 1$. Middle row: $\theta = 0.5$. Bottom row: $\theta = 2$. Left column: $d_x = 3$. Middle column: $d_x = 10$. Right column: $d_x = 100$.

optimization (S) ambiguity sets allow distributions with support different from $\hat{P}_n^{ER}(x)$, whereas the 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 more results for the S formulations because they are similar to those of the W formulations. We also do not consider the H formulations further.

The optimality gaps of the E+OLS and the ER-DRO+OLS estimators are not guaranteed to converge to zero whenever $\theta \neq 1$ due to model misspecification; however, the W+OLS and S+OLS formulations are able to effectively mitigate the impact of model misspecification, especially for $\theta = 0.5$, in this case study. Interestingly, choosing the radius ζ_n of the ambiguity sets of all three ER-DRO formulations using Algorithm 2 performs worse than the E+OLS formulation when $\theta = 2$ and n is large. This indicates that Algorithm 2 may not yield a good choice of the radius ζ_n for large sample sizes when model misspecification is significant. From Figure 2, we see that Algorithm 3 provides a better-performing alternative to Algorithm 3 in this regime. Note that relatively large optimality gaps for DRO at small sample sizes n is to be expected for this case study because of the risk-averse nature of the portfolio objective.

Evaluation of the different radius selection strategies[8]. Figure 2 compares the performance of the E+OLS formulation with the W+OLS formulation over the same range of parameter values as in Figure 1. The radius $\zeta_n(x)$ of the ambiguity set for the W+OLS formulation is determined using Algorithms 1 and 2 that pick covariate-independent radii and Algorithm 3 that picks a covariate-dependent radius. Note that all three strategies for choosing the radius $\zeta_n(x)$ are based on the same underlying Wasserstein ER-DRO formulation with OLS regression. In this small sample regime, the W+OLS formulations perform better than the E+OLS formulation across almost all cases—the only exception again is when Algorithm 2 is used for $\theta = 2$ and n large. The radius specified by Algorithm 2 performs better than the radius specified using Algorithm 1 in all other cases, with the difference being most significant for larger covariate dimensions and smaller sample sizes. The radius specified by Algorithm 2 performs better than the radius specified using Algorithm 3 for smaller sample sizes and covariate dimensions, and the converse holds for larger covariate dimensions and sample sizes. When $\theta \neq 1$, the E+OLS and W+OLS approaches are not guaranteed to yield consistent estimators because the regression model is misspecified; however, Figure 2 shows that the W+OLS formulation with the radius ζ_n specified by Algorithm 2 is able to effectively mitigate the impact of model misspecification for $\theta = 0.5$ in this case study. Similarly, W+OLS with Algorithm 3 is able to effectively mitigate the impact of model misspecification for both $\theta = 0.5$ and $\theta = 2$. Finally, as expected, the benefits of the ER-DRO formulations diminish with increasing sample size.

Impact of the prediction step. We now highlight the modularity of our ER-DRO framework by exploring the potential benefits of regularization-based methods for estimating f^*. Figure 3 compares the performance of the E+Lasso approach with the W+Lasso approach, whereas Figure 5 in Appendix C compares the performance of the E+Ridge approach with the W+Ridge approach. The radius ζ_n of the ambiguity set for these W formulations is determined using Algorithms 1, 2, and 3. We consider $d_x \in \{3, 10, 100\}$, vary the model degree θ, and vary the sample size among $n \in \{3(d_x + 1), 5(d_x + 1), 10(d_x + 1), 20(d_x + 1)\}$ in these experiments[9]. We consider smaller sample sizes because the Lasso and Ridge regression are most effective in this regime. These experiments also illustrate the modularity of our residuals-based formulations. The W+Lasso and W+Ridge formulations outperform the E+Lasso and E+Ridge formulations, respectively, when the sample size n is small relative to the covariate dimension d_x (except when Algorithm 2 is used for $\theta = 2$ and n large). Note that the y-axis limits are different across the subplots in Figures 3 and 5. Regularization as in the case of Lasso and Ridge regression reduces the variance of the regression estimate $\hat{f}_n$ at the expense of an increase in its bias. Since the ambiguity sets of our residuals-based ER-DRO formulations do not *explicitly* address the uncertainty in the coefficients of the regression estimate $\hat{f}_n$, trading the variance of the coefficient estimates for some bias can result in ER-DRO estimators with better out-of-sample performance.

Wasserstein-DRO certificates. Figure 4 compares the normalized optimal objective value $100(\hat{v}_n^{ER}(x) - v^*(x))$ of the E+OLS formulation with the normalized optimal objective value $100(\hat{v}_n^{DRO}(x) - v^*(x))$ of the W+OLS formulation 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 $n \in \{5(d_x + 1), 10(d_x + 1), 20(d_x + 1), 50(d_x + 1)\}$ in these experiments. We omit the results for smaller covariate dimensions for brevity. Note that the y-axis limits are different across the subplots. First, we see that the ER-SAA solutions are optimistically biased and the bias reduces with increasing sample size (cf. [11, 23, 40]). Second, we see that ER-DRO solutions tend to err on the side of caution, with the pessimism of the W+OLS formulation shrinking to zero for $\theta = 1$ and $\theta = 0.5$ as the sample size increases. Finally, the pessimistic bias of the W+OLS formulation does not rapidly shrink to zero for $\theta = 2$ because the radius ζ_n specified using Algorithm 2 does not shrink to zero for this case due to significant model misspecification (cf. Figure 7 in Appendix C).

[8] We do not include results for Algorithm 3 with $n = 20$ because it requires at least 30 samples (line 7 of Algorithm 3 needs at least 6 points for Lasso regression with 5-fold CV).

[9] Once again, we only report results for Algorithm 3 when $n \geq 30$.

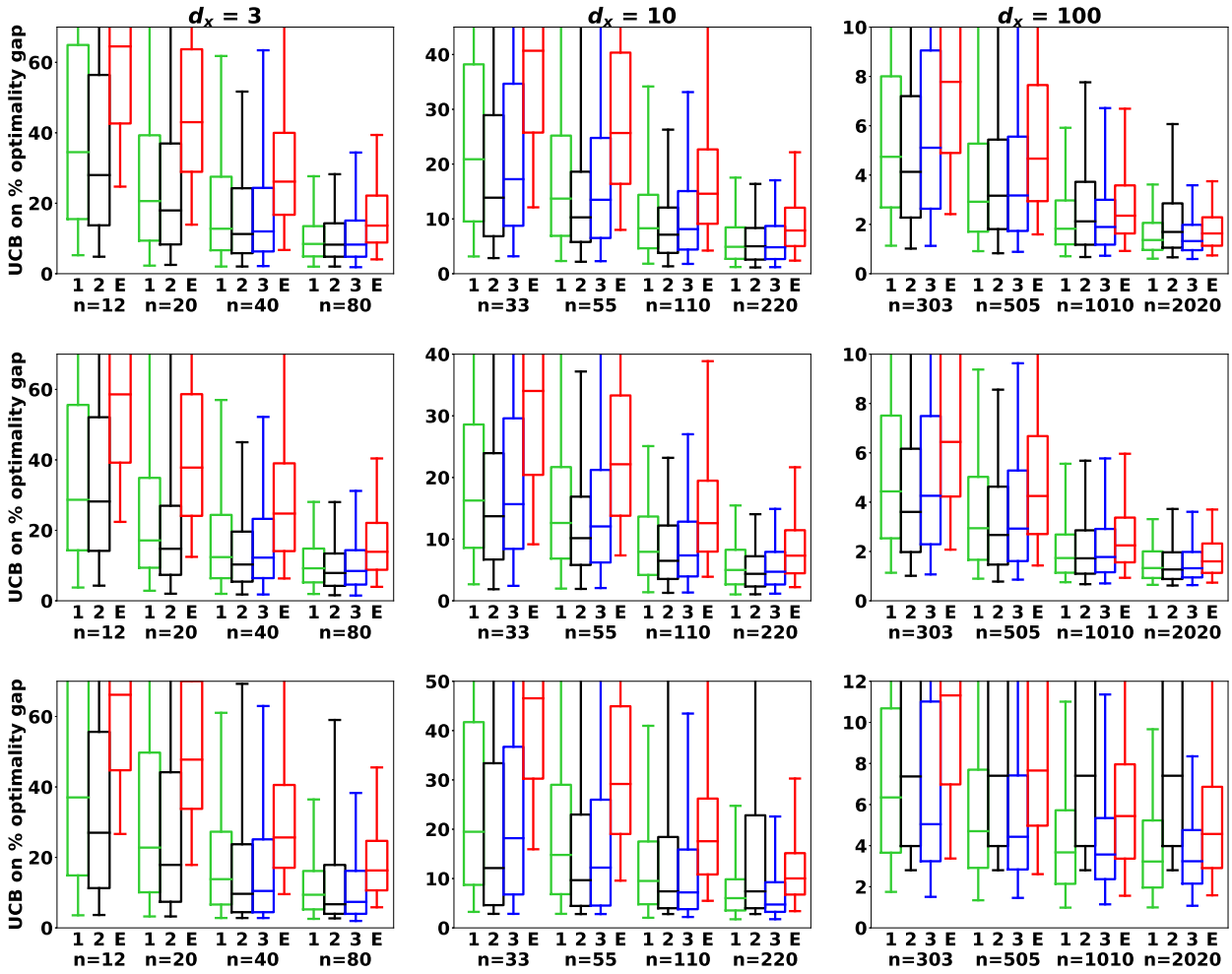

Figure 3: **(Wasserstein-DRO with the Lasso)** Comparison of the E+Lasso approach (E) with the tuning of the W+Lasso radius using Algorithms 1 (1), 2 (2), and 3 (3). Top row: $\theta = 1$. Middle row: $\theta = 0.5$. Bottom row: $\theta = 2$. Left column: $d_x = 3$. Middle column: $d_x = 10$. Right column: $d_x = 100$.

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 new data-driven approaches for sizing our ambiguity sets that do not require samples from the conditional distribution of the random variables. Numerical experiments illustrate that our residuals-based Wasserstein and sample robust optimization DRO formulations can significantly outperform the ER-SAA formulation in the limited data regime. We conclude that the ER-DRO and ER-SAA approaches are complementary. With limited data, the ER-DRO approach can yield better solutions. On the other hand, the value of ER-DRO over ER-SAA diminishes if there is ample data available, and the ER-SAA formulation remains tractable under milder assumptions on the true problem (3) compared to the Wasserstein and sample robust optimization-based ER-DRO formulations. In particular, these ER-DRO formulations generally result

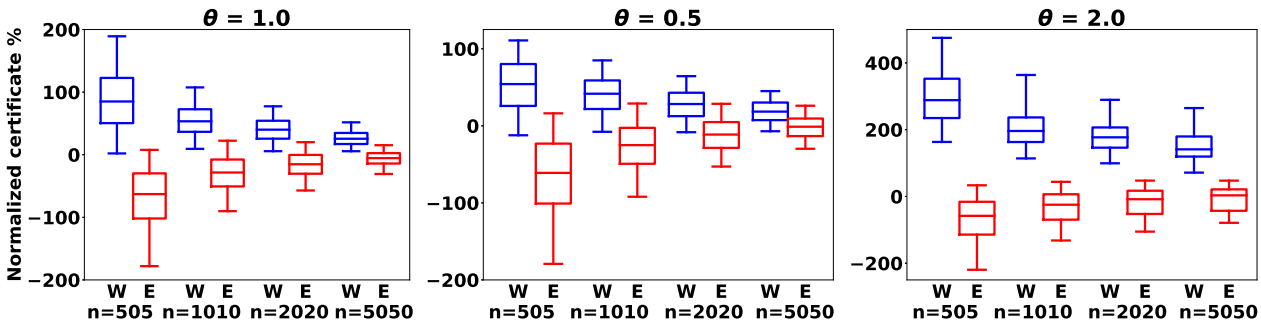

Figure 4: **(Wasserstein-DRO certificate)** Comparison of the E+OLS approach (E) with the covariate-independent tuning of the W+OLS Wasserstein radius using Algorithm 2 (W) for $d_x = 100$. Left: $\theta = 1$. Middle: $\theta = 0.5$. Right: $\theta = 2$.

in NP-hard formulations for two-stage stochastic programs and hence may require approximations [36, 43].

Deriving sharper finite sample guarantees for Wasserstein ER-DRO is an interesting avenue for future work. Extensions of the ER-SAA and ER-DRO formulations to multi-stage stochastic programming (cf. [10]), for the case when decisions affect the realizations of the random variables (cf. [7]), and for problems with stochastic constraints (cf. [33]) also merit further investigation.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR) under Contract No. DE-AC02-06CH11357. R.K. also gratefully acknowledges the support of the U.S. Department of Energy through the LANL/LDRD Program and the Center for Nonlinear Studies. This research was performed using the computing resources of the UW-Madison CHTC in the CS Department. 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. The authors thank the three anonymous reviewers for suggestions that helped improve the readability of this paper. R.K. also thanks Nam Ho-Nguyen for helpful discussions.

References

[1] G.-Y. Ban and C. Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67 (1):90–108, 2018.

[2] G.-Y. Ban, J. Gallien, and A. J. Mersereau. Dynamic procurement of new products with covariate information: The residual tree method. *Manufacturing & Service Operations Management*, 21(4):798–815, 2019.

[3] M. Bansal, K.-L. Huang, and S. Mehrotra. Decomposition algorithms for two-stage distributionally robust mixed binary programs. *SIAM Journal on Optimization*, 28(3):2360–2383, 2018.

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

[5] T. Bazier-Matte and E. Delage. Generalization bounds for regularized portfolio selection with market side information. *INFOR: Information Systems and Operational Research*, pages 1–28, 2020.

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

[7] D. Bertsimas and N. Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

[8] D. Bertsimas and B. Van Parys. Bootstrap robust prescriptive analytics. *Mathematical Programming*, pages 1–40, 2021.

[9] D. Bertsimas, V. Gupta, and N. Kallus. Robust sample average approximation. *Mathematical Programming*, 171(1-2): 217–282, 2018.

[10] D. Bertsimas, C. McCord, and B. Sturt. Dynamic optimization with side information. *European Journal of Operational Research*, 2022.

[11] D. Bertsimas, S. Shtern, and B. Sturt. A data-driven approach to multistage stochastic linear optimization. *Management Science*, 2022.

[12] D. Bertsimas, S. Shtern, and B. Sturt. Two-stage sample robust optimization. *Operations Research*, 70(1):624–640, 2022.

[13] J. Bezanson, A. Edelman, S. Karpinski, and V. Shah. Julia: a fresh approach to numerical computing. *SIAM Review*, 59 (1):65–98, 2017.

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

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

[16] J. Blanchet, K. Murthy, and N. Si. Confidence regions in Wasserstein distributionally robust estimation. *Biometrika*, 04 2021. ISSN 0006-3444. URL https://doi.org/10.1093/biomet/asab026.

[17] D. Boskos, J. Cortés, and S. Martínez. Data-driven ambiguity sets with probabilistic guarantees for dynamic processes. *IEEE Transactions on Automatic Control*, 66(7):2991–3006, 2020.

[18] D. Boskos, J. Cortés, and S. Martínez. Data-driven ambiguity sets for linear systems under disturbances and noisy observations. In *2020 American Control Conference (ACC)*, pages 4491–4496. IEEE, 2020.

[19] D. Boskos, J. Cortés, and S. Martínez. High-confidence data-driven ambiguity sets for time-varying linear systems. *arXiv preprint arXiv:2102.01142*, 2021.

[20] X. Dou and M. Anitescu. Distributionally robust optimization with correlated data from vector autoregressive processes. *Operations Research Letters*, 47(4):294–299, 2019.

[21] J. Duchi, P. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

[22] I. Dunning, J. Huchette, and M. Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59 (2):295–320, 2017.

[23] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[24] A. Esteban-Pérez and J. M. Morales. Distributionally robust stochastic programs with side information based on trimmings. *Mathematical Programming*, pages 1–37, 2021.

[25] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[26] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[27] R. Gao. 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.

[28] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[29] R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

[30] J.-y. Gotoh, M. J. Kim, and A. E. Lim. Calibration of distributionally robust empirical optimization models. *Operations Research*, 69(5):1630–1650, 2021.

[31] G. A. Hanasusanto and D. Kuhn. Robust data-driven dynamic programming. In *Advances in Neural Information Processing Systems*, pages 827–835, 2013.

[32] G. A. Hanasusanto and D. Kuhn. Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. *Operations Research*, 66(3):849–869, 2018.

[33] T. Homem-de-Mello and G. Bayraksan. Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.

[34] R. Kannan, G. Bayraksan, and J. Luedtke. Data-driven sample average approximation with covariate information. Optimization Online. URL: http://www.optimization-online.org/DB_HTML/2020/07/7932.html, 2020.

[35] R. Kannan, G. Bayraksan, and J. Luedtke. Heteroscedasticity-aware residuals-based contextual stochastic optimization. *arXiv preprint arXiv:2101.03139*, 2021.

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

[37] H. Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

[38] H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

[39] D. Lewandowski, D. Kurowicka, and H. Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

[40] W.-K. Mak, D. P. Morton, and R. K. Wood. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24(1-2):47–56, 1999.

[41] V. A. Nguyen, F. Zhang, J. Blanchet, E. Delage, and Y. Ye. Distributionally robust local non-parametric conditional estimation. *Advances in Neural Information Processing Systems*, 33:15232–15242, 2020.

[42] G. Pflug and D. Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[43] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

[44] P. Rigollet and J.-C. Hütter. High dimensional statistics. Lecture notes for MIT's 18.657 course, 2017. URL http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf.

[45] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[46] R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

[47] S. Sen and Y. Deng. 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.

[48] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[49] N. G. Trillos and D. Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. *Canadian Journal of Mathematics*, 67(6):1358–1383, 2015.

[50] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996.

[51] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[52] W. Xie. Tractable reformulations of two-stage distributionally robust linear programs over the type-∞ Wasserstein ball. *Operations Research Letters*, 48(4):513–523, 2020.

[53] H. Xu, C. Caramanis, and S. Mannor. A distributional interpretation of robust optimization. *Mathematics of Operations Research*, 37(1):95–110, 2012.

A Omitted Proofs

A.1 Proof of Proposition 1

We begin with the following useful results.

Lemma 22. Let $a_1, a_2, \ldots, a_d$ be positive constants with $a_j \geq 1$, $\forall j \in [d]$. Then, we have

$$\Big(\sum_{i=1}^{d} a_i^2\Big)^{\theta/2} \leq \sum_{i=1}^{d} a_i^{1+\frac{\theta}{2}}, \quad \forall \theta \in [1,2].$$

Proof. Let $F(\theta) := \big(\sum_{i=1}^{d} a_i^{1+\frac{\theta}{2}}\big)^{2/\theta}$ and $G(\theta) := \log(F(\theta)) = \frac{2}{\theta}\log\big(\sum_{i=1}^{d} a_i^{1+\frac{\theta}{2}}\big)$. The stated result holds if F (or equivalently, G) is nonincreasing on $\theta \in [1,2]$. We have

$$G'(\theta) = \frac{1}{\theta}\Big[\sum_{i=1}^{d} w_i \log(a_i) - \frac{2}{\theta}\log\Big(\sum_{i=1}^{d} a_i^{1+\frac{\theta}{2}}\Big)\Big] = \frac{1}{\theta}\Big[\log\Big(\prod_{i=1}^{d}(a_i)^{w_i}\Big) - \frac{2}{\theta}\log\Big(\sum_{i=1}^{d} a_i^{1+\frac{\theta}{2}}\Big)\Big],$$

where the nonnegative weight $w_i := \big(a_i^{1+\theta/2}\big)\big(\sum_{j=1}^{d} a_j^{1+\theta/2}\big)^{-1}$. Note that $w_i \in (0,1)$ and $\sum_{i=1}^{d} w_i = 1$. For $\theta \in [1,2]$, the above expression implies that $G'(\theta) \leq 0$ whenever $\prod_{i=1}^{d}(a_i)^{w_i} \leq \sum_{i=1}^{d} a_i^{1+\theta/2}$. We have

$$\prod_{i=1}^{d}(a_i)^{w_i} \leq \sum_{i=1}^{d} w_i a_i \leq \sum_{i=1}^{d} a_i \leq \sum_{i=1}^{d} a_i^{1+\frac{\theta}{2}},$$

where the first inequality follows from the weighted AM-GM inequality, the second inequality follows from the fact that $0 < w_i < 1$, $\forall i \in [d]$, and the final inequality follows from the assumption that $a_i \geq 1$, $\forall i \in [d]$. Consequently, $G'(\theta) \leq 0$, $\forall \theta \in [1,2]$, which implies F is nonincreasing on $\theta \in [1,2]$. □

Lemma 23. Let W be a sub-Gaussian random variable with variance proxy σ_w^2. Then

$$\mathbb{E}\Big[\exp\big(|W|^{1+\frac{\theta}{2}}\big)\Big] < +\infty, \quad \forall \theta \in (1,2).$$

Proof. We have

$$\mathbb{E}\Big[\exp\big(|W|^{1+\frac{\theta}{2}}\big)\Big] = \mathbb{E}\Big[\sum_{j=0}^{\infty} \frac{|W|^{(1+\frac{\theta}{2})j}}{j!}\Big] \tag{13}$$

Lemma 1.4 of [44] implies that

$$\mathbb{E}\Big[\frac{|W|^{(1+\frac{\theta}{2})j}}{j!}\Big] \leq \frac{\big(\sigma_w^2 e^{\frac{2}{e}} j\big)^{(1+\frac{\theta}{2})\frac{j}{2}}}{j!}, \quad \forall j \geq 2,$$

where $e := \exp(1)$. Therefore, inequality (13), the Fubini-Tonelli theorem, and the ratio test imply the stated result whenever $\lim_{j\to\infty} t_{j+1}/t_j < 1$, where $t_j := \left(\sigma_w^2 e^{\frac{2}{e}} j\right)^{(1+\frac{\theta}{2})\frac{j}{2}} (j!)^{-1}$. Let $C := \sigma_w^2 e^{\frac{2}{e}}$. We have

$$\begin{aligned}
\lim_{j\to\infty} \frac{t_{j+1}}{t_j} &= \lim_{j\to\infty} \frac{\left(C(j+1)\right)^{(1+\frac{\theta}{2})\frac{j+1}{2}}}{\left(Cj\right)^{(1+\frac{\theta}{2})\frac{j}{2}}} \frac{j!}{(j+1)!} \\
&= \lim_{j\to\infty} \frac{\left(C(j+1)\right)^{0.5(1+\frac{\theta}{2})}}{j+1} \left(\frac{j+1}{j}\right)^{(1+\frac{\theta}{2})\frac{j}{2}} \\
&= \lim_{j\to\infty} O(1)(j+1)^{0.5(\frac{\theta}{2}-1)} \lim_{j\to\infty} \left(\frac{j+1}{j}\right)^{(1+\frac{\theta}{2})\frac{j}{2}} \\
&= O(1) \lim_{j\to\infty} (j+1)^{0.5(\frac{\theta}{2}-1)} \lim_{j\to\infty} \left(1 + \frac{(1+\frac{\theta}{2})\frac{1}{2}}{(1+\frac{\theta}{2})\frac{j}{2}}\right)^{(1+\frac{\theta}{2})\frac{j}{2}} \\
&= 0 \times O(1) = 0,
\end{aligned}$$

where we use the fact $\theta \in (1,2)$ in the last step. □

We are now ready to prove Proposition 1. To establish $\mathbb{E}\left[\exp(\|\varepsilon\|^a)\right] < +\infty$, it suffices to show that $\mathbb{E}\left[\exp(\|\varepsilon\|^a)\mathbb{1}_{\{\|\varepsilon\|_\infty \geq 1\}}\right] < +\infty$, where $\mathbb{1}_{\{\|\varepsilon\|_\infty \geq 1\}} = 1$ if $\|\varepsilon\|_\infty \geq 1$ and 0 otherwise. Lemma 22 implies that

$$\mathbb{E}\left[\exp(\|\varepsilon\|^a)\right] \leq \mathbb{E}\left[\exp\left(\sum_{i=1}^{d_y} |\varepsilon_i|^{1+\frac{a}{2}}\right)\right] = \mathbb{E}\left[\prod_{i=1}^{d_y} \exp\left(|\varepsilon_i|^{1+\frac{a}{2}}\right)\right].$$

Independence of the components of ε further implies

$$\mathbb{E}\left[\exp(\|\varepsilon\|^a)\right] \leq \prod_{i=1}^{d_y} \mathbb{E}\left[\exp\left(|\varepsilon_i|^{1+\frac{a}{2}}\right)\right].$$

Consequently, it suffices to show that $\mathbb{E}\left[\exp\left(|\varepsilon_i|^{1+\frac{a}{2}}\right)\mathbb{1}_{\{|\varepsilon_i|\geq 1\}}\right] < +\infty$ for each $i \in [d_y]$, which follows from Lemma 23. □

A.2 Proof of Lemma 8

We require the following result (cf. [23, Lemma 3.7]).

Lemma 24. Suppose Assumptions 1, 2, and 6 hold and the samples $\{\varepsilon^i\}_{i=1}^n$ are i.i.d. Let $\{Q_n(x)\}$ be a sequence of distributions with $Q_n(x) \in \hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x))$. Then

$$\mathbb{P}\Big\{d_{W,p}(P_{Y|X=x}, Q_n(x)) \leq 2\zeta_n(\alpha_n, x)\Big\} \geq 1 - \alpha_n, \quad \text{for a.e. } x \in \mathcal{X}.$$

Consequently, we a.s. have for n large enough:

$$d_{W,p}(P_{Y|X=x}, Q_n(x)) \leq 2\zeta_n(\alpha_n, x), \quad \text{for a.e. } x \in \mathcal{X}.$$

Furthermore, $\{Q_n(x)\}$ converges a.s. converges to $P_{Y|X=x}$ under the Wasserstein metric

$$\mathbb{P}\Big\{\lim_{n\to\infty} d_{W,p}(P_{Y|X=x}, Q_n(x)) = 0\Big\} = 1.$$

Proof. Mirrors the proof of [23, Lemma 3.7] on account of Theorem 7. □

We now prove Lemma 8. From Theorem 7, we have for a.e. $x \in \mathcal{X}$ that

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

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

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

From Lemma 24, 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 $d_{W,p}(P_{Y|X=x}, Q_n(x)) \le 2\zeta_n(\alpha_n,x)$ for a.e. $x \in \mathcal{X}$.

Suppose Assumption 4 holds. 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

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

where the second inequality follows from Assumption 4 and the Kantorovich-Rubinstein theorem (cf. [36, Theorem 5]).

Suppose instead that Assumption 5 holds and $p \in [2,+\infty)$. 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 $p \in [2,+\infty)$, we a.s. have for n large enough and a.e. $x \in \mathcal{X}$ that

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

where the latter inequality follows from Assumption 5 and [27, Lemma 2] (see also [29]). □

A.3 Proof of Theorem 9

From Theorem 7, we have

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

From Lemma 24, 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 $Q_n(x) \in \hat{\mathcal{P}}_n(x;\zeta_n(\alpha_n,x))$. Theorem 6.9 of [51] then a.s. implies that $Q_n(x)$ converges weakly to $P_{Y|X=x}$ in the space of distributions with finite pth moments for a.e. $x \in \mathcal{X}$.

Lemma 8 implies a.s. that for all n large enough

$$v^*(x) \le g(\hat{z}_n^{DRO}(x);x) \le \hat{v}_n^{DRO}(x), \quad \text{for a.e. } x \in \mathcal{X}. \tag{14}$$

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

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

We suppress the dependence of $Q_n^*(x)$ on η for simplicity. We a.s. have for a.e. $x \in \mathcal{X}$

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

The first equality above follows from the fact that $Q_n^*(x)$ converges weakly to $P_{Y|X=x}$ (as noted above) and by Definition 6.8 of [51] (which holds by virtue of Assumption 3). Since $\eta > 0$ was arbitrary, we conclude that $\limsup_{n\to\infty} \hat{v}_n^{DRO}(x) \leq v^*(x)$ a.s. 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 $\operatorname{dist}(\hat{z}_n^{DRO}(x), S^*(x)) \xrightarrow{a.s.} 0$ for a.e. $x \in \mathcal{X}$. From (14) and the above conclusion, 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 a.s. 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 $\operatorname{dist}(\hat{z}_n^{DRO}(x), S^*(x))$ 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 $x \in \bar{\mathcal{X}}$, $\operatorname{dist}(\hat{z}_n^{DRO}(x), S^*(x))$ 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 $\operatorname{dist}(\hat{z}_n^{DRO}(x), S^*(x))$ 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.4 Proof of Theorem 10

Lemma 8 implies that inequality (14) a.s. holds for all n large enough. From Lemma 24, we a.s. have for any distribution $Q_n(x) \in \hat{\mathcal{P}}_n(x; \zeta_n(\alpha_n, x))$ and sample size n large enough that $d_{W,p}(P_{Y|X=x}, Q_n(x)) \leq 2\zeta_n(\alpha_n, x)$ for a.e. $x \in \mathcal{X}$.

If Assumption 4 holds, then the desired result follows from inequality (14) and part A of Lemma 8. On the other hand, if Assumption 4 holds and $p \geq 2$, then the desired result follows from inequality (14) and part B of Lemma 8. □

A.5 Proof of Lemma 12

Theorem 7 implies $g(\hat{z}_n^{DRO}(x); x) \leq \hat{v}_n^{DRO}(x)$ with probability at least $1-\alpha$ when $\zeta_n(\alpha, x)$ is chosen according to equation (9). Lemma 5 then yields

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

Suppose Assumption 4 holds. Following the proof of part A of Lemma 8 (see Lemma 24), we have for any $z^*(x) \in S^*(x)$

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

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

$$\alpha \geq O(1)\exp\left(-O(1)n\left(\tfrac{\kappa}{4L_1(z^*(x))}\right)^{1/\theta}\right)$$

with θ equal to $\min\{p/d_y, 1/2\}$ or p/a. Assumption 8 implies that we can choose the constant $\kappa^{(1)}_{p,n}(\alpha, x)$ in equation (9) such that for a.e. $x \in \mathcal{X}$, $2L_1(z^*(x))\kappa^{(1)}_{p,n}(\alpha, x) \leq \kappa/2$ 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 imply the existence of constants $\tilde{\Gamma}(\kappa, x), \tilde{\gamma}(\kappa, x) > 0$ such that the risk level $\alpha = \tilde{\Gamma}(\kappa, x)\exp(-n\tilde{\gamma}(\kappa, x))$ satisfies $2L_1(z^*(x))\zeta_n(\alpha, x) \leq \kappa$. Consequently, (10) holds.

Next, suppose instead that Assumption 5 holds. Following the proof of part B of Lemma 8 (see Lemma 24), we have for any $z^*(x) \in S^*(x)$

$$\mathbb{P}\big\{\hat{v}_n^{DRO}(x) > v^*(x) + \big(\mathbb{E}\left[\|\nabla c(z^*(x), Y)\|^2\right]\big)^{1/2}\zeta_n(\alpha, x) + 4L_2(z^*(x))\zeta_n^2(\alpha, x)\big\} \leq \alpha.$$

Therefore, if we pick $\alpha \in (0, 1)$ so that

$$\big(\mathbb{E}\left[\|\nabla c(z^*(x), Y)\|^2\right]\big)^{1/2}\zeta_n(\alpha, x) + 4L_2(z^*(x))\zeta_n^2(\alpha, x) \leq \kappa,$$

then $\mathbb{P}\left\{g(\hat{z}_n^{DRO}(x); x) > v^*(x) + \kappa\right\} \leq 2\alpha$. Similar to the analysis above, positive constants $\tilde{\Gamma}(\kappa, x)$ and $\tilde{\gamma}(\kappa, x)$ and inequality (10) can be obtained by bounding the smallest value of α using Assumption 8 and equation (9) so that

$$\big(\mathbb{E}\left[\|\nabla c(z^*(x), Y)\|^2\right]\big)^{1/2}\zeta_n(\alpha, x) + 4L_2(z^*(x))\zeta_n^2(\alpha, x) \leq \kappa. \qquad \square$$

A.6 Proof of Lemma 14

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

$$\sup_{z\in\mathcal{Z}}\left|\hat{g}_{s,n}^{ER}(z; x) - g(z; x)\right| \leq \sup_{z\in\mathcal{Z}}\sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))}\sum_{i=1}^n p_i\Big|\sup_{y\in\hat{\mathcal{Y}}_n^i(x;\mu_n(x))} c(z, y) - c\left(z, f^*(x) + \varepsilon^i\right)\Big| \\ + \sup_{z\in\mathcal{Z}}\left|g_{s,n}^*(z; x) - g_n^*(z; x)\right| + \sup_{z\in\mathcal{Z}}|g_n^*(z; x) - g(z; x)|. \tag{15}$$

Consider the first term on the r.h.s. of (15). We have for each $x \in \mathcal{X}$

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

where the first step follows from Assumption 9, the second step follows from the definition of the set $\hat{\mathcal{Y}}_n^i(x;\mu_n(x))$, the triangle inequality, and inequality (6), and the fourth step follows by applying the Cauchy-Schwarz inequality twice.

Next, consider the second term on the r.h.s. of (15). For each $x \in \mathcal{X}$

$$
\begin{aligned}
&\sup_{z\in\mathcal{Z}} |g_{s,n}^*(z;x) - g_n^*(z;x)| \\
&= \sup_{z\in\mathcal{Z}} \Big| \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n p_i c(z, f^*(x)+\varepsilon^i) - \frac{1}{n}\sum_{i=1}^n c(z, f^*(x)+\varepsilon^i) \Big| \\
&= \sup_{z\in\mathcal{Z}} \Big| \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \sum_{i=1}^n \Big(p_i - \frac{1}{n}\Big) c(z, f^*(x)+\varepsilon^i) \Big| \\
&\leq \sup_{p\in\mathfrak{P}_n(x;\zeta_n(x))} \left(n \sum_{i=1}^n \Big(p_i - \frac{1}{n}\Big)^2 \right)^{\frac{1}{2}} \sup_{z\in\mathcal{Z}} \left(\frac{1}{n}\sum_{i=1}^n \big(c(z, f^*(x)+\varepsilon^i)\big)^2 \right)^{\frac{1}{2}},
\end{aligned}
$$

where the inequality follows by Cauchy-Schwarz. □

A.7 Proof of Theorem 18

Since

$$
\begin{aligned}
\big\|\hat{v}_n^{DRO}(X) - v^*(X)\big\|_{L^q} &= \big\| \min_{z\in\mathcal{Z}} \hat{g}_{s,n}^{ER}(z;X) - \min_{z\in\mathcal{Z}} g(z;X) \big\|_{L^q} \\
&\leq \Big\| \sup_{z\in\mathcal{Z}} \big|\hat{g}_{s,n}^{ER}(z;X) - g(z;X)\big| \Big\|_{L^q}, \qquad (16)
\end{aligned}
$$

we look to establish uniform rates of convergence of $\hat{g}_{s,n}^{ER}(\cdot;X)$ to $g(\cdot;X)$ with respect to the L^q-norm on $\mathcal{X}$. From (15) and the triangle inequality, we have

$$
\begin{aligned}
\Big\| \sup_{z\in\mathcal{Z}} \big|\hat{g}_{s,n}^{ER}(z;X) - g(z;X)\big| \Big\|_{L^q} \leq{}& \Big\| \sup_{z\in\mathcal{Z}} \big|\hat{g}_{s,n}^{ER}(z;x) - g_{s,n}^*(z;X)\big| \Big\|_{L^q} + \\
& \Big\| \sup_{z\in\mathcal{Z}} \big|g_{s,n}^*(z;X) - g_n^*(z;X)\big| \Big\|_{L^q} + \\
& \Big\| \sup_{z\in\mathcal{Z}} \big|g_n^*(z;X) - g(z;X)\big| \Big\|_{L^q}. \qquad (17)
\end{aligned}
$$

We bound the terms on the r.h.s. of (17) using Lemma 14. Assumptions 9, 16, and 17 and $\|\mu_n(X)\|_{L^q} = O(n^{-r/2})$ imply the first term on the r.h.s. of inequality (11) satisfies:

$$
\begin{aligned}
&\Big\| \sup_{z\in\mathcal{Z}} \big|\hat{g}_{s,n}^{ER}(z;X) - g_{s,n}^*(z;X)\big| \Big\|_{L^q} \\
&\leq \left\| \sup_{z\in\mathcal{Z}} L(z) \left(\mu_n(X) + \left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(X)\|^2 \right)^{\frac{1}{2}} \right) \sup_{p\in\mathfrak{P}_n(X;\zeta_n(X))} \left(1 + n\sum_{i=1}^n \Big(p_i - \frac{1}{n}\Big)^2 \right)^{\frac{1}{2}} \right\|_{L^q} \\
&= O(1) \left[\|\mu_n(X)\|_{L^q} + \left\| \left(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(X)\|^2 \right)^{\frac{1}{2}} \right\|_{L^q} \right] \\
&= O(1) \left[\|\mu_n(X)\|_{L^q} + \Big\| f^*(X) - \hat{f}_n(X) \Big\|_{L^q} + \left\| \left(\frac{1}{n}\sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 \right)^{\frac{1}{2}} \right\|_{L^q} \right] \\
&= O_p(n^{-r/2}).
\end{aligned}
$$

Assumptions 16 and 18 imply that the second term on the r.h.s. of inequality (11) satisfies

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

Finally, Assumption 14 implies

$$\left\| \sup_{z\in\mathcal{Z}} |g^*_n(z;X) - g(z;X)| \right\|_{L^q} = O_p(n^{-1/2}).$$

Putting the above three inequalities together into inequality (17), we obtain

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

The first part of the stated result then follows from (16). The second part of the stated result follows from (16) and the fact that

$$\begin{aligned}
\left\| g(\hat{z}^{DRO}_n(X);X) - v^*(X) \right\|_{L^q} \leq\ & \left\| g(\hat{z}^{DRO}_n(X);X) - \hat{v}^{DRO}_n(X) \right\|_{L^q} + \\
& \left\| \hat{v}^{DRO}_n(X) - v^*(X) \right\|_{L^q}.
\end{aligned}$$

□

B Ambiguity sets satisfying Assumption 12

In Section 5, we outlined conditions under which phi-divergence ambiguity sets $\mathfrak{P}_n(x;\zeta_n(x))$ satisfy Assumption 12 for a suitable choice of the radius $\zeta_n(x)$. Lemma 25 below determines sharp bounds on the radius $\zeta_n(x)$ for some other families of ambiguity sets to satisfy Assumption 12. Before presenting the lemma, 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 [48]: 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 25. The following ambiguity sets satisfy Assumption 12 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 $\zeta_n(x) < 0.5$. 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) The stated result follows from the fact that

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

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

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

(c) Let $\bar{q} := \frac{1}{n}\sum_{i=1}^n q_i$. We have:

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

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

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

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

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

Algorithm 4 Estimating the 99% UCB on the optimality gap of a solution

Input: Covariate realization $X = x$ and data-driven solution $\hat{z}_n(x)$ for a particular realization of the data $\mathcal{D}_n$.
Output: $\hat{B}_{99}(x)$, which is a normalized estimate of the 99% UCB on the out-of-sample optimality gap of $\hat{z}_n(x)$.
for $k = 1, \ldots, 30$ **do**
    Draw 10^5 i.i.d. samples $\bar{\mathcal{D}}^k := \{\bar{\varepsilon}^{k,i}\}_{i=1}^{10^5}$ of ε according to P_ε.
    Estimate the optimal value $v^*(x)$ by solving the full-information SAA problem (4) using the scenarios $\{f^*(x) + \bar{\varepsilon}^{k,i}\}_{i=1}^{10^5}$ constructed with $\bar{\mathcal{D}}^k$
    Estimate the out-of-sample cost of the solution $\hat{z}_n(x)$ using the first 20,000 samples of $\bar{\mathcal{D}}^k$, i.e., $\hat{v}^k(x) := \frac{1}{20000}\sum_{i=1}^{2\times 10^4} c(\hat{z}_n(x), f^*(x) + \bar{\varepsilon}^{k,i})$
    Estimate the optimality gap of the $\hat{z}_n(x)$ as $\hat{G}^k(x) = \hat{v}^k(x) - \bar{v}^k(x)$.
end for
Construct the estimate of the 99% UCB on the optimality gap of $\hat{z}_n(x)$ as

$$\hat{B}_{99}(x) := 100\Big(\texttt{avg}(\{\hat{G}^k(x)\}) + 2.462\sqrt{\texttt{var}(\{\hat{G}^k(x)\})/30}\Big),$$

where $\texttt{avg}(\{\hat{G}^k(x)\})/\texttt{var}(\{\hat{G}^k(x)\})$ denote the mean/variance of estimates.

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 $p_i = \frac{1}{n}\big[1 + q_i - \frac{1}{n}\sum_{j=1}^n q_j\big]$, $i \in [n]$, 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

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

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

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

□

C Additional computational results

Algorithm 4 describes our procedure for estimating the 99% UCB on the optimality gap of our data-driven solutions using the multiple replication procedure [40]. We only use 20,000 of the generated 10^5 samples

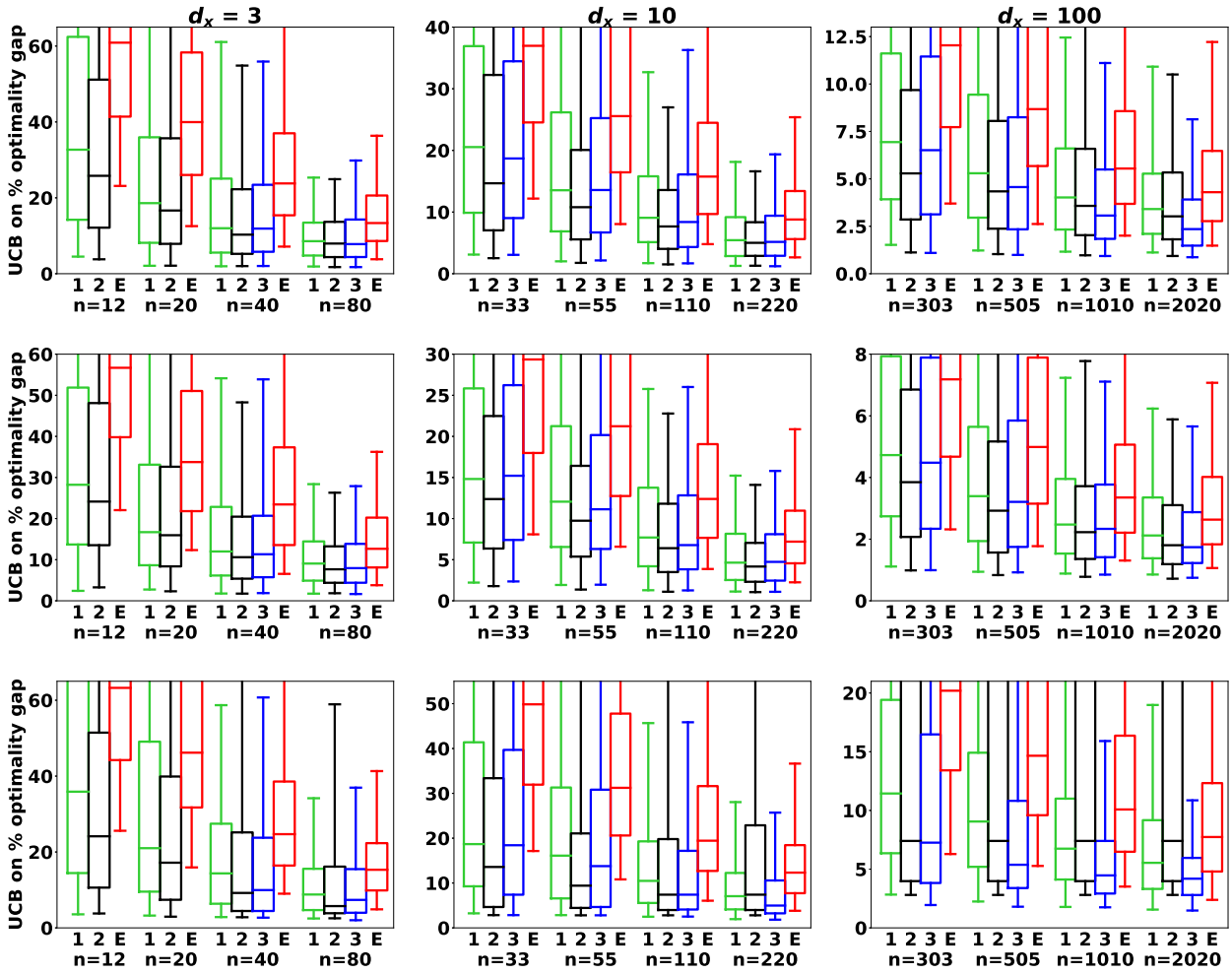

Figure 5: **(Wasserstein-DRO with Ridge regression)** Comparison of the E+Ridge approach (E) with tuning of the W+Ridge radius using Algorithms 1 (1), 2 (2), and 3 (3). Top row: $\theta = 1$. Middle row: $\theta = 0.5$. Bottom row: $\theta = 2$. Left column: $d_x = 3$. Middle column: $d_x = 10$. Right column: $d_x = 100$.

from the conditional distribution of Y given $X = x$ to compute these UCBs since they are sufficient to yield an accurate estimate of the optimality gaps. Unlike [34, Algorithm 1] that uses *relative* optimality gaps, we use *absolute* optimality gaps in our 99% UCB estimates to avoid division by small quantities when $v^*(x)$ is close to zero.

We compare Algorithm 2 with an "optimal covariate-independent" specification of the radius 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. We also benchmark Algorithm 3 against an "optimal covariate dependent" specification of $\zeta_n(x)$ that is determined by choosing $\zeta_n(x)$ such that the 99% UCBs are minimized. Determining these optimal covariate-independent and covariate-dependent radii $\zeta_n(x)$ is impractical because it requires 20,000 i.i.d. samples from the conditional distribution of Y given $X = x$ (which a decision-maker does not have). We consider it only to benchmark the performance of Algorithms 2 and 3.

"Optimal" tuning of the Wasserstein radius. Figure 6 compares the performance of the W+OLS formulations when the radius $\zeta_n(x)$ of the ambiguity set is determined using Algorithms 2 and 3 and optimal

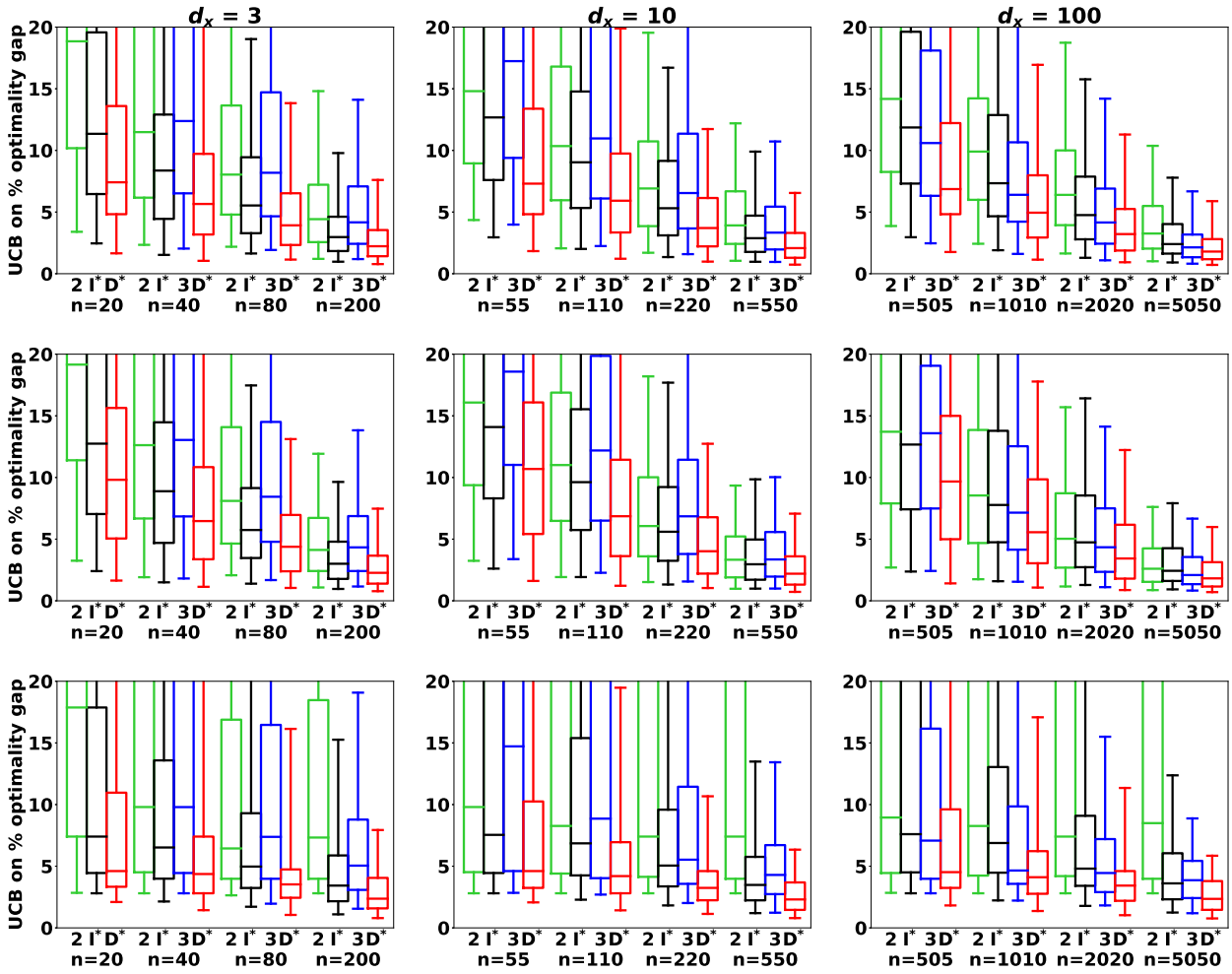

Figure 6: **(Comparison with "optimal" specification of the Wasserstein radius)** Comparison of the W+OLS approach with the optimal covariate-dependent (D*) and covariate-independent (I*) tuning of the W+OLS radius, and the tuning of the W+OLS radius using Algorithm 3 (3) and Algorithm 2 (2). Top row: $\theta = 1$. Middle row: $\theta = 0.5$. Bottom row: $\theta = 2$. Left column: $d_x = 3$. Middle column: $d_x = 10$. Right column: $d_x = 100$.

covariate-dependent and covariate-independent tuning. We vary the model degree θ, the covariate dimension among $d_x \in \{3, 10, 100\}$, and the sample size among $n \in \{5(d_x+1), 10(d_x+1), 20(d_x+1), 50(d_x+1)\}$ in these experiments. The radius specified by Algorithm 2 performs better than the radius specified using Algorithm 3 for smaller sample sizes and covariate dimensions, and the converse holds for larger covariate dimensions and sample sizes. These results indicate that while covariate-dependent tuning theoretically has potential to yield better results than the covariate-independent tuning of Algorithm 2, Algorithm 3 is only able to obtain good estimates of the optimal covariate-dependent radius $\zeta_n(x)$ for relatively large sample sizes n. The difference between the performance of Algorithm 2 and the optimal covariate-independent tuning reduces with increasing sample size and covariate dimension except for $\theta = 2$. The difference between the performance of Algorithm 3 and optimal covariate-dependent tuning of the radius also reduces with increasing covariate dimension and sample size.

Comparison of the radii specified by Algorithms 1, 2, and 3. Figure 7 compares the radii specified by Algorithms 1, 2, and 3 with the optimal covariate-dependent radius and optimal covariate-independent

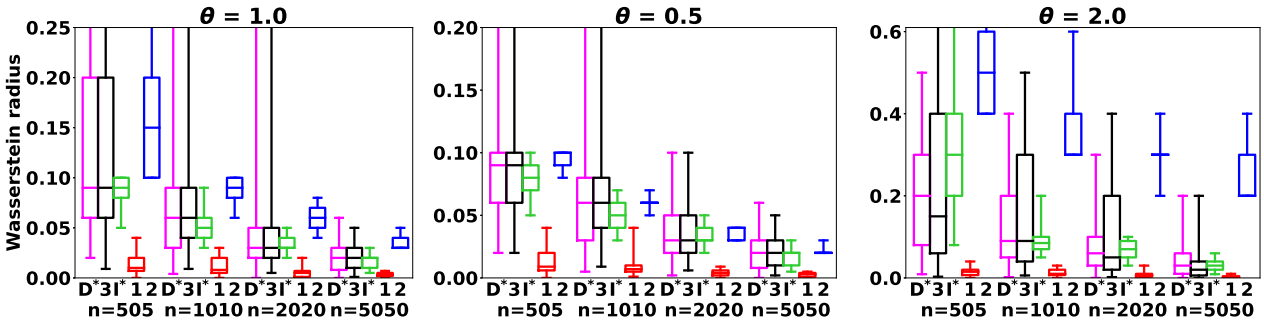

Figure 7: **(Comparison of the radii specified by Algorithms 1, 2, and 3)** Comparison of the optimal covariate-dependent tuning (D*) and optimal covariate-independent tuning (I*) of the W+OLS radius, the covariate-dependent tuning of the W+OLS radius using Algorithm 3 (3), and the covariate-independent 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: $\theta = 2$.

radius for the W+OLS formulation. We consider $d_x = 100$, vary the model degree θ, and vary the sample size among $n \in \{5(d_x + 1), 10(d_x + 1), 20(d_x + 1), 50(d_x + 1)\}$ in these experiments. Note that the y-axis limits are different across the subplots. 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 2 that the resulting ER-DRO estimators typically do not perform as well as the estimators obtained when the radius ζ_n is specified using Algorithms 2 and 3. Second, we see that the covariate-independent specifications of the radius result in more narrow distributions overall 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 radius determined using Algorithm 3 appears to converge to the distribution of the optimal covariate-dependent radius as the sample size increases. Similarly, the distribution of the radius determined using Algorithm 2 also appears to converge to the distribution of the optimal covariate-independent radius as n increases (except for the case when $\theta = 2$). 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 $\theta \neq 1$). 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.

Comparison with the jackknife-based formulations. Figure 8 compares the performance of the ER-SAA+OLS approach and the jackknife-based SAA (J-SAA+OLS) approach [34] 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 $n \in \{3(d_x+1), 5(d_x+1), 10(d_x+1), 20(d_x+1)\}$ in these experiments. As observed in [34], the J-SAA+OLS formulation performs better than the E+OLS formulation in the small sample size regime. Figure 8 shows that the W+OLS formulations outperform the J-SAA+OLS formulation (except when using Algorithm 2 for $\theta = 2$ and large n). This is 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 jackknife formulations do 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.

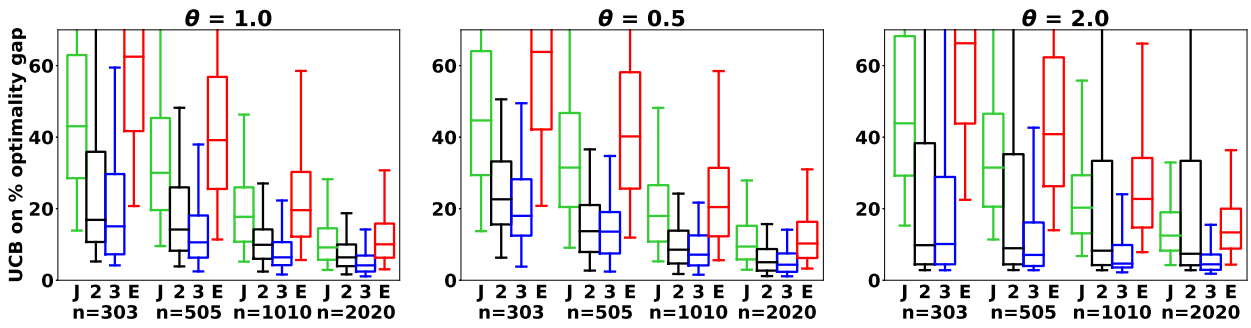

Figure 8: **(Comparison of Wasserstein-DRO with J-SAA)** Comparison of the E+OLS (E) and J+OLS (J) approaches with tuning of the W+OLS radius using Algorithms 2 (2) and 3 (3) for $d_x = 100$. Left: $\theta = 1$. Middle: $\theta = 0.5$. Right: $\theta = 2$.