Distributionally Robust Losses Against Mixture Covariate Shifts

John C. Duchi[1,2] Tatsunori Hashimoto[1] Hongseok Namkoong[3]

Stanford University
Departments of [1]Statistics, [2]Electrical Engineering,
and [3]Management Science and Engineering

{jduchi,thashim, hnamk}@stanford.edu

Abstract

Modern large-scale datasets are often collected over heterogeneous subpopulations, such as multiple demographic groups or multiple text corpora. Minimizing average loss over such datasets fails to guarantee uniformly low losses across all subpopulations. We propose a convex procedure that controls the worst-case performance over all subpopulations of a certain size. Our procedure comes with finite-sample optimality guarantees on the worst off subpopulation, and converges at the standard nonparametric rate. Empirically, we observe on lexical similarity and recidivism prediction tasks that our worst-case procedure learns models that do well against unseen subpopulations.

1 Introduction

When we train models over heterogeneous data, a basic goal is to train models that perform uniformly well across all subpopulations instead of just on average. For example, in natural language processing (NLP), large-scale corpora are often consist of data from multiple domains, each domain varying in difficulty and frequently containing large proportions of easy examples [19, 54]. In spite of this, we seek models that perform well on each sub-corpus rather than achieving good (average) performance by focusing on the easy examples and domains. Standard statistical learning approaches optimize average performance, however, only learning to accurately predict easy examples and sacrificing predictive performance on hard subpopulations [55].

The growing use of machine learning systems in core socioeconomic decision-making problems such as loan-service and recidivism prediction highlights the importance of models that perform well over different demographic groups [6]. In the face of this need, a number of authors observe that optimizing average performance often yields models that perform poorly on minority subpopulations [3, 30, 36, 15, 59, 67]. When datasets contain demographic information, a natural approach is to optimize worst-case loss or equalize losses over groups. But in many tasks—such as language identification or video analysis [67, 15]—privacy concerns preclude recording demographic or other sensitive information, limiting the applicability of methods that require knowledge of group identities.

In this work, we develop procedures that control performance over *all* large enough subpopulations, agnostic to the distribution of each subpopulation. We study a worst-case formulation over over *all* large enough subpopulations in the data, and provide procedures that automatically focus on the difficult subsets of the dataset. Our procedure guarantees a uniform level of performance across subpopulations by hedging against unseen covariate shifts, potentially even in the presence of confounding.

In classical statistical learning and prediction problems [33], we wish to predict a target $Y \in \mathcal{Y}$ from a covariate vector $X \in \mathcal{X} \subset \mathbb{R}^d$ drawn from some underlying population via

$(X, Y) \sim P$, and for a loss $\ell : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_+$ we wish to find a model $\theta \in \Theta$ minimizing $\mathbb{E}_P[\ell(\theta; (X, Y))]$. In contrast, we consider an elaborated setting in which the observed data comes from a mixture model and we evaluate model losses on a component (subpopulation) from this mixture. More precisely, we assume that

$$X \sim P_X := \alpha Q_0 + (1 - \alpha) Q_1,$$

where the subpopulation proportion $\alpha \in (0, 1)$, while the subpopulations Q_0 and Q_1 are unknown. The classical formulation does little to ensure equitable performance for both data X from Q_0 and Q_1, especially for small α. Thus for a fixed conditional distribution $P_{Y|X}$[1] and the loss ℓ, we instead seek $\theta \in \Theta$ that minimizes the expected loss under the latent subpopulation Q_0

$$\underset{\theta \in \Theta}{\text{minimize}} \, \mathbb{E}_{X \sim Q_0}[\mathbb{E}[\ell(\theta; (X, Y)) \mid X]]. \tag{1}$$

We call this loss minimization under *mixture covariate shifts*.

Since the latent mixture weight and components are unknown, it is impossible to compute the loss (1) from observed data. Thus we postulate a lower bound $\alpha_0 \in (0, \frac{1}{2})$ on the subpopulation proportion α and consider the set of potential minority subpopulations

$$\mathcal{P}_{\alpha_0, X} := \{Q_0 : P_X = \alpha Q_0 + (1 - \alpha) Q_1 \;\; \text{for some } \alpha \geq \alpha_0 \text{ and distribution } Q_1 \text{ on } \mathcal{X}\}.$$

Concretely, our goal is to minimize worst-case subpopulation risk $\mathcal{R}$,

$$\underset{\theta \in \Theta}{\text{minimize}} \left\{ \mathcal{R}(\theta) := \sup_{Q_0 \in \mathcal{P}_{\alpha_0, X}} \mathbb{E}_{X \sim Q_0}[\mathbb{E}[\ell(\theta; (X, Y)) \mid X]] \right\}. \tag{2}$$

The worst-case formulation (2) is a distributionally robust optimization problem [9, 62] where we consider the worst-case loss over mixture covariate shifts, and we term the methodology we develop around this formulation *marginal* distributionally robust optimization (DRO), as we seek robustness only to shifts in the marginals over the covariates X. For datasets with heterogeneous subpopulations (e.g. NLP corpora), the worst-case subpopulation corresponds to a group that is "hard" under the current model θ. This approach—as we discuss further in the sequel—has connections with numerous desiderata and fields, including covariate shift problems, distributional robustness, fairness, and causal inference.

In some instances, the worst-case subpopulation (2) may be too conservative; the distribution of X may shift only on some components, or we may only care to achieve uniform performance across one variable. As an example, popular computer-vision datasets draw images mostly from western Europe and the United States [61], but one may wish for models that perform uniformly well over different geographic locations. In such cases, when one wishes to consider distributional shifts only on a subset of variables X_1 (e.g. geographic location) of the covariate vector $X = (X_1, X_2)$, we may simply redefine X as X_1, and Y as (X_2, Y) in the problem (2). All of our subsequent discussion generalizes in a straightforward way to such scenarios.

On the other hand, because of confounding, the assumption that the conditional distribution $P_{Y|X}$ does not change across groups may be too optimistic. While for machine learning tasks where human annotators use X to generate the label Y the assumption is appropriate, many problems include *unmeasured* confounding variables that affect the label Y and vary across subpopulations. For example, in a recidivism prediction task, the feature X may be the type of crime, the label Y is represents re-offending, and the subgroup may be race; without measuring

[1] We assume that the regular conditional probability distribution $P_{Y|X}$ exists without mention.

additional variables, such as income or location, $P_{Y|X}$ is likely to differ between groups. To address this issue, in Section 5 we generalize our proposed worst-case loss (2) to incorporate worst-case confounding shifts via a robust optimization formulation, providing finite-sample upper bounds on worst-case loss whose tightness depends on the effect of the unmeasured confounders on the conditional risk $\mathbb{E}[\ell(\theta; (X, Y)) \mid X]$.

1.1 Overview of results

In the rest of the paper, we construct a tractable finite sample approximation to the worst-case problem (2), and show that it allows learning models $\theta \in \Theta$ that do *uniformly well* over subpopulations. Our starting point is the duality result (see Section 2.1)

$$\mathcal{R}(\theta) := \sup_{Q_0 \in \mathcal{P}_{\alpha_0, X}} \mathop{\mathbb{E}}_{X \sim Q_0} [\mathbb{E}[\ell(\theta; (X, Y)) \mid X]] = \inf_{\eta} \left\{ \frac{1}{\alpha_0} \mathop{\mathbb{E}}_{X \sim P_X} [(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta)_+] + \eta \right\}.$$

For convex losses, the dual form yields a single convex loss minimization in the variables (θ, η) for minimizing $\mathcal{R}(\theta)$. When we (approximately) know the conditional risk $\mathbb{E}[\ell(\theta; (X, Y)) \mid X]$—for example, when we have access to replicate observations Y for each X—it is reasonably straightforward to develop estimators for the risk (2) (see Section 2.2).

Estimating the conditional risk via replication is infeasible in scenarios in which X corresponds to a unique individual (similar to issues in estimation of conditional treatment effects [20]). Our approach to this issue begins with the variational representation

$$\mathbb{E}[(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta)_+] = \sup_{h : \mathcal{X} \to [0,1]} \mathbb{E}_P[h(X)(\ell(\theta; (X, Y)) - \eta)]. \tag{3}$$

As the space $\{h : \mathcal{X} \to [0, 1]\}$ is too large to effectively estimate the quantity (3), we consider approximations via easier-to-control function spaces $\mathcal{H} \subset \{h : \mathcal{X} \to \mathbb{R}\}$ and study the convex minimization problem

$$\mathop{\text{minimize}}_{\theta \in \Theta, \eta} \left\{ \frac{1}{\alpha_0} \sup_{h \in \mathcal{H}} \mathbb{E}_P[h(X)(\ell(\theta; (X, Y)) - \eta)] + \eta \right\}. \tag{4}$$

By choosing $\mathcal{H}$ appropriately—e.g. as a reproducing kernel Hilbert space [11, 22] or a collection of bounded Hölder continuous functions—we can develop analytically and computationally tractable approaches to minimizing both (4) and (2).

If instead of simply considering sub-populations, we also allow shifts in the underlying sub-populations, we arrive at more "robust" choices than the problem (2). In particular, a combination of our results and those of Duchi and Namkoong [23] yields the following corollary. Recalling the Rényi divergence [69] of order q, defined by

$$D_q(P \| Q) := \frac{1}{q - 1} \log \int \left(\frac{dP}{dQ} \right)^q dQ,$$

then our Lemma 2.1 and [23, Lemma 1] show that defining

$$\mathcal{P}_{\Delta, X, q} := \{Q : D_q(Q \| P_X) \le \Delta\},$$

then for $1/p + 1/q = 1$ and $p \in (1, \infty)$ we have

$$\sup_{Q \in \mathcal{P}_{\Delta, X, q}} \mathop{\mathbb{E}}_{X \sim Q} [\mathbb{E}[\ell(\theta; (X, Y)) \mid X]] = \inf_{\eta} \left\{ \exp(\Delta/p) \mathop{\mathbb{E}}_{X \sim P_X} \left[(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta)_+^p \right]^{1/p} + \eta \right\}.$$

Abstracting away the particular choice of uncertainty in P_X, for $p \in [1, \infty]$, the infimum

$$\inf_{\eta \geq 0} \left\{ \frac{1}{\alpha_0} \left(\mathbb{E}_{X \sim P_X} \left[(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta)_+^p \right] \right)^{1/p} + \eta \right\} \tag{5}$$

always upper bounds the basic uniform mixture performance risk (2). As we show in Section 4, for Lipschitz conditional risks $x \mapsto \mathbb{E}[\ell(\theta; (X, Y)) \mid X = x]$, Eq. (5) is equal to a variant of the problem (4) where we take $\mathcal{H}$ to be a particular collection of Hölder-continuous functions allowing estimation from data. Because our robustness approach in this paper is new, there is limited analysis—either empirical or theoretical—of similar problems. Consequently, we perform some initial empirical evaluation on simulations to suggest the appropriate approximation spaces $\mathcal{H}$ in the dual form (4) (see Section 3), which informs our theoretical development and more detailed empirical evaluation to follow.

We develop an empirical surrogate to the risk (4) in Section 4. Our main theoretical result—Theorem 2, applied to the estimator (20)—shows that the model $\widehat{\theta}_n^{\text{rob}} \in \mathbb{R}^d$ minimizing this empirical surrogate achieves

$$\sup_{Q_0 \in \mathcal{P}_{\alpha_0, X}} \mathbb{E}_{X \sim Q_0}[\mathbb{E}[\ell(\widehat{\theta}_n^{\text{rob}}; (X, Y)) \mid X]] \leq R_p^* + O\left(n^{-\frac{p-1}{d+1}}\right),$$

with high probability, where

$$R_p^* := \inf_{\theta \in \Theta, \eta \geq 0} \left\{ \frac{1}{\alpha_0} \left(\mathbb{E}_{X \sim P_X} \left[(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta)_+^p \right] \right)^{1/p} + \eta \right\} \tag{6}$$

whenever $x \mapsto \mathbb{E}[\ell(\theta; (X, Y)) \mid X = x]$ is suitably smooth. In a rough sense, then, we expect that p trades between approximation (via R_p^*) and estimation error.

While our convergence guarantee gives a nonparametric rate $O(n^{-\frac{p-1}{d+1}})$, we empirically observe that our procedure achieves low worst-case losses even when the dimension d is large. We conjecture that this is due to the fact that our empirical approximation to the L^p norm bound (5) is an *upper bound* with error only $O(n^{-\frac{1}{4}})$, but a *lower bound* at the conservative rate $O(n^{-\frac{p-1}{d+1}})$. Such results—which we present at the end of Section 4—seem to point to the conservative nature of our convergence guarantee in practical scenarios. In our careful empirical evaluation on different semantic similarity assessment and recidivism prediction tasks (Section 6), we observe that our procedure learns models that do uniformly well across unseen subpopulations of difficult examples and minority groups.

1.2 Related work

Several important issues within statistics and machine learning are closely related to our goals of uniform performance across subpopulations. We touch on a few of these connections here, but only briefly, in the hope that further linking them may yield alternative approaches and deeper insights.

Covariate shifts A number of authors have studied the case where a target distribution of interest is different to the data-generating distribution—known as covariate shift or sample selection bias [63, 8, 66]. Much of the work focuses on the domain adaptation setting where the majority of observed data consists of samples $(X, Y) \sim P$ from a source domain. These methods require (often unsupervised) samples from an a priori *fixed* target domain, and apply

importance weight methods to reweight the observations when training a model for the target domain [65, 13, 29, 38]. When there are multiple domains, representation based methods can also identify sufficient statistics which are not affected by covariate shifts [35, 28].

On the other hand, our worst-case formulation assumes no knowledge of the latent group distribution Q_0 (unknown target) and controls performance on the worst subpopulation of size larger than α_0. Kernel-based adversarial losses [70, 45, 46] minimize the worst-case loss over the set of importance weighted distributions where the importance weights lie within a reproducing kernel Hilbert space. These methods are similar in that they consider a worst-case loss, but these worst-case weights provide no guarantees (even asymptotically) for latent subpopulations.

Distributionally robust optimization A large body of work on distributionally robust optimization (DRO) methods [10, 12, 42, 50, 24, 51, 26, 60, 14, 64, 43] solves a worst-case problem over the *joint distribution on* (X, Y). Our *marginal DRO* formulation (2) is a departure from previous DRO methods in that we study distributional shifts in the marginal covariate distribution $X \sim P_X$. Concretely, we can formulate an analogue [10, 51] of our marginal formulation (2) by letting

$$\mathcal{P}_{\alpha_0,(X,Y)} := \{Q_0 : P = \alpha Q_0 + (1-\alpha)Q_1 \text{ for some } \alpha \geq \alpha_0 \text{ and probability } Q_1 \text{ on } \mathcal{X} \times \mathcal{Y}\}.$$

The *joint DRO* objective

$$\sup_{Q_0 \in \mathcal{P}_{\alpha_0,(X,Y)}} \mathbb{E}_{(X,Y)\sim Q_0}[\ell(\theta; (X, Y))] \tag{7}$$

upper bounds the *marginal* worst-case formulation (2). The joint DRO objective (7) is frequently too conservative (see Section 2). By providing a tighter bound on the worst-case loss (2) under mixture covariate shifts, our finite-sample procedure (20) to come achieves better performance on unseen subpopulations (see Sections 4 and 6). For example, the joint DRO bound (7) applied to zero-one loss for classification may result in a degenerate non-robust estimator that upweights *all* misclassified examples [37], but our marginal DRO formulation mitigates these issues by using the underlying metric structure.

Fairness A growing literature recognizes the challenges of fairness within statistical learning [25, 31, 41, 40, 32], which motivates our approach as well. Among the many approaches to this problem, researchers have proposed that models with similar behavior across demographic subgroups are fair [25, 40]. The closest approach to our work is the use of Lipschitz constraints as a way to constrain the labels predicted by a model [25]. Rather than directly constraining the prediction space, we use the Lipschitz continuity of the conditional risk to derive upper bounds on model performance. The gap between joint DRO and marginal DRO relates to "gerrymandering" [40]: fair models can be unreasonably pessimistic by guaranteeing good performance against minority subpopulations with high *observed loss*—which can be a result of random noise—rather than high *expected loss* [25, 40, 34]. Our marginal DRO approach alleviates such gerrymandering behavior relative to the joint DRO formulation (7); see Section 4 for a more detailed discussion.

Causal inference Learning models that perform well under interventions is a common goal in causal inference, and one formulation of causality is as a type of invariance across environmental changes [53], and in this context, our formulation seeking models θ with low loss across marginal distributions on X is, in some sense, an analogue of observational studies

in causal inference. Bühlmann, Meinshausen, and colleagues have proposed a number of procedures similar in spirit to our marginal DRO formulation (2), though the key difference in their approaches is that they assume that underlying environmental changes or groups are *known*. Their maximin effect methods find linear models that perform well over heterogeneous data relative to a fixed baseline with known or constrained population structure [48, 57, 18], while anchor regression [58] fits regression models that perform well under small perturbations to feature values. Heinze-Deml and Meinshausen [35] consider worst-case covariate shifts, but assume a decomposition between causal and nuisance variables, with replicate observations sharing identical causal variables. Peters et al. [53] use heterogeneous environments to discover putative causal relationships in data, identifying robust models and suggesting causal links. Our work, in contrast, studies models that are robust to *mixture covariate shifts*, a new type of restricted interventions over all large enough subpopulations.

2 Estimating Performance Under Mixture Covariate Shift

We begin by reformulating the worst-case loss over mixture covariate shifts (2) via its dual (Section 2.1). We then consider a simpler setting in which we can collect replicate labels Y for individual feature vectors X—essentially, the analogue of a randomized study in causal inference problems—showing that in this case appropriate sample averages converge quickly to the worst-case loss (2) (Section 2.2). Although this procedure provides a natural gold standard when $x \mapsto \mathbb{E}[\ell(\theta; (x, Y)) \mid X = x]$ is estimable, it is impossible to implement when large sets of replicate labels are unavailable. This motivates empirical fitting procedure we propose in Eq. (20) to come, which builds out of the tractable upper bounds we present in Section 4.

2.1 Upper bounds for mixture covariate shift

Taking the dual of the inner maximization problem over covariate shifts (2), we have the following dual.

Lemma 2.1. *If* $\mathbb{E}[|\mathbb{E}[\ell(\theta; (X, Y)) \mid X]|] < \infty$, *then*

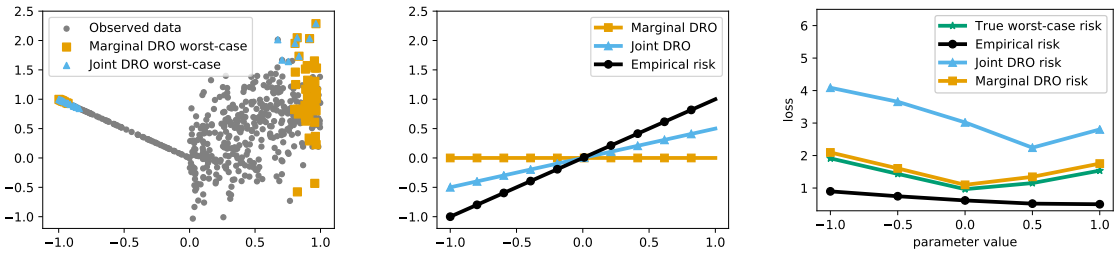
$$\sup_{Q_0(x) \in \mathcal{P}_{\alpha_0, X}} \mathop{\mathbb{E}}_{X \sim Q_0} [\mathbb{E}[\ell(\theta; (X, Y)) \mid X]] = \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha_0} \mathop{\mathbb{E}}_{X \sim P_X} \left[(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta)_+ \right] + \eta \right\}. \tag{8}$$

If additionally $0 \le \mathbb{E}[\ell(\theta; (X, Y)) \mid X] \le M$ *w.p. 1, the infimizing* η *lies in* $[0, M]$.

See Appendix A.1 for the proof. The dual form (8) is the conditional value at risk (CVaR) of the conditional risk $\mathbb{E}[\ell(\theta; (X, Y)) \mid X]$; the CVaR is a common measure of risk in the portfolio and robust optimization literatures [56, 62], but there it applies to an *unconditional* loss, making it (as we discuss in the sequel) conservative for the problems we consider.

The marginal DRO formulation (8) makes clear that the joint DRO problem (7) is conservative, as an identical calculation gives that Eq. (7) is equal to $\inf_\eta \{ \frac{1}{\alpha_0} \mathbb{E}[(\ell(\theta; (X, Y)) - \eta)_+] + \eta \}$, which is greater than the marginal DRO (8) unless Y is a function of X. In Section 4, we provide an approximation to the marginal DRO dual form (8), and one of our contributions is to show that our procedure has better theoretical and empirical performance than conservative estimators using the joint DRO objective (7).

To illustrate the advantages of the marignal distributionally robust approach, consider a misspecified linear regression problem, where we predict $\widehat{Y} = X\theta$ and use absolute the loss



(a) Data for the one-dimensional regression problem (circle) and the worst case distribution Q_0 for joint and marginal DRO (triangle / square).

(b) Best fit lines according to each loss. Only marginal DRO selects a line which fits both the $X > 0$ and $X < 0$ groups.

(c) Loss values for regression coefficients. Unlike Marginal DRO, ERM dramatically underestimates and joint DRO overestimates the worst case loss (2).

Figure 1: Toy problem of L_1 regression through origin.

$\ell(\theta; (x, y)) = |\theta x - y|$. The following mixture model generates the data:

$$Z \sim \text{Bernoulli}(0.15), \quad X = (1 - 2Z) \cdot \text{Uniform}([0, 1]), \\ Y = |X| + \mathbf{1}\{X \geq 0\} \cdot \varepsilon \quad \text{where } \varepsilon \sim \mathsf{N}(0, 1). \tag{9}$$

We plot observations from this model in Figure 1a, where 85% of the points are on the right, and have high noise. The model with the best uniform performance is near $\theta = 0$ which incurs similar losses between left and right groups. In contrast, the empirical risk minimizer ($\theta = 1$) incurs a high loss of 1 on the left group and $\sqrt{2/\pi}$ on the right one.

Empirical risk minimization tends to ignore the left group since it is a minority, resulting in high loss on the minority group $X < 0$ (Figure 1b). The joint DRO solution (7) minimizes losses over a worst-case distribution Q_0 consisting of examples which receive high loss (blue triangles), which tend to be samples on $X > 0$ due to noise. This results in a loose upper bound on the true worst-case risk as seen in Figure 1c. Our proposed estimator selects a worst-case distribution consisting of examples with high *conditional risk* $\mathbb{E}[\ell(\theta; (X, Y)) \mid X]$ (Figure 1a, orange squares). This worst-case distribution is not affected by the noise level, and results in a close approximation to the true loss (Figure 1c).

2.2 Estimation via replicates

A natural approach to estimating the dual form (8) is a two phase strategy, where we draw $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_X$ and then—for each X_i in the sample—draw a secondary sample of size m i.i.d. from the conditional $Y \mid X = X_i$; we then use these empirical samples to estimate $\mathbb{E}[\ell(\theta; (X, Y)) \mid X = X_i]$. While it is not always possible to collect replicate labels for a single X, *human annotated* data, which is common in machine learning applications [47, 5, 54], allows replicate measurement, where we may ask multiple annotators to label the same X.

Let us show how this procedure can yield explicit finite sample bounds with error at most $O(n^{-1/2} + m^{-1/2})$ for the population marginal robust risk (2) when the losses are bounded. We begin with the following

Assumption A1. *For some $M < \infty$, we have $\ell(\theta; (x, y)) \in [0, M]$ for all $\theta \in \Theta, x \in \mathcal{X}, y \in \mathcal{Y}$.*

The following estimate approximates the worst-case loss (2) for a fixed value of θ.

Proposition 1. *Let Assumption A1 hold. There exists a universal constant C such that for any fixed $\theta \in \Theta$, with probability at least $1 - \delta$*

$$\left| \sup_{Q_0 \in \mathcal{P}_{\alpha_0, X}} \mathop{\mathbb{E}}_{X \sim Q_0} [\mathbb{E}[\ell(\theta; (X, Y)) \mid X]] - \inf_{\eta \in [0, M]} \left\{ \frac{1}{\alpha_0 n} \sum_{i=1}^{n} \left(\frac{1}{m} \sum_{j=1}^{m} \ell(\theta; (X_i, Y_{i,j})) - \eta \right)_+ + \eta \right\} \right|$$
$$\le C \frac{M}{\alpha_0} \sqrt{\frac{1 + \log \frac{1}{\delta}}{\min\{m, n\}}}.$$

See Appendix A.2 for the proof. The estimator in Proposition 1 approximates the worst-case loss (2) well for large enough m and n. However—similar to the challenges of making causal inferences from observational data and estimating conditional treatment effects—it is frequently challenging or impossible to collect replicates for individual observations X, as each X represents an unrepeatable unique measurement. Consequently, the quantity in Proposition 1 is a type of gold standard, but achieving it is likely practically challenging.

3 Variational Approximation to Worst-Case Loss

The difficulty of collecting replicate data, coupled with the conservativeness of the joint DRO objective (7) for approximating the worst-case loss $\mathcal{R}(\theta)$, impel us to study tighter approximations that do not depend on replicates. Recalling the variational representation (3), we attempt to minimize

$$\mathcal{R}(\theta) = \frac{1}{\alpha_0} \sup_{h : \mathcal{X} \to [0,1]} \mathbb{E}_P[h(X)(\ell(\theta; (X, Y)) - \eta)] + \eta.$$

As we note in the introduction, this quantity is challenging to work with, so we restrict h to subsets $\mathcal{H}$ of $h : \mathcal{X} \to \mathbb{R}$, developing many possible approximations. The advantage of this formulation and its related relaxations (4) is that it replaces the dependence on the conditional risk with an expectation over the joint distribution on (X, Y), which we may estimate using the empirical distribution, as we describe in the next section. In this case, the lack of a "standard" choice motivates us to perform experiments to direct our development; we provide several example approximations in the next section, evaluating each empirically, and then perform a more careful theoretical analysis of the procedure providing the best empirical performance.

3.1 Example approximations and empirical variants

Each choice of a collection of functions $\mathcal{H} \subset \{h : \mathcal{X} \to [0, 1]\}$ to approximate the variational form (3) in the formulation (4) yields a new optimization problem. For each choice of $\mathcal{H}$ we propose below, we consider an empirical approximation $\widehat{\mathcal{H}}$, the subset of $\mathcal{H}$ restricted to mapping $\{X_1, \ldots, X_n\} \to \mathbb{R}$ instead of $\mathcal{X} \to \mathbb{R}$, solving the empirical alternative

$$\mathop{\text{minimize}}_{\theta \in \Theta, \eta} \left\{ \frac{1}{\alpha_0} \sup_{h \in \widehat{\mathcal{H}}} \mathbb{E}_{\widehat{P}_n}[h(X)(\ell(\theta; (X, Y)) - \eta)] + \eta \right\}. \tag{10}$$

We design our proposals so the dual of the inner supremum (10) is computable. When the conditional risk $x \mapsto \mathbb{E}[\ell(\theta; (X, Y)) \mid X = x]$ is smooth, we can provide generalization bounds for our procedures. We omit detailed development for our first two procedures—which we believe are natural proposals, justifying a bit of discussion—as in our empirical evaluations, neither is as effective as the last procedure, which controls the L^p *upper bound* (5) on $\mathcal{R}(\theta)$.

Reproducing Hilbert kernel spaces (RKHS) Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a reproducing kernel [11, 4] generating the reproducing kernel Hilbert space $\mathcal{H}_K$ with associated norm $\|\cdot\|_K$. For any $R \in \mathbb{R}_+$, we can define a norm ball

$$\mathcal{H}_{K,R} := \{h \in \mathcal{H}_K : \|h\|_K \le R, h \in [0,1]\}$$

and consider the variational approximation (4) with $\mathcal{H} = \mathcal{H}_{K,R}$. To approximate the population variational problem $\sup_{h \in \mathcal{H}_{K,R}} \mathbb{E}[h(X)(\ell(\theta;(X,Y)) - \eta)]$, we consider a restriction of the same kernel K to the sample space $\{X_1, \ldots, X_n\}$. Let $K_n = \{K(X_i, X_j)\}_{1 \le i,j \le n}$ be the Gram matrix evaluated on samples $X_1, \ldots, X_n$, and define the empirical approximation

$$\widehat{\mathcal{H}}_{K,R} := \left\{ h \in [0,1]^n : h = K_n \alpha \;\text{ for some } \alpha \in \mathbb{R}^n \text{ such that } \frac{1}{n^2} \alpha^\top K_n \alpha \le R \right\}.$$

(Recall that if $h(x) = \sum_{i=1}^n K(x, X_i)\alpha_i$, then $\|h\|_K^2 = \frac{1}{n^2}\alpha^\top K_n \alpha$.) To compute the empirical problem (10) with $\widehat{\mathcal{H}} = \widehat{\mathcal{H}}_{k,R}$, we take the dual of the inner supremum. Simplifying the dual form—whose derivation is a standard exercise in convex optimization [17]—we get

$$\underset{\theta \in \Theta, \eta \in \mathbb{R}, \beta \in \mathbb{R}^n}{\text{minimize}} \left\{ \frac{1}{\alpha_0 n} \sum_{i=1}^n \left(\ell(\theta;(X_i,Y_i)) - \eta + \beta_i\right)_+ + \frac{1}{n}\sqrt{R\beta^\top K_n \beta} \right\}.$$

For convex losses $\ell(\theta;(X,Y))$, this is a convex optimization problem in (θ, β, η),

Hölder continuous functions (bounded Hölder) Recall that a function $f : \mathcal{X} \to \mathbb{R}$ is (α, c)-Hölder continuous for $\alpha \in (0,1]$ and $c > 0$ if $|f(x) - f(x')| \le c\|x - x'\|^\alpha$ for all $x, x' \in \mathcal{X}$. Instead of the space of bounded functions, we restrict attention to Hölder continuous functions

$$\mathcal{H}_{c,\alpha} := \{h : \mathcal{X} \to [0,1] \mid h \text{ is } (\alpha, c)\text{-Hölder continuous}\}, \tag{11}$$

and consider the variational problem (4) with $\mathcal{H} = \mathcal{H}_{c,\alpha}$. The empirical plug-in of $\mathcal{H}_{c,\alpha}$ is

$$\widehat{\mathcal{H}}_{c,\alpha} := \{h : \{X_1, \ldots, X_n\} \to [0,1] \mid h \text{ is } (\alpha, c)\text{-Hölder continuous}\},$$

and we obtain an empirical plug-in procedure (10) with $\widehat{\mathcal{H}} = \widehat{\mathcal{H}}_{c,\alpha}$. Taking the dual of the inner supremum problem, we arrive at the following dual reformulation of the empirical variational problem

$$\underset{\theta \in \Theta, \eta, B \in \mathbb{R}_+^{n \times n}}{\text{minimize}} \left\{ \frac{1}{\alpha_0 n} \sum_{i=1}^n \left(\ell(\theta;(X_i,Y_i)) - \frac{1}{n}\sum_{j=1}^n (B_{ij} - B_{ji}) - \eta \right)_+ + \frac{c}{\epsilon n^2} \sum_{i,j=1}^n \|X_i - X_j\|^\alpha B_{ij} \right\}.$$

For convex losses $\theta \mapsto \ell(\theta;(X,Y))$, this is again a convex optimization problem in (θ, B, η), and is always smaller than the empirical joint DRO formulation (7).

L^p norm bounded Hölder functions (L^p Hölder) We consider the closely related function class consisting of L^p bounded Hölder functions. This class of functions is motivated through a L^p-norm bound (5) on the first term of the dual objective (8). For some $p \in (1,2]$ and $q = \frac{p}{p-1}$ we have

$$\begin{aligned}
\mathbb{E}_{X \sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+\right] &\le \left(\mathbb{E}_{X \sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+^p\right]\right)^{1/p} \\
&= \sup_h \left\{\mathbb{E}\left[h(X)(\ell(\theta;(X,Y)) - \eta)\right] \mid h : \mathcal{X} \to \mathbb{R}_+,\ \mathbb{E}[h(X)^q] \le 1\right\}.
\end{aligned} \tag{12}$$

If $x \mapsto \mathbb{E}[\ell(\theta; (X, Y)) \mid X = x]$ is Hölder continuous, then the function

$$h^\star(x) := \frac{(\mathbb{E}[\ell(\theta; (X, Y)) \mid X = x] - \eta)_+^{p-1}}{\left(\mathbb{E}_{X \sim P_X}\left[(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta)_+^p\right]\right)^{1/q}} \tag{13}$$

attaining the supremum in the variational form (12) is Hölder continuous with constant dependent on the magnitude of the denominator. In this case, carefully selecting the smoothness constant and L^p norm radius allows us to ensure that $\mathcal{H}$ contains $h^\star$ and derive upper bound guarantees for the resulting estimator (see Section 4 for more).

Minimizing the L^p upper bound rather than the original variational objective (alternatively, seeking higher-order robustness than the CVaR of the conditional risk $\mathbb{E}[\ell(\theta; (X, Y)) \mid X]$ as in our discussion of the quantity (5)) incurs approximation error. In practice, our experience is that this gap has limited effect, and the following lemma—whose proof we defer to Section A.3—quantifies the approximation error in inequality (12).

Lemma 3.1. *Let Assumption A1 hold and $Z(X) = \mathbb{E}[\ell(\theta; (X, Y)) \mid X]$. For $\eta \in [0, M]$*

$$\left(\mathbb{E}_{X \sim P_X}\left[(Z(X) - \eta)_+^p\right]\right)^{1/p} \le \min\Bigg\{(M - \eta)^{1/q}\left(\mathbb{E}\left(Z(X) - \eta\right)_+\right)^{1/p}, \\ \mathbb{E}\left(Z(X) - \eta\right)_+ + p^{1/p}(M - \eta)^{1/q}\left(\mathbb{E}\left|(Z(X) - \eta)_+ - \mathbb{E}[(Z(X) - \eta)_+]\right|\right)^{1/p}\Bigg\}.$$

Empirically (see the next section) we find this function class performs favorably to the other potential approximations; we consequently focus on it in the sequel.

3.2 Empirical Comparison of Variational Procedures

A priori it is unclear whether one class or another in our choices in Section 3.1 should yield better performance; at least in this point, our theoretical understanding provides similarly limited guidance. To that end, we perform a small simulation study to direct our coming deeper theoretical and empirical evaluation, discussing the benefits and drawbacks of various choices of $\mathcal{H}$ through the example we introduce in Figure 1a, Section 2.1.

We consider an elaborated version of the data mechanism (9) to incorporate higher dimensionality. Consider the data generating distribution

$$\begin{aligned} Z \sim \mathsf{Bern}(0.15), \quad X_1 = (1 - 2Z) \cdot \mathsf{Uni}([0, 1]), \quad X_2, \ldots, X_d \overset{\text{iid}}{\sim} \mathsf{Uni}([-1, 1]) \\ Y = |X_1| + \mathbf{1}\left\{X_1 \ge 0\right\} \cdot \varepsilon, \quad \varepsilon \sim \mathsf{N}(0, 1). \end{aligned} \tag{14}$$

Our goal is to predict Y via $\widehat{Y} = \theta^\top X$, and we use the absolute loss $\ell(\theta; (x, y)) = |y - \theta^\top x|$. We provide details of the experimental setup, such as the estimators and optimizers, in Section 6. In brief, we perform a grid search over all hyperparameters (Lipschitz estimates and kernel scales) for each method over 4 orders of magnitude; for the RKHS-based estimators, we test Gaussian, Laplacian, and Matern kernels, none of which have qualitative differences from one another, so we present results only for the Gaussian kernel $K(x, x') = \exp(-\frac{1}{2\sigma^2}\|x - x'\|_2^2)$.

Dimension dependence We first investigate the dimension dependence of the estimators, increasing d in the model (14) from $d = 2$ to 50 with a fixed sample size $N = 5000$.

Under model (14), we consider marginal distributionally robust objective (2) and evaluate the excess risk $\mathcal{R}(\theta) - \inf_\theta \mathcal{R}(\theta)$ for the choice $\alpha_0 = .15$, the hardest 15% of the data. As d grows, we expect estimation over Hölder continuous functions to become more difficult, and for the RKHS-based estimator to outperform the others. Figure 2 bears out this intuition (plotting the excess risk): high dimensionality induces less degradation in the RKHS approach than the others. Yet the absolute performance of the Hölder-based methods is better, which is unsurprising, as we are approximating a discontinuous indicator function.

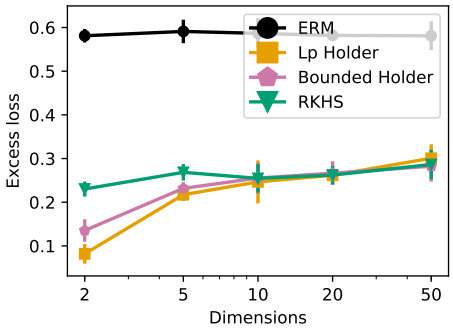


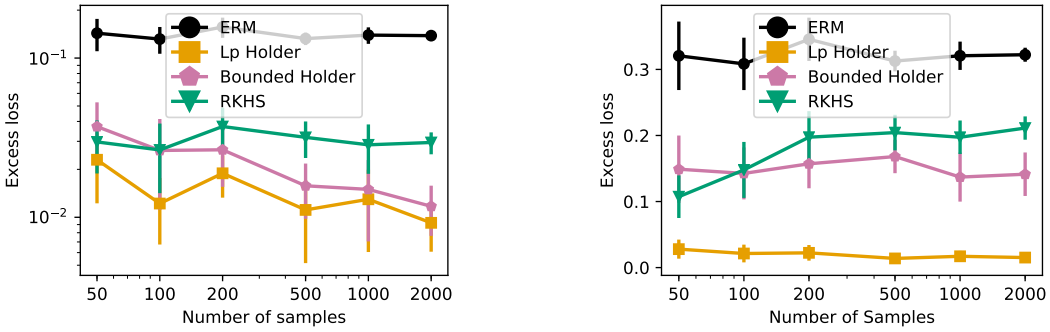
Figure 2. Variational estimates based on RKHS are less affected by the dimensionality of the problems, but perform worse than the Hölder continuous function approaches overall.

Sample size dependence We also consider the sample dependence of the estimators, fitting models using losses with robustness level set to $\alpha_0 = .15$, then evaluating their excess risk $\mathcal{R}(\widehat{\theta}) - \inf_\theta \mathcal{R}(\theta)$ using $\alpha_0 \in \{.05, .15\}$, so that we can see the effects of misspecification, as it is unlikely in practice that we know the precise minority population size against which to evaluate. Unlike the L^p-Hölder class (Eq. (12)), the bounded c-Hölder continuous function class (11) can approximate the population optimum of the original variational problem (3) as $n \to \infty$ and $c \to \infty$. Because of this, we expect that as the sample size grows, and c is set optimally, the bounded Hölder class will perform well.

On example (9) with $d = 1$, we observe in Figure 3a that both Hölder continuous class estimators with constants set via a hold-out set perform well as n grows, achieving negligible excess error. In contrast, the RKHS approach incurs high loss even with large sample sizes. Although both Hölder class approaches perform similarly when the robustness level α_0 is set properly, we find that the L^p-Hölder class is *substantially* better when the test time robustness level changes. The L^p-Hölder based estimator is the only one which provides reasonable estimators when training with $\alpha_0 = 0.15$ and testing with $\alpha_0 = 0.05$ (Figure 3b). Motivated by these practical benefits, we study finite sample properties of the L^p-bound estimator in what follows.

4 Tractable Risk Bounds for L^p Variational Problem

In this section, we formally develop and analyze our marginal DRO estimator $\widehat{\theta}_n^{\text{rob}}$, which solves an empirical approximation of the upper bound (5). Our estimator smooths the losses over nearby observations and upweights observations with high smoothed risk. Motivated by the empirical successes in the previous section, we also provide a number of generalization guarantees for our procedure.



(a) $\alpha_0 = 0.15$ for both train and test. (b) $\alpha_0 = 0.15$ for training, $\alpha_0 = 0.05$ for testing.

Figure 3. Performance of the function classes in low dimensional high sample size settings with well-specified robustness level (left) and misspecified robustness level (right).

4.1 A restricted variational formulation

As the space of all measurable functions $h : \mathcal{X} \to \mathbb{R}_+$ in the variational form (12) is too large, we consider a restriction to Lipschitzian risks.

Assumption A2. *For all $\theta \in \Theta$, the mappings $(x, y) \mapsto \ell(\theta; (x, y))$ and $x \mapsto \mathbb{E}[\ell(\theta; (x, Y)) \mid X = x]$ are L-Lipschitz.*

Recall that a function $f : \mathcal{X} \to \mathbb{R}$ is (α, c)-Hölder continuous for $\alpha \in (0, 1]$ and $c > 0$ if $|f(x) - f(x')| \le c \|x - x'\|^\alpha$ for all $x, x' \in \mathcal{X}$. To ease notation let $\mathcal{H}_{L,p}$ denote the space of Hölder continuous functions

$$\mathcal{H}_{L,p} := \left\{h : \mathcal{X} \to \mathbb{R},\ (p-1, L^{p-1})\text{-Hölder continuous}\right\}. \tag{15}$$

As we show below, if Assumption A2 holds and the denominator in the expression (13) has lower bound $\epsilon > 0$, then $\epsilon h^\star \in \mathcal{H}_{L,p}$, and we can approximate the variational form (12) by solving an analogous problem over smooth functions. Otherwise, we can bound the L^p-norm (12) by ϵ^{q-1}, which is small for small values of ϵ. Formally, if we define a variational objective over smooth functions $\mathcal{H}_{L,p}$

$$R_{p,\epsilon,L}(\theta, \eta) := \sup_{h \in \mathcal{H}_{L,p}} \left\{ \mathbb{E}\left[\frac{h(X)}{\epsilon}(\ell(\theta; (X, Y)) - \eta)\right] \;\middle|\; h \ge 0,\ (\mathbb{E}[h(X)^q])^{1/q} \le \epsilon \right\}, \tag{16}$$

we have a tight approximation to the variational form (12):

Lemma 4.1. *Let Assumptions A1, A2 hold and let $p \in (1, 2]$. Then, for any $\theta \in \Theta$ and $\eta \in \mathbb{R}$,*

$$\left(\mathbb{E}_{X \sim P_X}\left[(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta)_+^p\right]\right)^{1/p} = \inf_{\epsilon \ge 0} \left\{R_{p,\epsilon,L}(\theta, \eta) \vee \epsilon^{q-1}\right\}$$

and for any $\epsilon > 0$,

$$\left(R_{p,\epsilon,L}(\theta, \eta) \vee \epsilon^{q-1}\right) - \epsilon^{q-1} \le \left(\mathbb{E}_{X \sim P_X}\left[(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta)_+^p\right]\right)^{1/p}.$$

See Section A.4 for the proof.

4.2 The empirical estimator

Since the variational approximation $R_{p,\epsilon,L}$ does not depend on the conditional risk $\mathbb{E}[\ell(\theta;(X,Y)) \mid X]$, its empirical plug-in is a natural finite-sample estimator. Defining

$$\widehat{\mathcal{H}}_{L,p} := \left\{ h \in \mathbb{R}^n : \; h(X_i) - h(X_j) \le L^{p-1} \left\| X_i - X_j \right\|^{p-1} \quad \text{for all} \quad i,j \in [n] \right\}, \tag{17}$$

we consider the estimator

$$\widehat{R}_{p,\epsilon,L}(\theta,\eta) := \sup_{h \in \widehat{\mathcal{H}}_{L,p}} \left\{ \mathbb{E}_{\widehat{P}_n}\left[\frac{h(X)}{\epsilon} (\ell(\theta;(X,Y)) - \eta) \right] \;\middle|\; h \ge 0, \left(\mathbb{E}_{\widehat{P}_n}[h^q(X)] \right)^{1/q} \le \epsilon \right\}. \tag{18}$$

The following lemma shows that the plug-in $\widehat{R}_{p,\epsilon,L}(\theta,\eta)$ can be computed by a convex program.

Lemma 4.2. *For a sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ define the empirical loss*

$$\begin{aligned} \widehat{R}_{p,\epsilon,L}(\theta,\eta,B) := & \left(\frac{p-1}{n} \sum_{i=1}^n \left(\ell(\theta;(X_i,Y_i)) - \frac{1}{n} \sum_{j=1}^n (B_{ij} - B_{ji}) - \eta \right)_+^p \right)^{1/p} \\ & + \frac{L^{p-1}}{\epsilon n^2} \sum_{i,j=1}^n \left\| X_i - X_j \right\|^{p-1} B_{ij}. \end{aligned} \tag{19}$$

Then $\widehat{R}_{p,\epsilon,L}(\theta,\eta) = \inf_{B \ge 0} \widehat{R}_{p,\epsilon,L}(\theta,\eta,B)$ for all $\epsilon > 0$.

See Appendix A.5 for a proof of this lemma. We can interpret the dual variables B_{ij} as a transport plan for transferring the loss from example i to j in exchange for a distance dependent cost. The objective $\widehat{R}_{p,\epsilon,L}(\theta,\eta,B)$ thus consists of any losses larger than η after smoothing with B.

Noting that $\widehat{R}_{p,\epsilon,L}(\theta,\eta,B)$ is jointly convex in (η,B)—and jointly convex in (θ,η,B) if the loss $\theta \mapsto \ell(\theta;(X,Y))$ is convex—we consider the empirical minimizer

$$\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon} \in \operatorname*{argmin}_{\theta \in \Theta} \inf_{\eta \in [0,M], B \in \mathbb{R}_+^{n \times n}} \left\{ \frac{1}{\alpha_0} \left(\widehat{R}_{p,\epsilon,L}(\theta,\eta,B) \vee \epsilon^{q-1} \right) + \eta \right\} \tag{20}$$

as an approximation to the worst-case mixture covariate shift problem (2). We note that $\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}$ interpolates between the marginal and joint DRO solution; as $L \to \infty$, $B \to 0$ in the infimum over $\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}$ and $\widehat{R}_{p,\epsilon,L}(\theta,\eta) \to (\frac{p-1}{n} \sum_{i=1}^n (\ell(\theta;(X_i,Y_i)) - \eta)_+^p)^{1/p}$, an existing empirical approximation to the joint DRO problem [51].

4.3 Generalization and uniform convergence

We now turn to uniform convergence guarantees based on concentration of Wasserstein distances, which show that the empirical minimizer $\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}$ in expression (20) is an approximately optimal solution to the population bound (5). First, we prove that the empirical plug-in (18) converges to its population counterpart at the rate $O(n^{-\frac{p-1}{d+1}})$. For $\alpha \in (0,1]$, define the Wasserstein distance $W_\alpha(Q_1,Q_2)$ between two probability distributions Q_1, Q_2 on a metric space $\mathcal{Z}$ by

$$W_\alpha(Q_1,Q_2) := \sup \left\{ |\mathbb{E}_{Q_1}[h] - \mathbb{E}_{Q_2}[h]| \mid h : \mathcal{Z} \to \mathbb{R}, \; (\alpha,1)\text{-Hölder continuous} \right\}.$$

The following result—whose proof we defer to Appendix A.6—shows that the empirical plug-in (18) is at most $W_{p-1}(P,\widehat{P}_n)$-away from its population version.

Lemma 4.3. *Let Assumptions A1 and A2 hold. Assume that* $\operatorname{diam}(\mathcal{X}) + \operatorname{diam}(\mathcal{Y}) \le R$. *Then for* $p \in (1, 2]$ *and* $q = p/(p-1)$,

$$\sup_{\theta\in\Theta,\eta\in[0,M]} \left|\epsilon \vee \widehat{R}_{p,\epsilon,L}(\theta,\eta) - \epsilon \vee R_{p,\epsilon,L}(\theta,\eta)\right| \le B_\epsilon W_{p-1}(\widehat{P}_n, P)$$

where

$$B_\epsilon := LR + \epsilon^{-1}2^{q-1}L\left(2M + (q-1)LR\right) + \epsilon^{-q}2^{q-1}RML^2 + \epsilon^{q-2}(q-1)2^{q-2}L. \tag{21}$$

Our final bound follows from the fact that the Wasserstein distance between empirical and population distribution converges at rate $n^{-(p-1)/(d+1)}$. (See Appendix A.7 for proof.)

Theorem 2. *Let Assumptions A1, A2 hold. If* $p \in (1, 2]$, $d > 1$, *and* $\operatorname{diam}(\mathcal{X}) + \operatorname{diam}(\mathcal{Y}) \le R$, *then for any* $t > 0$, *with probability at least* $1 - c_1 \exp\left(-c_2 n(t^{\frac{d+1}{p-1}} \wedge t^2)\right)$

$$\begin{aligned}
&\sup_{Q_0(x)\in\mathcal{P}_{\alpha_0,X}} \mathbb{E}_{X\sim Q_0}[\mathbb{E}[\ell(\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}; (X,Y)) \mid X]] \\
&\le \inf_{\eta\in[0,M]} \left\{ \frac{1}{\alpha_0}\left(\mathbb{E}\left[\left(\mathbb{E}[\ell(\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}; (X,Y)) \mid X] - \eta\right)_+^p\right]\right)^{1/p} + \eta \right\} \\
&\le \inf_{\theta\in\Theta,\eta\in[0,M]} \left\{ \frac{1}{\alpha_0}\left(\mathbb{E}\left[(\mathbb{E}[\ell(\theta; (X,Y)) \mid X] - \eta)_+^p\right]\right)^{1/p} + \eta \right\} + \frac{\epsilon^{q-1}}{\alpha_0} + \frac{2B_\epsilon t}{\alpha_0}
\end{aligned} \tag{22}$$

where c_1 *and* c_2 *are positive constants that depend on* M, d, p.

Theorem 2 makes concrete how the power p trades off between approximation and estimation error in our concentration bounds for the worst-case loss (2) under mixture covariate shift. As $p \to 1$, the value R_p^* defined by expression (6) approaches the optimal value $\inf_{\theta\in\Theta} \sup_{Q_0\in\mathcal{P}_{\alpha_0,X}} \mathbb{E}_{X\sim Q_0}[\mathbb{E}[\ell(\theta; (X,Y)) \mid X]]$ so that approximation error goes down, but estimation becomes more difficult.

Upper bounds at faster rates Theorem 2 shows that the empirical estimator $\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}$ is approximately optimal with respect to the L^p-bound (5), but with a conservative $O(n^{-\frac{p-1}{d+1}})$-rate of convergence. On the other hand, we can still show that $\widehat{R}_{p,\epsilon,L}(\theta,\eta)$ provides an *upper bound* to the worst-case loss under mixture covariate shifts (2) at the faster rate $O(n^{-\frac{1}{4}})$. This provides a conservative estimate on the performance under the worst-case subpopulation; see Section A.8 for the proof of the following result.

Proposition 3. *Let Assumptions A1 and A2 hold. There exist numerical constants* $c_1, c_2 < \infty$ *such that the following holds. Let* $\theta \in \Theta$, $\epsilon > 0$, *and* $p \in (1, 2]$. *Then with probability at least* $1 - 2\gamma$, *uniformly over* $\eta \in [0, M]$

$$\mathbb{E}[(\mathbb{E}[\ell(\theta; (X,Y)) \mid X] - \eta)_+^p]^{1/p} \le \max\left\{\epsilon^{q-1}, (1+\tau_n)^{q-1}\widehat{R}_{p,\epsilon,L_n(\gamma)}(\theta,\eta) + \frac{c_1 M^2}{\epsilon^{q-1}}\sqrt{\frac{1}{n}\log\frac{1}{\gamma}}\right\}$$

where $\tau_n := c_2 M^2 \epsilon^{-q}\sqrt{\frac{1}{n}\log\frac{1}{\gamma}}$ *and* $L_n(\gamma) := L(1 + \tau_n(\gamma, \epsilon))^{-1/q}$.

We can similarly prove a uniform variant of Proposition 3 with respect to $\theta \in \Theta$, but we omit the details for brevity.

5 Risk bounds under confounding

Our assumption that $P_{Y|X}$ is fixed for each of our marginal populations over X is analagous to the frequent assumptions in causal inference that there are no unmeasured confounders [39]. To the extent that this is true—for example, in machine learning tasks where the label Y is a human annotation of the covariate X—minimizing worst-case losses over covariate shifts is natural, but the assumption may fail in real-world problems. For example, in predicting crime recidivism Y based on the type X of crime committed and race Z of the individual, unobserved confounders C (e.g. income, location, education) likely vary with race. Consequently, we provide a parallel to our earlier development that provides a sensitivity analysis to unmeasured (hidden) confounding.

Let us formalize. Let $C \in \mathcal{C}$ be a random variable, and in analogy to (2) we define

$$\mathcal{P}_{\alpha_0,(X,C)} := \left\{Q_0 : \exists\ \alpha \geq \alpha_0,\ \text{measure } Q_1 \text{ on } (\mathcal{X} \times \mathcal{C}) \text{ s.t. } P_{(X,C)} = \alpha Q_0 + (1-\alpha)Q_1\right\}. \quad (23)$$

Our goal is then to minimize the worst-case loss under mixture covariate shifts

$$\operatorname*{minimize}_{\theta \in \Theta} \sup_{Q_0 \in \mathcal{P}_{\alpha_0,(X,C)}} \mathop{\mathbb{E}}_{(X,C)\sim Q_0}[\mathbb{E}[\ell(\theta;(X,Y)) \mid X, C]]. \quad (24)$$

Since the confounding variable C is unobserved, we extend our robustness approach to control the degree of confounding—assuming a bounded effect—and derive conservative upper bounds on the worst-case loss.

We make the following boundedness definition and assumption on the effects of C.

Definition 1. *The triple (X, Y, C) is at most δ-*confounded for the loss ℓ *if*

$$\|\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \mathbb{E}[\ell(\theta;(X,Y)) \mid X, C]\|_{L^\infty(P)} \leq \delta.$$

Assumption A3. *The triple (X, Y, C) is at most δ-confounded for the loss ℓ.*

Paralleling our earlier development, we can apply a variational argument to bound the worst-case confounded risk (24).

Confounded variational problem Under confounding, a development completely parallel to Lemma 2.1 and Hölder's inequality yields the dual

$$\begin{aligned}
&\sup_{Q_0 \in \mathcal{P}_{\alpha_0,(X,C)}} \mathop{\mathbb{E}}_{(X,C)\sim Q_0}[\mathbb{E}[\ell(\theta;(X,Y)) \mid X, C]] \\
&\leq \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha_0} \left(\mathbb{E}_{(X,C)\sim P_{(X,C)}} \left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X, C] - \eta)_+^p \right] \right)^{\frac{1}{p}} + \eta \right\}
\end{aligned}$$

for all $p \geq 1$. Taking the variational form of the L^p-norm for $p \in (1, 2]$ yields

$$\begin{aligned}
&\left(\mathbb{E}_{P_{X,C}} \left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X, C] - \eta)_+^p \right]\right)^{1/p} \\
&= \sup_h \left\{ \mathbb{E}\left[h(X,C)(\mathbb{E}[\ell(\theta;(X,Y)) \mid X, C] - \eta)\right] \;\middle|\; h \geq 0, \mathbb{E}[h^q(X,C)] \leq 1 \right\}.
\end{aligned} \quad (25)$$

Instead of the somewhat challenging variational problem over h, we we reparameterize our problem in as $h(X) + f(X, C)$, where h is smooth and f is a bounded residual term, which—by taking a worst case over bounded f—allows us to provide an upper bound on the worst-case

problem (23). Let $\mathcal{H}_{L,p}$ be the space of Hölder functions (15) and $\mathcal{F}_{\delta,p}$ be the space of bounded functions

$$\mathcal{F}_{\delta,p} := \left\{ f : \mathcal{X} \times \mathcal{C} \to \mathbb{R} \text{ measurable}, \;\; \|f(X,C)\|_{L^\infty(P)} \le \delta^{p-1} \right\}.$$

Then defininining the analogue of the unconfounded variational objective (16)

$$R_{p,\epsilon,L,\delta}(\theta,\eta) := \tag{26}$$
$$\sup_{h+f\ge 0} \left\{ \mathbb{E}\left[\frac{h(X)+f(X,C)}{\epsilon} (\ell(\theta;(X,Y)) - \eta) \right] \mid h \in \mathcal{H}_{L,p}, f \in \mathcal{F}_{\delta,p}, \|h+f\|_{L^q(P)} \le \epsilon \right\},$$

the risk $R_{p,\epsilon,L,\delta}$ is ϵ-close to the variational objective (25):

Lemma 5.1. *Let Assumptions A1, A2, and A3 hold. Then, for any $\theta \in \Theta$ and $\eta \in \mathbb{R}$, we have*

$$\left(\mathbb{E}_{P_{X,C}}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C] - \eta)_+^p\right]\right)^{1/p} = \inf_{\epsilon\ge 0} \left\{ R_{p,\epsilon,L,\delta}(\theta,\eta) \vee \epsilon^{q-1} \right\} \tag{27}$$

and for any $\epsilon > 0$,

$$(R_{p,\epsilon,L,\delta}(\theta,\eta) \vee \epsilon^{q-1}) - \epsilon^{q-1} \le \left(\mathbb{E}_{P_{X,C}}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C] - \eta)_+^p\right]\right)^{1/p}.$$

See Appendix A.9 for proof.

Confounded estimator By replacing $\mathcal{H}_{L,p}$ with the empirical version $\widehat{\mathcal{H}}_{L,p}$ (the set of Hölder functions on the empirical distribution) and $\mathcal{F}_{\delta,p}$ with the empirical counterpart $\mathcal{F}_{\delta,p,n} := \{f \in \mathbb{R}^n \mid \max_{i\le n} |f(X_i,C_i)| \le \delta^{p-1}\}$, we may take the obvious empirical plug-in $\widehat{R}_{p,\epsilon,L,\delta}(\theta,\eta)$ of the population quantity (26). In this case, a duality argument provides the following analogue of Lemma 4.2, which follows because the class $\mathcal{F}_{\delta,p,n}$ simply corresponds to an $\|\cdot\|_\infty$ constraint on a vector in $\mathbb{R}^n$.

Lemma 5.2. *For any $\epsilon > 0$ and $(X_1,Y_1),\ldots,(X_n,Y_n)$, we have*

$$\widehat{R}_{p,\epsilon,L,\delta}(\theta,\eta) = \inf_{B\in\mathbb{R}_+^{n\times n}} \left\{ \left(\frac{p-1}{n} \sum_{i=1}^n \left(\ell(\theta;(X_i,Y_i)) - \frac{1}{n}\sum_{j=1}^n (B_{ij}-B_{ji}) - \eta \right)_+^p \right)^{1/p} \right.$$
$$\left. + \frac{L^{p-1}}{\epsilon n^2} \sum_{i,j=1}^n \|X_i - X_j\| B_{ij} + \frac{2\delta^{p-1}}{\epsilon n^2} \sum_{i,j=1}^n |B_{ij}| \right\}.$$

See Appendix A.11 for the proof. The lemma is satisfying in that it smoothly interpolates, based on the degree of confounding δ, between marginal distributionally robust optimization (when $\delta = 0$, as in Lemma 4.2) and the fully robust joint DRO setting as $\delta \uparrow \infty$, which results in the choice $B = 0$.

Upper bound on confounded objective In analogy with Proposition 3, the empirical plug-in $\widehat{R}_{p,\epsilon,L,\delta}(\theta,\eta)$ is an upper bound on the population objective under confounding. Although our estimator only provides an upper bound, it provides practical procedures for controlling the worst-case loss (23) when Assumption A3 holds, as we observe in the next section. The next proposition, whose proof we provide in Section A.10, shows the upper bound.

Proposition 4. *Let Assumptions A1, A2, and A3 hold. There exist universal constants $c_1, c_2 < \infty$ such that the following holds. Let $\theta \in \Theta$, $\epsilon > 0$, and $p \in (1, 2]$. Then with probability at least $1 - 2\gamma$, uniformly in $\eta \in [0, M]$*

$$\left(\mathbb{E}_{P_{X,C}}\left(\mathbb{E}[\ell(\theta; (X, Y)) \mid X, C] - \eta\right)_+^p\right)^{1/p}$$
$$\le \max\left\{\epsilon^{q-1},\ (1 + \tau_n(\gamma, \epsilon))^{1/q}\ \widehat{R}_{p,\epsilon,L_n(\gamma),\delta_n(\gamma)}(\theta, \eta) + \frac{c_1 M^2}{\epsilon^{q-1}}\sqrt{\frac{\log\frac{1}{\gamma}}{n}}\right\}$$

where $\tau_n(\gamma, \epsilon) := \frac{c_2 M^2}{\epsilon^{q-1}}\sqrt{\frac{1}{n}\log\frac{1}{\gamma}}$, $\delta_n(\gamma) := \delta(1 + \tau_n(\gamma, \epsilon))^{-1/q}$, *and* $L_n(\gamma) := L(1 + \tau_n(\gamma, \epsilon))^{-1/q}$.

6 Experiments

To complete the paper, we provide an empirical investigation of the procedure (20), focusing on two main aspects of our results. First, our theoretical results exhibit nonparametric rates of convergence, so it is important to understand whether these penalties are real and cause degradation in empirical performance and the extent to which the procedure is effective. Second, by virtue of focusing on examples with high conditional risk $\mathbb{E}[\ell(\theta; (X, Y)) \mid X]$, we expect our procedure should improve performance on minority groups and hard subpopulations.

To investigate both of these issues, we begin by studying simulated data (Section 6.1) so that we can more precisely evaluate true convergence, and we see that in moderately high dimensions, our procedure outperforms both empirical risk minimization (ERM) and joint DRO on worst-off subpopulations; we perform a parallel simulation study in Section 6.2 for the confounded case (minimizing $\widehat{R}_{p,\epsilon,L,\delta}(\theta, \eta)$ of Lemma 5.2). After our simulation study, we continue to assess efficacy of our procedure on real data, using our method to predict semantic similarity (Sections 6.3) and crime recidivism (Section 6.4). In both of these experiments, our results are consistent with our expectation that our procedure (20) typically achieves better performance over unseen subpopulations than other methods.

Important in each of our procedures is the choice of hyperparameters. We must choose a Lipschitz constant L, mixture size α, risk level ϵ, and moment parameter p. In our experiments, we see that cross-validation is attractive and effective. We treat the value L/ϵ as a single hyperparameter to estimate via a hold-out set or, in effort to demonstrate sensitivity to the parameter, plot results across a range of L/ϵ. As the objective (20) is convex standard methods apply; we use (sub)gradient descent to optimize the problem parameters over θ, η, B. In each experiment, we compare our marginal DRO method against two baselines: empirical risk minimization (ERM) and joint DRO (7). ERM minimizes the empirical risk $\text{minimize}_{\theta\in\Theta}\, \mathbb{E}_{\widehat{P}_n}[\ell(\theta; (X, Y))]$, and provides very weak guarantees on subpopulation performance. Joint DRO (7) is the only existing method that provides an upper bound to the worst-case risk; we evaluate empirical plug-ins of the L_p dual of joint DRO,

$$\inf_{\eta\in\mathbb{R}}\left\{\frac{1}{\alpha_0}\mathbb{E}[(\ell(\theta; (X, Y)) - \eta)_+^p]^{1/p} + \eta\right\}.$$

6.1 Simulation study: the unconfounded case

Our first simulation study focuses on the unconfounded procedure (20), where the known ground truth allows us to carefully measure the effects of the problem parameters (n, d, α) and sensitivity to the smoothness assumption L. We focus on the regression example from Sec. 3.2.

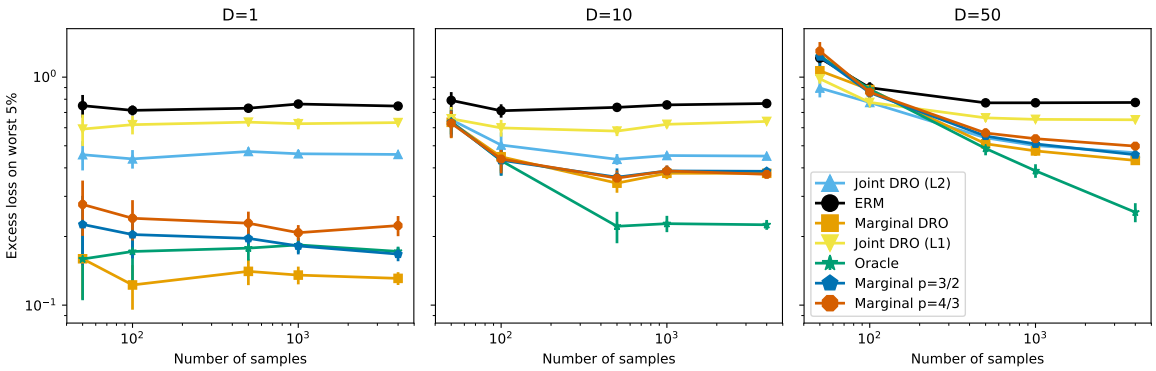


Figure 4: Dimension and sample size dependence of robust loss surrogates

Because the uniform distribution exhibits the slowest convergence of empirical distributions for Wasserstein distance [27]—and as Wasserstein convergence underpins our $n^{-1/d}$ rates in Lemma 4.3—we use the uniform distribution over covariates X.

The simulation distribution (14) captures several aspects of loss minimization in the presence of heterogeneous subpopulations. The two subpopulations ($X_1 \geq 0$ and $X_1 < 0$) constitute a majority and minority group, and minimizing the risk of the majority group comes at the expense of risk for the minority group. The two subpopulations also define an oracle model, obtained by minimizing the maximum loss over the two groups.

Our procedure (20) is an empirical approximation to the problem

$$\inf_{\theta} \sup_{Q_0 \in \mathcal{P}_{\alpha_0, X}} \mathop{\mathbb{E}}_{X \sim Q_0} \left[\mathbb{E}[|\theta^\top X - Y| \mid X] \right].$$

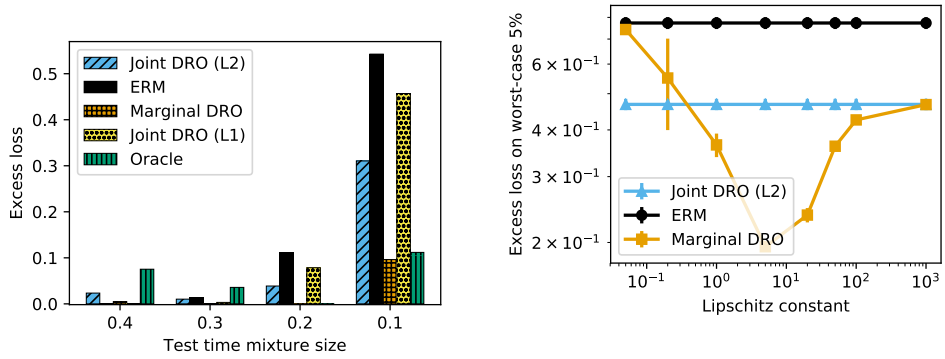
We choose L/ϵ by cross-validation on a held-out set with 1000 examples and 100 replicates (i.e. repeated measurements of Y).

we do not use any additional regularization such as an L_2 penalty since we are in the $d \ll n$ regime.

Effect of the p-norm bound We evaluate the difference in model quality as a function of p, which controls the tightness of the p-norm upper bound. Our convergence guarantees in Theorem 2 are looser for p near 1, though such values achieve smaller asymptotic bias to the true sub-population risk, while values of p near 2 suggest a more favorable sample size dependence in the theorem.

In Figure 4, we plot the results of experiments for each suggested procedure, where the horizontal axis of each plot indexes sample size and the vertical axis the excess loss over the population minimizer on the worst 5% of the population; each plot corresponds to a different dimension $d \in \{1, 10, 50\}$. The plots suggest that the choice of $p = 2.0$ (Marginal DRO) outperforms other choices of p, and performance generally seems to degrade as $p \downarrow 1$ with this trend holding across all dimensions tested. Conveniently, when $p = 2$ the optimizationn problem (20) is a second-order cone problem (SOCP), which standard mathematical programming solvers support [17]. We consequently focus on the $p = 2$ case for the remainder of this section.

Sample size and dimension dependence In our second simulation study, we examine pessimistic $O(n^{-1/d})$ convergence rate of our estimator (20) results in substantially worse



(a) Loss for various worst-case group sizes. (b) Marginal DRO losses across L/ϵ

Figure 5. Sensitivity of marginal DRO losses to the size of the worst-case group (left) and Lipschitz constant estimate (right).

estimates compared to ERM and joint DRO (7), both of which have convergence rates scaling at worst as $1/\sqrt{n}$ [23]. We fix $\alpha = .2$ (sub-population size 20%). In low dimensions ($d = 1$ to $d = 10$) convergence to the true, optimal function value—which we can compute exactly—is relatively fast, and marginal DRO becomes substantially better with as few as 500 samples (Figure 4). In higher dimensions ($d = 50$), marginal DRO convergence is slower, but it is only worse than the joint DRO solution when $n = d = 100$. At large sample sizes $n > 1000$, we find that marginal DRO begins to strictly outperform the two baselines. (Again the vertical axis of the plots displays the excess loss over the population minimizer on the worst 5% of the population.)

Sensitivity to robustness level We rarely know the precise minority proportion α, so that in practice one usually provides a postulated lower bound α_0; we thus investigate sensitivty to its specification. We fix the true proportion $\alpha = 0.3$, and show results of varying the test-time mixture size in Figure 5a. Marginal DRO obtains a loss within 1.2 times the oracle model regardless of the test-time mixture size, while both ERM and joint DRO incur substantially higher losses on the tails.

Sensitivity to Lipschitz constant Finally, the empirical bound (22) requires an estimate of the Lipschitz constant of the conditional risk. We vary the estimate L/ϵ in Figure 5a, showing that the marginal robustness formulation has some sensitivity to the parameter, though there is a range of several orders of magnitude through which it outperforms the joint DRO procedures. The behavior that it exhibits is expected, however: the choice $L = 0$ reduces the marginal DRO procedure to ERM in the bound (22), while the choice $L = \infty$ results in the joint DRO approach.

6.2 Simulation study: the confounded case

To complement our results in the unconfounded case, we now extend our simulation experiment by adding unmeasured confounders, investigating the risk upper bounds of Lemma 5.2 and Proposition 4. We use a generative model nearly identical to the unconfounded model (14),

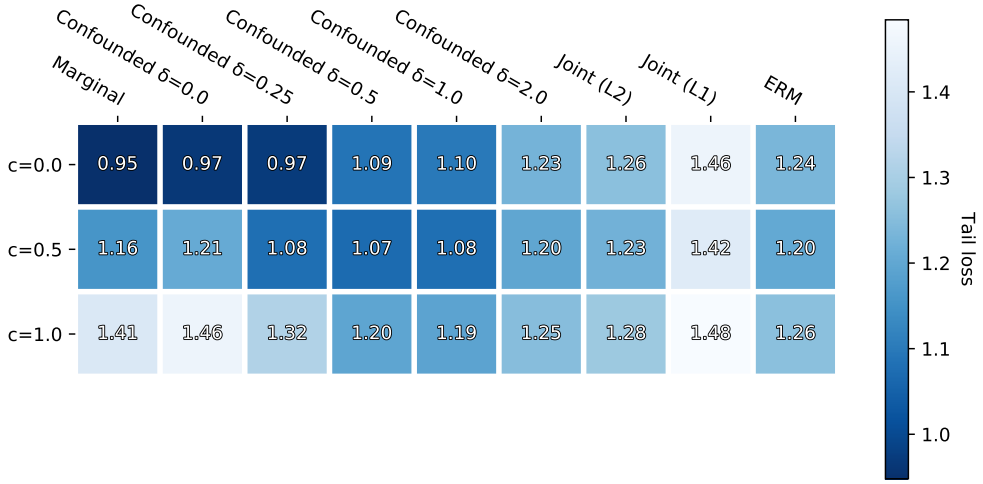


Figure 6. Marginal DRO bounds under confounding. The confounded risk bound extension in Lemma 5.2 controls the risk under bounded confounding.

introducing a confounder C:

$$Z \sim \mathsf{Bern}(\alpha), \quad X_1 = (1-2Z) \cdot \mathsf{Uni}([0,1]), \quad X_2, \ldots, X_d \overset{\text{iid}}{\sim} \mathsf{Uni}([0,1])$$
$$Y = |X_1| + \mathbf{1}\{X_1 \geq 0\} \cdot C, \quad C \sim \mathsf{Uni}(\{-1, 0.5, 0, 0.5, 1\}).$$

We evaluate our model on the worst-case confounded loss for a *fixed* c,

$$\inf_{\theta} \sup_{Q_0 \in \mathcal{P}_{\alpha_0,(X)}} \mathop{\mathbb{E}}_{X \sim Q_0} \mathbb{E}\left[|\theta^\top X - Y| \mid X, C = c\right].$$

The choice of c is precisely the bound on the shift in $\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \mathbb{E}[\ell(\theta;(X,Y)) \mid X, C]$ in Definition 1 (corresponding to δ in the definition). Lemma 5.2 provides a bound on the worst case loss under confounding depending on the amount of shift, and we compare this to the expected worst-case loss (25) at test time. When we set the postulated level δ of confounding in $\widehat{R}_{p,\epsilon,L,\delta}(\theta,\eta)$ (Lemma 5.2) to be equal to $|c|$—so that it is exactly correct—the procedure that minimizes $\widehat{R}_{p,\epsilon,L,\delta}(\theta,\eta)$ achieves lower losses compared to other baselines, including marginal DRO (no confounding), L^2 joint DRO (fully counfounded with $\delta = \infty$), and empirical risk minimization.

6.3 Semantic similarity prediction

We now present the first of our real-world evaluations of marginal distributionally robust optimization, studying the simpler case in which we have multiple measured outcome labels Y for each covariate X so that it is possible to accurately estimate the true worst-case losses of our model via Proposition 1. We consider a lexical semantic similarity prediction dataset (WS353 [2]) where the features are pairs of words, and labels are a set of 13 human annotations rating the word similarity on a $0-10$ scale. We represent each word pair as the difference $(x_1 - x_2)$ of the word vectors (derived from the representation [52]) x_1 and x_2 associated to each word. (We could allow more general models, but we wish only to measure similarity, and we wish to keep the dimension of the resulting models reasonably small.)

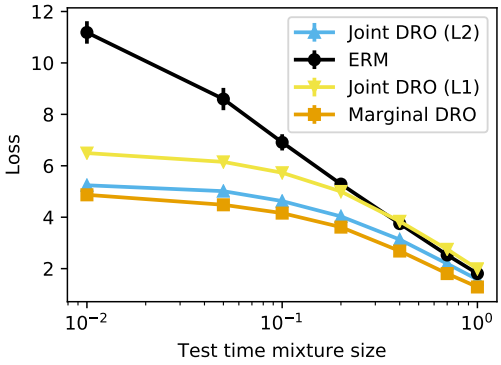


Figure 7. Semantic similarity prediction task, with prediction error over subgroups (y-axis) evaluated over varying worst case mixture group sizes (x-axis).

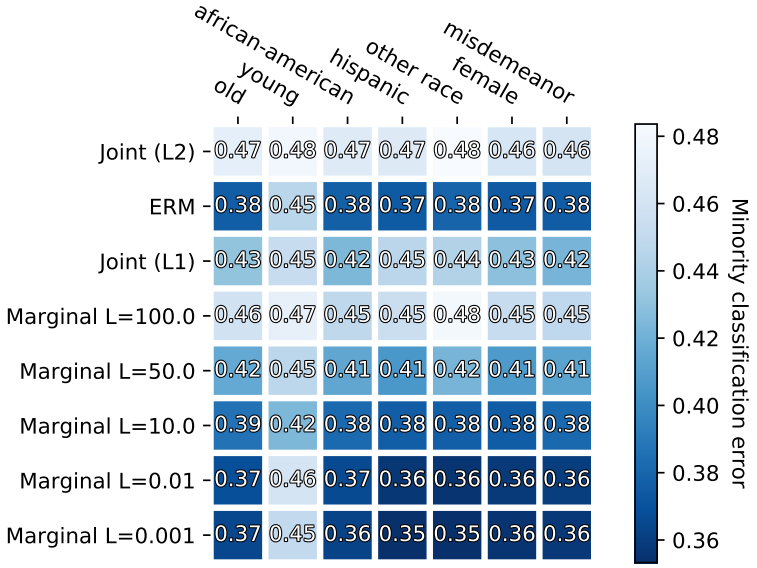


Figure 8. Recidivism prediction task as a covariate shift problem. Models trained on the full dataset are evaluated on subgroups.

We cast this as a regression task of predicting a scalar similarity Y with a word-pair vector X via the quadratic model $x \mapsto x^\top \Theta_1 x + \theta_2$, $\Theta_1 \in \mathbb{R}^{d \times d}$ and $\theta_2 \in \mathbb{R}$. For training, we minimize the absolute deviation loss over 1989 individual annotations, and evaluate the loss with respect to 246 averaged annotations. We set the minority proportion for both DRO procedures at $\alpha_0 = 0.3$, and tuned the Lipschitz constant via a held out set on the multi-annotator data; all methods used the same ridge regularization parameter tuned for the *ERM model*.

All methods achieve low average error over the entire dataset, but ERM, joint DRO and marginal DRO have very different behaviors for small subgroups. ERM incurs extremely large errors at $\sim 5\%$ of the test set, resulting in near random prediction. Applying the joint DRO estimator reduces error by nearly half and marginal DRO reduces this even further (Figure 7).

6.4 Recidivism prediction

Finally, we show that marginal DRO can control the loss over a minority group on a recidivism prediction task. The COMPAS recidivism dataset [21] is a classification task where examples are individual convicts, features consist of binary demographic labels (such as African American or not) and description of their crimes, and the label is whether they commit another crime (recidivism). We use the fairML toolkit version of this dataset [1]. Classification algorithms for recidivism have systematically discriminated against minority groups, and this dataset is a benchmark for testing such discrimination [6]. We consider this dataset from the perspective of achieving uniform performance across various groups. There are 10 binary variables in data, each indicating a potential split of the data into minority and majority group (e.g. young vs. not young, or African-American vs. not African-American), of which 7 have enough ($n > 10$) observations in each split to make reasonable error estimates. We train a model over the full population (using all the features), and for each of the 7 frequent enough demographic indicator variables, we evaluate loss on a held-out test set suffered on the associated majority group and associate minority group. We then report the maximum of these two losses.

Our goal is to ensure that the classification accuracy remains high *without* explicitly splitting the data on particular demographic labels (though we include them in our models as they have predictive power). We treat this problem as a classification problem with the logistic loss, using a 70% v. 30% train-test split. We set $\alpha_0 = 0.4$ for all DRO methods, and apply L_2-regularization to all models where the regularization parameter was tuned for the *ERM model*. To evaluate sensitivity, we present the loss across a range of Lipschitz constants from 100 to 0.001.

In Figure 8, we present the loss on the seven frequent enough demographic splits. For each attribute on the horizontal axis, we evaluate classification error (on the test set) of the model when the attribute is true and false, reporting the worse value of the two. Unlike the regression tasks considered earlier, joint DRO (both L_2 and L_1) performs worse than ERM on almost all demographic splits. On the other hand, we find that the use of marginal DRO with the appropriate smoothness constant $L \in \{10^{-2}, 10^{-3}\}$ reduces classification errors between $1-2\%$ on the worst case group across various demographics, with the largest error reduction of 3% occurring in the young vs. old demographic split.

References

[1] J. A. Adebayo. Fairml : Toolbox for diagnosing bias in predictive modeling. Master's thesis, Massachusetts Institute of Technology, 2016.

[2] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2009.

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, and G. Chen. Deep speech 2: end-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 173–182, 2016.

[4] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, May 1950.

[5] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007. URL http://www.ics.uci.edu/~mlearn/MLRepository.html.

[6] S. Barocas and A. D. Selbst. Big data's disparate impact. *104 California Law Review*, 3: 671–732, 2016.

[7] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[8] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 20*, pages 137–144, 2007.

[9] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

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

[11] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[12] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *Mathematical Programming, Series A*, 167(2):235–292, 2018. URL http://arxiv.org/abs/1401.0212.

[13] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[14] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *arXiv:1604.01446 [math.PR]*, 2016. URL https://arxiv.org/abs/1604.01446.

[15] S. L. Blodgett, L. Green, and B. O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of Empirical Methods for Natural Language Processing*, pages 1119–1130, 2016.

[16] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[17] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[18] P. Bühlmann and N. Meinshausen. Magging: maximin aggregation for inhomogeneous large-scale data. *Proceedings of the IEEE*, 104(1):126–135, 2016.

[19] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2017)*, 2017.

[20] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[21] A. Chouldechova. A study of bias in recidivism prediciton instruments. *Big Data*, pages 153–163, 2017.

[22] N. Cristianini and J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[23] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv:1810.08750 [stat.ML]*, 2018.

[24] J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv:1610.03425 [stat.ML]*, 2016.

[25] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pages 214–226, 2012.

[26] P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming, Series A*, to appear, 2017.

[27] 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.

[28] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf. Domain adaptation with conditional transferable components. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 2839–2848, 2016.

[29] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. In J. Q. nonero Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors, *Dataset Shift in Machine Learning*, chapter 8, pages 131–160. MIT Press, 2009.

[30] P. J. Grother, G. W. Quinn, and P. J. Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST Interagency/Internal Reports (NISTIR)*, 7709, 2010.

[31] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 30*, 2016.

[32] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[33] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[34] Ú. Hébert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv:1711.08513 [cs.LG]*, 2017.

[35] C. Heinze-Deml and N. Meinshausen. Grouping-by-id: Guarding against adversarial domain shifts, 2017.

[36] D. Hovy and A. Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, volume 2, pages 483–488, 2015.

[37] W. Hu, G. Niu, I. Sato, and M. Sugiayma. Does distributionally robust supervised learning give robust classifiers? *arXiv:1611.02041v4 [stat.ML]*, 2018. URL https://arxiv.org/abs/1611.02041v4.

[38] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 20*, pages 601–608, 2007.

[39] G. Imbens and D. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[40] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv:1711.05144 [cs.LG]*, 2018.

[41] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems 31*, 2017.

[42] H. Lam and E. Zhou. Quantifying input uncertainty in stochastic optimization. In *Proceedings of the 2015 Winter Simulation Conference*. IEEE, 2015.

[43] J. Lee and M. Raginsky. Minimax statistical learning and domain adaptation with Wasserstein distances. *arXiv:1705.07815 [cs.LG]*, 2017.

[44] J. Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *arXiv:1804.10556 [math.ST]*, 2018.

[45] A. Liu and B. Ziebart. Robust classification under sample selection bias. In *Advances in Neural Information Processing Systems 28*, pages 37–45, 2014.

[46] A. Liu and B. Ziebart. Robust covariate shift prediction with general losses and feature views. *arXiv:1712.10043 [cs.LG]*, 2017. URL https://arxiv.org/abs/1712.10043.

[47] M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330, 1994.

[48] N. Meinshausen and P. Bühlmann. Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4):1801–1830, 2015.

[49] G. J. Minty. On the extension of lipschitz, lipschitz-hölder continuous, and monotone functions. *Bulletin of the American Mathematical Society*, 76(2):334–339, 1970.

[50] T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii. Distributional smoothing with virtual adversarial training. *arXiv:1507.00677 [stat.ML]*, 2015.

[51] H. Namkoong and J. C. Duchi. Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems 31*, 2017.

[52] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of Empirical Methods for Natural Language Processing*, 2014.

[53] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B*, 78(5):947–1012, 2016.

[54] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods for Natural Language Processing*, 2016.

[55] P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2018.

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

[57] D. Rothenhäusler, N. Meinshausen, and P. Bühlmann. Confidence intervals for maximin effects in inhomogeneous large-scale data. In *Statistical Analysis for High-Dimensional Data*, pages 255–277. Springer, 2016.

[58] D. Rothenhäusler, P. Bühlmann, N. Meinshausen, and J. Peters. Anchor regression: heterogeneous data meets causality. *arXiv:1801.06229 [stat.ME]*, 2018.

[59] P. Sapiezynski, V. Kassarnig, and C. Wilson. Academic performance prediction in a gender-imbalanced environment. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, volume 1, pages 48–51, 2017.

[60] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems 29*, pages 1576–1584, 2015.

[61] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv:1711.08536 [stat.ML]*, 2017.

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

[63] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[64] A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *Proceedings of the Sixth International Conference on Learning Representations*, 2018.

[65] A. J. Storkey and M. Sugiyama. Mixture regression for covariate shift. In *Advances in Neural Information Processing Systems 19*, pages 1337–1344, 2006.

[66] M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

[67] R. Tatman. Gender and dialect bias in YouTube's automatic captions. In *First Workshop on Ethics in Natural Langauge Processing*, volume 1, pages 53–59, 2017.

[68] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

[69] T. van Erven and P. Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[70] J. Wen, C.-N. Yu, and R. Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *Proceedings of the 31st International Conference on Machine Learning*, pages 631–639, 2014.

A Proofs

A.1 Proof of Lemma 2.1

We begin by deriving a likelihood ratio reformulation, where we use $W := \mathbb{E}[\ell(\theta; (X, Y)) \mid X]$ to ease notation

$$\sup_{Q_0 \in \mathcal{P}_{\alpha_0, X}} \mathbb{E}_{X \sim Q_0}[W] = \sup_{L} \left\{ \mathbb{E}_P[LW] \middle| L : \Omega \to [0, 1/\alpha_0], \text{ measurable, } \mathbb{E}_p[L] = 1 \right\}. \tag{28}$$

To see that inequality "$\leq$" holds, let $Q_0 \in \mathcal{P}_{\alpha_0, X}$ be a probability over $\mathcal{X}$. Note that Q_0 induces a distribution over $(\Omega, \sigma(X))$, which we denote by the same notation for simplicity. Since $P_X = \alpha_0 Q_0 + (1 - \alpha_0) Q_1$ for some probability Q_1 and $\alpha_0 \in (0, 1)$, we have $Q_0 \ll P_X$. Letting $L := \frac{dQ_0}{dP_X}$, it belongs to the constraint set in the right hand side and we conclude $\leq$ holds. To see the reverse inequality "$\geq$", for any likelihood ratio $L : \Omega \to [0, 1/\alpha_0]$, let $Q_0 := P_X L$ so that $Q_0(A) := \mathbb{E}_{P_X}[\mathbf{1}\{A\} L]$ for all $A \in \sigma(X)$. Noting $Q_1 := \frac{1}{1-\alpha_0} P_X - \frac{\alpha_0}{1-\alpha_0} Q_0$ defines a probability measure and $P_X = \alpha_0 Q_0 + (1 - \alpha_0) Q_1$, we conclude that inequality $\geq$ holds.

Next, the following lemma gives a variational form for conditional value-at-risk, which corresponds to the worst-case loss (2) under mixture covariate shifts.

Lemma A.1 ([62, Example 6.19])**.** *For any random variable $W : \mathcal{X} \to \mathbb{R}$ with $\mathbb{E}|W| < \infty$,*

$$\begin{aligned} &\sup_{L} \left\{ \mathbb{E}_P[LW] \middle| L : \Omega \to [0, \frac{1}{\alpha_0}], \text{ measurable, } \mathbb{E}_p[L] = 1 \right\} \\ &= \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha_0} \mathbb{E}_{X \sim P_X} \left[(W - \eta)_+ \right] + \eta \right\}. \end{aligned}$$

From the reformulation (28) and Lemma A.1, we obtain the first result.

We now show that when $W \in [0, M]$,

$$\inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha_0} \mathbb{E}[(W - \eta)_+] + \eta \right\} = \inf_{\eta \in [0, M]} \left\{ \frac{1}{\alpha_0} \mathbb{E}[(W - \eta)_+] + \eta \right\}. \tag{29}$$

Noting that $\eta \mapsto g(\eta) := \frac{1}{\alpha_0} \mathbb{E}[(W - \eta)_+] + \eta$ is strictly increasing on $[M, \infty)$ since $g(\eta) = \eta$ for $\eta \in [M, \infty)$, we have

$$\inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha_0} \mathbb{E}[(W - \eta)_+] + \eta \right\} = \inf_{\eta \leq M} \left\{ \frac{1}{\alpha_0} \mathbb{E}[(W - \eta)_+] + \eta \right\}.$$

Further, for $\eta \leq 0$, we have

$$g(\eta) = \frac{1}{\alpha_0} \mathbb{E}[W] + \left(\frac{1}{\alpha_0} - 1 \right) |\eta| \geq \frac{1}{\alpha_0} \mathbb{E}[W] = g(0).$$

We conclude that the equality (29) holds.

A.2 Proof of Proposition 1

Using the dual of Lemma 2.1, we show that with probability at least $1-\gamma$,

$$\sup_{\eta\in[0,M]}\left|\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)_+\right]-\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{m}\sum_{j=1}^m\ell(\theta;(X_i,Y_{i,j}))-\eta\right)_+\right|$$
$$\le CM\sqrt{\frac{1+|\log\delta|}{\min\{m,n\}}}. \tag{30}$$

As the above gives a uniform approximation to the dual objective $\frac{1}{\alpha_0}\mathbb{E}[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)_+]+\eta$, the proposition will then follow.

To show the result (30), we begin by noting that

$$\sup_{\eta\in[0,M]}\left|\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)_+\right]-\frac{1}{n}\sum_{i=1}^n\left(\frac{1}{m}\sum_{j=1}^m\ell(\theta;(X_i,Y_{i,j}))-\eta\right)_+\right|$$
$$\le\sup_{\eta\in[0,M]}\left|\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)_+\right]-\frac{1}{n}\sum_{i=1}^n(\mathbb{E}[\ell(\theta;(X_i,Y))\mid X=X_i]-\eta)_+\right|$$
$$+\sup_{\eta\in[0,M]}\frac{1}{n}\sum_{i=1}^n\left|(\mathbb{E}[\ell(\theta;(X_i,Y))\mid X=X_i]-\eta)_+-\left(\frac{1}{m}\sum_{j=1}^m\ell(\theta;(X_i,Y_{i,j}))-\eta\right)_+\right|. \tag{31}$$

To bound the first term in the bound (31), note that since $\eta\mapsto(Z-\eta)_+$ is 1-Lipschitz, a standard symmetrization and Rademacher contraction argument [16, 7] yields

$$\sup_{\eta\in[0,M]}\left|\mathbb{E}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)_+\right]-\frac{1}{n}\sum_{i=1}^n(\mathbb{E}[\ell(\theta;(X_i,Y))\mid X=X_i]-\eta)_+\right|\le C\sqrt{\frac{M^2}{n}(1+t)}$$

with probability at least $1-e^{-t}$. To bound the second term in the bound (31), we first note that

$$\sup_{\eta\in[0,M]}\left|(\mathbb{E}[\ell(\theta;(X,Y))\mid X=X_i]-\eta)_+-\left(\frac{1}{m}\sum_{j=1}^m\ell(\theta;(X_i,Y_{i,j}))-\eta\right)_+\right|$$
$$\le\left|\mathbb{E}[\ell(\theta;(X_i,Y_{i,j}))\mid X=X_i]-\frac{1}{m}\sum_{j=1}^m\ell(\theta;(X_i,Y_{i,j}))\right|$$

since $|(x-\eta)_+-(x'-\eta)_+|\le|x-x'|$. The preceding quantity has bound M, and using that its expectation is at most $M/\sqrt{m}$ the bounded differences inequality implies the uniform concentration result (30).

A.3 Proof of Lemma 3.1

Since $Z,\eta\in[0,M]$, we have

$$\mathbb{E}_{X\sim P_X}\left[(Z(X)-\eta)_+^p\right]\le(M-\eta)^{p-1}\mathbb{E}_{X\sim P_X}\left[(Z(X)-\eta)_+\right]$$

which gives the first bound. To get the second bound, note that for a L-Lipschitz function f, we have $\mathbb{E}[f(X)] \le f(\mathbb{E}[X]) + L\mathbb{E}|X - \mathbb{E}[X]|$. Since $f(x) = x^p$ is $p(M-\eta)^{p-1}$-Lipschitz on $[0, M-\eta]$, we get

$$\mathbb{E}_{X\sim P_X}\left[(Z(X)-\eta)_+^p\right] \le \left(\mathbb{E}_{X\sim P_X}\left[(Z(X)-\eta)_+\right]\right)^p + p(M-\eta)^{p-1}\mathbb{E}\left|(Z(X)-\eta)_+ - \mathbb{E}[(Z(X)-\eta)_+]\right|.$$

Taking $1/p$-power on both sides, we obtain the second bound.

A.4 Proof of Lemma 4.1

First, we argue that

$$\begin{aligned}
&\sup_h \left\{ \mathbb{E}\left[h(X)(\ell(\theta;(X,Y))-\eta)\right] \;\middle|\; h:\mathcal{X}\to\mathbb{R},\ \text{measurable},\ h\ge 0,\ \mathbb{E}[h^q(X)]\le 1 \right\} \\
&\le \sup_{h\in\mathcal{H}_{L,p}} \left\{ \mathbb{E}\left[\frac{h(X)}{\epsilon}(\ell(\theta;(X,Y))-\eta)\right] \vee \epsilon^{q-1} \;\middle|\; h\ge 0,\ \left(\mathbb{E}[h^q(X)]\right)^{1/q} \le \epsilon \right\} \qquad (32)
\end{aligned}$$

and for any $\epsilon > 0$. We consider an arbitrary but fixed θ and η.

Suppose that $\epsilon^{q-1} \ge \left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)_+^p\right]\right)^{1/p}$, then

$$\begin{aligned}
\epsilon^{q-1} &\ge \left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)_+^p\right]\right)^{1/p} \\
&= \sup_h \left\{ \mathbb{E}\left[h(X)(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)\right] \;\middle|\; h:\mathcal{X}\to\mathbb{R}\ \text{measurable},\ h\ge 0,\ \mathbb{E}[h^q(X)]\le 1 \right\} \\
&\ge \sup_{h\in\mathcal{H}_{L,p}} \left\{ \mathbb{E}\left[\frac{h(X)}{\epsilon}(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)\right] :\ h\ge 0,\ \left(\mathbb{E}[h^q(X)]\right)^{1/q} \le \epsilon \right\},
\end{aligned}$$

and we have the upper bound. On the other hand, assume $\epsilon^{q-1} \le \left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)_+^p\right]\right)^{1/p}$. The inner supremum in Eq (12) is attained at $h^\star$ defined in expression (13), and from Assumption A2, for any $x, x' \in \mathcal{X}$

$$\begin{aligned}
|h^\star(x) - h^\star(x')| &\le \frac{1}{\epsilon}\left|(\mathbb{E}[\ell(\theta;(X,Y))\mid X=x]-\eta)_+^{p-1} - \left(\mathbb{E}[\ell(\theta;(X,Y))\mid X=x']-\eta\right)_+^{p-1}\right| \\
&\le \frac{1}{\epsilon}\left|(\mathbb{E}[\ell(\theta;(X,Y))\mid X=x]-\eta)_+ - \left(\mathbb{E}[\ell(\theta;(X,Y))\mid X=x']-\eta\right)_+\right|^{p-1} \\
&\le \frac{1}{\epsilon}\left|\mathbb{E}[\ell(\theta;(X,Y))\mid X=x] - \mathbb{E}[\ell(\theta;(X,Y))\mid X=x']\right|^{p-1} \\
&\le \frac{L^{p-1}}{\epsilon}\left\|x-x'\right\|^{p-1}
\end{aligned}$$

where we used $\epsilon \le \left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)_+^p\right]\right)^{1/q}$ in the first inequality. Thus, we conclude that $\epsilon h^\star$ is in $\mathcal{H}_{L,p}$, and obtain the equality

$$\begin{aligned}
&\left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)_+^p\right]\right)^{1/p} \\
&= \sup_h \left\{ \mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X]-\eta)h(X)\right] \middle| h:\mathcal{X}\to\mathbb{R}\ \text{measurable},\ h\ge 0,\ \mathbb{E}[h^q(X)]\le 1,\ \text{and}\ \epsilon h\in\mathcal{H}_{L,p} \right\} \\
&= \sup_{h\in\mathcal{H}_{L,p}} \left\{ \mathbb{E}\left[\frac{h(X)}{\epsilon}(\ell(\theta;(X,Y))-\eta)\right] \;\middle|\; h\ge 0,\ \left(\mathbb{E}[h^q(X)]\right)^{1/q} \le \epsilon \right\}
\end{aligned}$$

where we did a change of variables h to h/ϵ in the last equality. This yields the bound (32).

Now, for $\epsilon = \left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+^p\right]\right)^{1/q}$, the bound (32) is actually a equality. This proves the first claim. To show the second claim, it remains to show that

$$\begin{aligned}
&\sup_{h\in\mathcal{H}_{L,p}} \left\{ \mathbb{E}\left[\frac{h(X)}{\epsilon}(\ell(\theta;(X,Y)) - \eta)\right] \vee \epsilon^{q-1} \;\middle|\; h \geq 0,\ \ (\mathbb{E}[h^q(X)])^{1/q} \leq \epsilon \right\} - \epsilon^{q-1} \\
&\leq \sup_h \left\{ \mathbb{E}\left[h(X)(\ell(\theta;(X,Y)) - \eta)\right] \;\middle|\; h : \mathcal{X} \to \mathbb{R},\ \text{measurable},\ h \geq 0,\ \mathbb{E}[h^q(X)] \leq 1 \right\}.
\end{aligned}$$

If $\epsilon^{q-1} \geq \left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+^p\right]\right)^{1/p}$, then the left hand side is less than or equal to 0 by the same logic above. If $\epsilon^{q-1} \leq \left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+^p\right]\right)^{1/p}$, then we have

$$\begin{aligned}
&\sup_{h\in\mathcal{H}_{L,p}} \left\{ \mathbb{E}\left[\frac{h(X)}{\epsilon}(\ell(\theta;(X,Y)) - \eta)\right] \vee \epsilon^{q-1} \;\middle|\; h \geq 0,\ \ (\mathbb{E}[h^q(X)])^{1/q} \leq \epsilon \right\} \\
&= \sup_h \left\{ \mathbb{E}\left[h(X)(\ell(\theta;(X,Y)) - \eta)\right] \;\middle|\; h : \mathcal{X} \to \mathbb{R},\ \text{measurable},\ h \geq 0,\ \mathbb{E}[h^q(X)] \leq 1 \right\},
\end{aligned}$$

so the result follows.

A.5 Proof of Lemma 4.2

We take the dual of the following optimization problem

$$\begin{aligned}
\underset{h\in\mathbb{R}^n}{\text{maximize}} \quad & \frac{1}{n}\sum_{i=1}^n \frac{h_i}{\epsilon}(\ell(\theta;(X_i,Y_i)) - \eta) \\
\text{subject to} \quad & h_i \geq 0 \ \text{ for all } \ i \in [n], \quad \frac{1}{n}\sum_{i=1}^n h_i^q \leq \epsilon^q, \\
& h_i - h_j \leq L^{p-1}\left\|X_i - X_j\right\|^{p-1} \ \text{ for all } \ i,j \in [n]
\end{aligned}$$

where $h_i := h(X_i)$. To ease notation, we do a change of variables $h_i \leftarrow \frac{h_i}{\epsilon}$

$$\begin{aligned}
\underset{h\in\mathbb{R}^n}{\text{maximize}} \quad & \frac{1}{n}\sum_{i=1}^n h_i(\ell(\theta;(X_i,Y_i)) - \eta) \\
\text{subject to} \quad & h_i \geq 0 \ \text{ for all } \ i \in [n], \quad \frac{1}{n}\sum_{i=1}^n h_i^q \leq 1, \\
& h_i - h_j \leq \frac{L^{p-1}}{\epsilon}\left\|X_i - X_j\right\|^{p-1} \ \text{ for all } \ i,j \in [n].
\end{aligned} \tag{33}$$

For $\gamma \in \mathbb{R}_+^n$, $\lambda \geq 0$, $B \in \mathbb{R}_+^{n\times n}$, the associated Lagrangian is given by

$$\begin{aligned}
\mathcal{L}(h,\gamma,\lambda,B) := \frac{1}{n}\sum_{i=1}^n h_i(\ell(\theta;(X_i,Y_i)) - \eta) + \gamma^\top h + \frac{\lambda}{q}\left(1 - \frac{1}{n}\sum_{i=1}^n h_i^q\right) \\
+ \frac{1}{n^2}\left(\frac{L^{p-1}}{\epsilon}\operatorname{tr}(B^\top D) - h^\top(B\mathbb{1} - B^\top\mathbb{1})\right)
\end{aligned}$$

where $D \in \mathbb{R}^{n\times n}$ is a matrix with entries $D_{ij} = \|X_i - X_j\|^{p-1}$. From strong duality, we have that the primal optimal value (33) is equal to $\inf_{\gamma\in\mathbb{R}^n_+,\lambda\geq 0, B\in\mathbb{R}^{n\times n}_+} \sup_h \mathcal{L}(h,\gamma,\lambda,B)$.

Since $h \mapsto \mathcal{L}(h,\gamma,\lambda,B)$ is a quadratic, a bit of algebra shows that

$$\sup_h \mathcal{L}(h,\gamma,\lambda,B) = \frac{\lambda}{q} + \frac{L^{p-1}}{\epsilon n^2}\operatorname{tr}(B^\top D) + \frac{1}{q\lambda^{p-1}n}\sum_{i=1}^n \left(\ell(\theta;(X_i,Y_i)) - \eta - \frac{1}{n}(B\mathbb{1} - B^\top\mathbb{1})_i + \gamma_i\right)^p.$$

From complementary slackness,

$$\inf_{\gamma\in\mathbb{R}^n_+}\sup_h \mathcal{L}(h,\gamma,\lambda,B) = \frac{\lambda}{q} + \frac{L^{p-1}}{\epsilon n^2}\operatorname{tr}(B^\top D) + \frac{1}{q\lambda^{p-1}n}\sum_{i=1}^n \left(\ell(\theta;(X_i,Y_i)) - \eta - \frac{1}{n}(B\mathbb{1} - B^\top\mathbb{1})_i\right)_+^p.$$

Finally, taking infimum with respect to $\lambda \geq 0$, we obtain

$$\inf_{\lambda\geq 0,\gamma\in\mathbb{R}^n_+}\sup_h \mathcal{L}(h,\gamma,\lambda,B) = \frac{L^{p-1}}{\epsilon n^2}\operatorname{tr}(B^\top D) + \left(\frac{p-1}{n}\sum_{i=1}^n \left(\ell(\theta;(X_i,Y_i)) - \eta - \frac{1}{n}(B\mathbb{1} - B^\top\mathbb{1})_i\right)_+^p\right)^{1/p}.$$

Unpacking the matrix notation, we obtain the result.

A.6 Proof of Lemma 4.3

From the extension theorem for Hölder continuous functions [49, Theorem 1], any $(p-1,$ $L^{p-1})$-Hölder continuous function $h : \{X_1,\dots,X_n\} \to \mathbb{R}$ extends to a $(p-1,\ L^{p-1})$-Hölder continuous $\bar{h} : \mathbb{R}^d \to \mathbb{R}$ with $\operatorname{range}(\bar{h}) \subseteq \operatorname{range}(h)$ so that $h = \bar{h}$ on $\{X_1,\dots,X_n\}$. Since $h \geq 0$ implies $\bar{h} \geq 0$, we have

$$\begin{aligned}\widehat{R}_{p,\epsilon,L}(\theta,\eta) &= \sup_{h\in\widehat{\mathcal{H}}_{L,p}}\left\{\mathbb{E}_{\widehat{P}_n}\left[\frac{h(X)}{\epsilon}(\ell(\theta;(X,Y)) - \eta)\right] :\ h\geq 0,\ \left(\mathbb{E}_{\widehat{P}_n}[h^q(X)]\right)^{1/q} \leq \epsilon\right\}\\ &= \sup_{h\in\mathcal{H}_{L,p}}\left\{\mathbb{E}_{\widehat{P}_n}\left[\frac{h(X)}{\epsilon}(\ell(\theta;(X,Y)) - \eta)\right] :\ h\geq 0,\ \left(\mathbb{E}_{\widehat{P}_n}[h^q(X)]\right)^{1/q} \leq \epsilon\right\}.\end{aligned}$$

To ease notation, define the function

$$R_{c,p,\epsilon,L}(\theta,\eta) := \sup_{h\in\mathcal{H}_{L,p}}\left\{\mathbb{E}\left[\frac{h(X)}{\epsilon}(\ell(\theta;(X,Y)) - \eta)\right] \;\middle|\; h\geq 0,\ (\mathbb{E}[h^q(X)])^{1/q} \leq c\epsilon\right\}$$

so that $R_{p,\epsilon,L} = R_{1,p,\epsilon,L}$. First, we establish a bound between $R_{p,\epsilon,L}$ and $R_{c,p,\epsilon,L}$ for $c \approx 1$.

Claim A.2.

$$\epsilon^{q-1} \vee R_{p,\epsilon,L}(\theta,\eta) \leq \left(\frac{\epsilon}{c}\right)^{q-1} \vee \left\{R_{c,p,\epsilon,L}(\theta,\eta) + (1-c)\left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+^p\right]\right)^{1/p}\right\} \quad \text{if } c<1$$

$$\epsilon^{q-1} \vee R_{c,p,\epsilon,L}(\theta,\eta) \leq c^{q-1}\epsilon^{q-1} \vee \left\{R_{p,\epsilon,L}(\theta,\eta) + (c-1)\left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+^2\right]\right)^{1/p}\right\} \quad \text{if } c>1.$$

Proof of Claim We only prove the bound when $c < 1$ since the proof is similar when $c > 1$. In the case that

$$\left(\frac{\epsilon}{c}\right)^{q-1} \leq \left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+^p\right]\right)^{1/q},$$

the maximizers $h^\star$ and $ch^\star$ (for $h^\star$ defined in expression (13) is contained in the constraint sets that define $R_{p,\epsilon,L}$ and $R_{c,p,\epsilon,L}$ respectively. Hence,

$$R_{p,\epsilon,L}(\theta,\eta) = \left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+^p\right]\right)^{1/p} \text{ and}$$
$$R_{c,p,\epsilon,L}(\theta,\eta) = c\left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+^p\right]\right)^{1/p}$$

and the desired bound holds. Otherwise, we have

$$\epsilon^{q-1} \vee R_{p,\epsilon,L}(\theta,\eta) \le \epsilon^{q-1} \vee \left(\mathbb{E}_{X\sim P_X}\left[(\mathbb{E}[\ell(\theta;(X,Y)) \mid X] - \eta)_+^p\right]\right)^{1/p} \le \left(\frac{\epsilon}{c}\right)^{q-1}.$$

□

Using the two bounds, we now bound $R_{p,\epsilon,L}$ by its empirical counterpart.

To get a upper bound, let us first take $c_1 := (1-\widehat{\delta}_n)^{1/q}$, where we define $\widehat{\delta}_n$

$$\widehat{\delta}_n := \frac{q}{2} \wedge q\epsilon^{-q}L^{p-1}\left((LR)^{p-1}+\epsilon\right)^{q-1} W_{p-1}(\widehat{P}_n, P).$$

Noting that $(1-\delta)^{-1/q} \le 1 + \frac{4}{q}\delta$ for $\delta \in (0, \frac{1}{2}]$, and $1-(1-\delta)^{1/q} \le \frac{2}{q}\delta$, the first bound in Claim A.2 yields

$$\epsilon^{q-1} \vee R_{p,\epsilon,L}(\theta,\eta) \le \epsilon^{q-1} \vee R_{c_1,p,\epsilon,L}(\theta,\eta) + 2^{q-1}\epsilon + \frac{2M}{q}\widehat{\delta}_n. \tag{34}$$

To bound $R_{c_1,p,\epsilon,L}(\theta,\eta)$ by $\widehat{R}_{p,\epsilon,L}(\theta,\eta)$, we first note

$$R_{c_1,p,\epsilon,L}(\theta,\eta) \le \sup_{h\in\mathcal{H}_{L,p}} \left\{ \mathbb{E}\left[\frac{h(X)}{\epsilon}(\ell(\theta;(X,Y)) - \eta)\right] \;\middle|\; h \ge 0,\ \mathbb{E}_{\widehat{P}_n}[h^q(X)] \le \epsilon^q \right\}. \tag{35}$$

Indeed, for $h \in \mathcal{H}_{L,p}$ satisfying $\mathbb{E}_Q[h(X)^q] \le \epsilon^q$ for some probability measure Q, $h^q : \mathcal{X} \to \mathbb{R}$ is bounded by $((LR)^{p-1}+\epsilon)^{q-1}$. Hence, we have for all $x, x' \in \mathcal{X}$

$$|h^q(x) - h^q(x')| \le q\max\left\{h(x), h(x')\right\}^{q-1}|h(x) - h(x')| \le qL^{p-1}\left((LR)^{p-1}+\epsilon\right)^{q-1}\left\|x-x'\right\|^{p-1}.$$

From the definition of the Wasserstein distance W_{p-1},

$$\sup_{h\in\mathcal{H}_{L,p}} \left|\mathbb{E}_{\widehat{P}_n}[h^q(X)] - \mathbb{E}[h^q(X)]\right| \le qL^{p-1}\left((LR)^{p-1}+\epsilon\right)^{q-1} W_{p-1}(\widehat{P}_n, P),$$

which implies that for any $h \in \mathcal{H}_{L,p}$ satisfying $\mathbb{E}[h^q(X)] \le c_1^q\epsilon^q$

$$\mathbb{E}_{\widehat{P}_n}[h^q(X)] \le \mathbb{E}[h^q(X)] + qL^{p-1}\left((LR)^{p-1}+\epsilon\right)^{q-1} W_{p-1}(\widehat{P}_n, P) \le \epsilon^q.$$

To further bound the expression (35), we check that for any $\theta \in \Theta$ and $\eta \in [0, M]$, the map

$(x,y) \mapsto \frac{h(x)}{\epsilon}(\ell(\theta;(x,y)) - \eta)$ is Holder continuous. By Assumption A2, we observe

$$\begin{aligned}
&\left|\frac{h(x)}{\epsilon}(\ell(\theta;(x,y)) - \eta) - \frac{h(x')}{\epsilon}(\ell(\theta;(x',y')) - \eta)\right| \\
&\leq \frac{h(x)}{\epsilon}\left|\ell(\theta;(x,y)) - \ell(\theta;(x',y'))\right| + \left|\ell(\theta;(x',y')) - \eta\right| \frac{|h(x) - h(x')|}{\epsilon} \\
&\leq \frac{(LR)^{p-1} + \epsilon}{\epsilon} L \left\|(x,y) - (x',y')\right\| + \frac{ML^{p-1}}{\epsilon}\left\|x - x'\right\|^{p-1} \\
&= \frac{(LR)^{p-1} + \epsilon}{\epsilon} LR \frac{\|(x,y) - (x',y')\|}{R} + \frac{ML^{p-1}}{\epsilon}\left\|x - x'\right\|^{p-1} \\
&\leq \frac{(LR)^{p-1} + \epsilon}{\epsilon} LR \frac{\|(x,y) - (x',y')\|^{p-1}}{R^{p-1}} + \frac{ML^{p-1}}{\epsilon}\left\|x - x'\right\|^{p-1} \\
&\leq \frac{1}{\epsilon}\left\{LR^{2-p}((LR)^{p-1} + \epsilon) + ML^{p-1}\right\}\left\|(x,y) - (x',y')\right\|^{p-1}
\end{aligned}$$

for all $(x,y),(x',y') \in \mathcal{X} \times \mathcal{Y}$. Using the definition of the Wasserstein distance to bound right hand side of (35),

$$R_{c_1,p,\epsilon,L}(\theta,\eta) \leq \widehat{R}_{p,\epsilon,L}(\theta,\eta) + \frac{1}{\epsilon}\left\{LR^{2-p}((LR)^{p-1} + \epsilon) + ML^{p-1}\right\} W_{p-1}(\widehat{P}_n, P).$$

Plugging in the preceding display in the bound (34), we get

$$\begin{aligned}
\epsilon^{q-1} \vee R_{p,\epsilon,L}(\theta,\eta) \leq \epsilon^{q-1} \vee \widehat{R}_{p,\epsilon,L}(\theta,\eta) &+ 2^{q-1}\epsilon \\
&+ M\left(1 \wedge 2\epsilon^{-q}L^{p-1}\left((LR)^{p-1} + \epsilon\right)^{q-1} W_{p-1}(\widehat{P}_n, P)\right) \\
&+ \frac{1}{\epsilon}\left\{LR^{2-p}((LR)^{p-1} + \epsilon) + ML^{p-1}\right\} W_{p-1}(\widehat{P}_n, P).
\end{aligned}$$

To obtain the lower bound, let $c_2 := (1 + \widehat{\delta}'_n)^{1/q}$ where

$$\widehat{\delta}'_n = q\epsilon^{-q}L^{p-1}\left((LR)^{p-1} + \epsilon\right)^{q-1} W_{p-1}(\widehat{P}_n, P).$$

From the second bound in Claim A.2,

$$\epsilon^{q-1} \vee R_{c_2,p,\epsilon,L}(\theta,\eta) \leq \epsilon^{q-1} \vee R_{p,\epsilon,L}(\theta,\eta) + \left(\frac{\epsilon^{q-1}}{p} + \frac{M}{q}\right)\widehat{\delta}'_n$$

holds, and from a similar argument as before, we have

$$\begin{aligned}
R_{c_2,p,\epsilon,L}(\theta,\eta) &\geq \sup_{h \in \mathcal{H}_{L,p}}\left\{\mathbb{E}\left[\frac{h(X)}{\epsilon}(\ell(\theta;(X,Y)) - \eta)\right] \;\middle|\; h \geq 0,\ \mathbb{E}_{\widehat{P}_n}[h^q(X)] \leq \epsilon^q\right\} \\
&\geq \widehat{R}_{p,\epsilon,L}(\theta,\eta) - \frac{1}{\epsilon}\left\{LR^{2-p}((LR)^{p-1} + \epsilon) + ML^{p-1}\right\} W_{p-1}(\widehat{P}_n, P).
\end{aligned}$$

We conclude that

$$\begin{aligned}
\epsilon^{q-1} \vee \widehat{R}_{p,\epsilon,L}(\theta,\eta) \leq \epsilon^{q-1} &\vee R_{p,\epsilon,L}(\theta,\eta) \\
&+ \left((q-1)L^{p-1}\epsilon^{-1} + M\epsilon^{-q}L^{p-1}\right)\left((LR)^{p-1} + \epsilon\right)^{q-1} W_{p-1}(\widehat{P}_n, P) \\
&+ \frac{1}{\epsilon}\left(LR^{2-p}((LR)^{p-1} + \epsilon) + ML^{p-1}\right) W_{p-1}(\widehat{P}_n, P).
\end{aligned}$$

A.7 Proof of Theorem 2

We use the following concentration result for the Wasserstein distance between an empirical distribution and its population counterpart.

Lemma A.3 ([27, Theorem 2]). *Let $p \in (1,2]$, and $d > 1$. Then, for any $t > 0$*

$$\mathbb{P}\left(W_{p-1}(P, \widehat{P}_n) \geq t\right) \leq c_1 \exp\left(-c_1 n (t^{\frac{d+1}{p-1}} \wedge t^2\right)$$

where c_1 and c_2 are positive constants that depend on M, d, p.

See [27, 44] for general concentration results.

Let $B_\epsilon := LR + \epsilon^{-1} 2^{q-1} L\,(2M + (q-1)LR) + \epsilon^{-q} 2^{q-1} RML^2 + \epsilon^{q-2}(q-1)2^{q-2}L$ to ease notation. From Lemmas 4.3 and A.3, for any fixed $\epsilon > 0$, with probability at least $1 - c_1 \exp\left(-c_2 n (t^{\frac{d+1}{p-1}} \wedge t^2\right)$

$$\begin{aligned}
\sup_{Q_0(x) \in \mathcal{P}_{\alpha_0, X}} \mathbb{E}_{X \sim Q_0}[\mathbb{E}[\ell(\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}; (X,Y)) \mid X]] &\leq \inf_{\eta \in [0,M]} \left\{ \frac{1}{\alpha_0} \left(R_{p,\epsilon,L}(\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}, \eta) \vee \epsilon^{q-1} \right) + \eta \right\} \\
&\leq \inf_{\eta \in [0,M]} \left\{ \frac{1}{\alpha_0} \left(\widehat{R}_{p,\epsilon,L}(\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}, \eta) \vee \epsilon^{q-1} \right) + \eta \right\} + \frac{B_\epsilon t}{\alpha_0} \\
&\leq \inf_{\eta \in [0,M]} \left\{ \frac{1}{\alpha_0} \left(\widehat{R}_{p,\epsilon,L}(\theta, \eta) \vee \epsilon^{q-1} \right) + \eta \right\} + \frac{B_\epsilon t}{\alpha_0}
\end{aligned}$$

for any $\theta \in \Theta$, where we used the fact that $\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}$ is an empirical minimizer.

Applying uniform convergence of $\widehat{R}_{p,\epsilon,L}(\theta, \eta)$ to $R_{p,\epsilon,L}$ again (Lemmas 4.3 and A.3), we get

$$\begin{aligned}
&\sup_{Q_0(x) \in \mathcal{P}_{\alpha_0, X}} \mathbb{E}_{X \sim Q_0}[\mathbb{E}[\ell(\widehat{\theta}^{\mathrm{rob}}_{n,\epsilon}; (X,Y)) \mid X]] \\
&\leq \inf_{\eta \in [0,M]} \left\{ \frac{1}{\alpha_0} \left(R_{p,\epsilon,L}(\theta, \eta) \vee \epsilon \right) + \eta \right\} + \frac{2B_\epsilon t}{\alpha_0} \\
&\leq \inf_{\eta \in [0,M]} \left\{ \frac{1}{\alpha_0} \left(\mathbb{E}_{X \sim P_X}\left[(\mathbb{E}[\ell(\theta; (X,Y)) \mid X] - \eta)_+^p \right] \right)^{1/p} + \eta \right\} + \frac{\epsilon^{q-1}}{\alpha_0} + \frac{2B_\epsilon t}{\alpha_0}
\end{aligned}$$

with probability at least $1 - 2c_1 \exp\left(-c_2 n (t^{\frac{d+1}{p-1}} \wedge t^2\right)$, where we used the second bound of Lemma 4.1. Taking infimum over $\theta \in \Theta$, we obtain the result.

A.8 Proof of Proposition 3

Since our desired bound holds trivially if $\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X,Y)) \mid X] - \eta\right)_+^p\right)^{1/p} \leq \epsilon^{q-1}$, we assume $\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X,Y)) \mid X] - \eta\right)_+^p\right)^{1/p} \geq \epsilon^{q-1}$. First, we rewrite the left hand side as

$$\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X,Y)) \mid X] - \eta\right)_+^p\right)^{1/p} = \frac{\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X,Y)) \mid X] - \eta\right)_+^p}{\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X,Y)) \mid X] - \eta\right)_+^p\right)^{1/q}} = \frac{\mathbb{E}[Z(\theta, \eta; (X,Y))]}{\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X,Y)) \mid X] - \eta\right)_+^p\right)^{1/q}}$$

where for convenience we defined

$$Z(\theta, \eta; (X,Y)) := \left(\ell(\theta; (X,Y)) - \eta\right)\left(\mathbb{E}[\ell(\theta; (X,Y)) \mid X] - \eta\right)_+^{p-1}.$$

Now, note that $\eta \mapsto Z(\theta, \eta; (X, Y))$ is pM-Lispchitz. Applying a standard bracketing number argument for uniform concentration of Lipschitz functions [68, Theorem 2.7.11]

$$\sup_{\eta \in [0,M]} \left| \mathbb{E}[Z(\theta, \eta; (X, Y))] - \mathbb{E}_{\widehat{P}_n}[Z(\theta, \eta; (X, Y))] \right| \leq c_1 M^2 \sqrt{\frac{\log \frac{1}{\gamma}}{n}}$$

with probability at least $1 - \gamma$, where c_1 is some universal constant. We conclude that with probability at least $1 - \gamma$, for all $\eta \in [0, M]$

$$\begin{aligned} \left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta\right)_+^p\right)^{1/p} \leq \left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta\right)_+^p\right)^{-1/q} \mathbb{E}_{\widehat{P}_n}[Z(\theta, \eta; (X, Y))] \\ + \left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta\right)_+^p\right)^{-1/q} c_1 M^2 \sqrt{\frac{\log \frac{1}{\gamma}}{n}}. \end{aligned} \tag{36}$$

Next, we upper bound the first term by our empirical objective $\widehat{R}_{p,\epsilon,L}(\theta, \eta)$

$$\begin{aligned} &\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta\right)_+^p\right)^{-1/q} \mathbb{E}_{\widehat{P}_n}[Z(\theta, \eta; (X, Y))] \\ &= (1 + \tau_n(\gamma, \epsilon))^{1/q} \, \mathbb{E}_{\widehat{P}_n}\left[\frac{h_\eta^\star(X)}{(1 + \tau_n(\gamma, \epsilon))^{1/q}} (\ell(\theta; (X, Y)) - \eta)\right]. \end{aligned}$$

Uniform concentration of Lipschitz functions [68, Theorem 2.7.11] implies that there exists a universal constant $c_2 > 0$ such that with probability at least $1 - \gamma$

$$\mathbb{E}_{\widehat{P}_n}\left(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta\right)_+^p \leq \mathbb{E}\left(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta\right)_+^p + c_2 M^2 \sqrt{\frac{1}{n} \log \frac{1}{\gamma}}$$

for all $\eta \in [0, M]$. Recalling the definition (13) of $h_\eta^\star(x)$ (where we now make the dependence on η explicit), we have

$$\mathbb{E}_{\widehat{P}_n}[h_\eta^\star(X)^q] \leq 1 + c_2 M^2 \left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta\right)_+^p\right)^{-1} \sqrt{\frac{1}{n} \log \frac{1}{\gamma}}. \tag{37}$$

with probability at least $1 - \gamma$.

Recalling the definition (17) of $\widehat{\mathcal{H}}_{L,p}$, since $x \mapsto h_\eta^\star(x)$ is $\frac{L}{\epsilon}$-Lipschitz, we get

$$\begin{aligned} &\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta\right)_+^p\right)^{-1/q} \mathbb{E}_{\widehat{P}_n}[Z(\theta, \eta; (X, Y))] \\ &\leq (1 + \tau_n(\gamma, \epsilon))^{1/q} \sup_{h \in \mathcal{H}_{L_n(\gamma), n}} \left\{ \mathbb{E}_{\widehat{P}_n}\left[\frac{h(X)}{\epsilon} (\ell(\theta; (X, Y)) - \eta)\right] \;\middle|\; \mathbb{E}_{\widehat{P}_n}[h^q(X)] \leq \epsilon^q \right\} \end{aligned}$$

with probability at least $1 - \gamma$, where we used the bound (37) in the second inequality. Combining the preceeding display with the bound (36), with probability at least $1 - 2\gamma$,

$$\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta; (X, Y)) \mid X] - \eta\right)_+^p\right)^{1/p} \leq (1 + \tau_n(\gamma, \epsilon))^{1/q} \, \widehat{R}_{p,\epsilon,L_n(\gamma)}(\theta, \eta) + \frac{c_1 M^2}{\epsilon^{q-1}} \sqrt{\frac{1}{n} \log \frac{1}{\gamma}}$$

for all $\eta \in [0, M]$.

A.9 Proof of Lemma 5.1

First, note that variational form for the $L^p(P)$-norm gives

$$\begin{aligned}&\left(\mathbb{E}_{(X,C)\sim P_{X,C}}\left[(\mathbb{E}[\ell(\theta;(X,Y))\mid X,C]-\eta)_+^p\right]\right)^{1/p} \\ &=\sup_h\left\{\mathbb{E}[\bar{h}(X,C)(\ell(\theta;(X,Y))-\eta)]:\ \bar{h}:\mathcal{X}\times\mathcal{C}\to\mathbb{R}\text{ measurable},\ \bar{h}\geq 0,\ \mathbb{E}[\bar{h}(X,C)^q]\leq 1\right\}.\end{aligned} \tag{38}$$

For ease of notation, let

$$e(x):=\mathbb{E}[\ell(\theta;(X,Y))\mid X=x],\quad e(x,c):=\mathbb{E}[\ell(\theta;(X,Y))\mid X=x,C=c].$$

We first show the equality (27). To see that "$\geq$" direction holds, let $\epsilon:=\left(\mathbb{E}\left(e(X,C)-\eta\right)_+^p\right)^{1/q}$. Then, we have

$$\begin{aligned}R_{p,\epsilon,L,\delta}(\theta,\eta)&\leq\sup_{h,f\ \text{measurable}}\left\{\mathbb{E}\left[(h(X)+f(X,C))(\ell(\theta;(X,Y))-\eta)\right]:h+f\geq 0,\ \mathbb{E}[(h(X)+f(X,C))^q]\leq 1\right\}\\&=\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y))\mid X,C]-\eta\right)_+^p\right)^{1/p}\end{aligned} \tag{39}$$

where we used the variational form (38) in the last inequality.

For the "$\leq$" inequality, fix an arbitrary $\epsilon>0$. If $\left(\mathbb{E}\left(e(X,C)-\eta\right)_+^p\right)^{1/q}\leq\epsilon$, then the bound follows. Otherwise, consider $\left(\mathbb{E}\left(e(X,C)-\eta\right)_+^p\right)^{1/q}>\epsilon>0$. Note that the supremum in the variational form (38) is attained by

$$\bar{h}^\star(x,c):=\frac{(e(x,c)-\eta)_+^{p-1}}{\left(\mathbb{E}\left(e(X,C)-\eta\right)_+^p\right)^{1/q}}.$$

Now, define

$$\begin{aligned}h^\star(x)&:=\frac{(e(x)-\eta)_+^{p-1}}{\left(\mathbb{E}\left(e(X,C)-\eta\right)_+^p\right)^{1/q}},\\ f^\star(x,c)&:=\frac{1}{\left(\mathbb{E}\left(e(X,C)-\eta\right)_+^p\right)^{1/q}}\left((e(x,c)-\eta)_+^{p-1}-(e(x)-\eta)_+^{p-1}\right)\end{aligned}$$

so that $\bar{h}^\star=h^\star+f^\star$. Since $\epsilon h^\star\in\mathcal{H}_{L,p}$ and $\|f^\star(X,C)\|_{L^\infty(P)}\leq\frac{\delta^{p-1}}{\epsilon}$, $h^\star$ and $f^\star$ are in the feasible region of the maximization problem (26). We conclude that

$$\begin{aligned}&\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y))\mid X,C]-\eta\right)_+^p\right)^{1/p}\\&\leq\inf_{\epsilon\geq 0}\left\{\epsilon\vee\sup_{h,f\ \text{meas.}}\left\{\mathbb{E}\left[(h(X)+f(X,C))(\ell(\theta;(X,Y))-\eta)\right]:\right.\right.\\&\qquad\left.\left.h+f\geq 0,\ \mathbb{E}[(h(X)+f(X,C))^q]\leq 1,\ \epsilon h\in\mathcal{H}_{L,p},\ \|f(X,C)\|_{L^\infty(P)}\leq\frac{\delta^{p-1}}{\epsilon}\right\}\right\}.\end{aligned}$$

Rescaling the supremum problem by $1/\epsilon$, we obtain the first result (27).

To show the second result, fix an arbitrary $\epsilon>0$. If $\left(\mathbb{E}\left(e(X,C)-\eta\right)_+^p\right)^{1/q}\leq\epsilon$, then from our upper bound (39)

$$(R_{p,\epsilon,L,\delta}(\theta,\eta)\vee\epsilon^{q-1})-\epsilon^{q-1}\leq 0$$

so that our desired result trivially holds. On the other hand, if $\left(\mathbb{E}\left(e(X,C)-\eta\right)_+^p\right)^{1/q} > \epsilon$, then

$$R_{p,\epsilon,L,\delta}(\theta,\eta) = \left(\mathbb{E}_{(X,C)\sim p_{X,C}}\left[\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^p\right]\right)^{1/p}$$

from our argument above, so the desired result again holds.

A.10 Proof of Proposition 4

We proceed similarly as in the proof of Proposition 3. Letting

$$Z(\theta,\eta;(X,C,Y)) := \left(\ell(\theta;(X,Y))-\eta\right)\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^{p-1},$$

rewrite the L^p-norm as

$$\begin{aligned}\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^p\right)^{1/p} &= \frac{\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^p}{\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^p\right)^{1/q}} \\ &= \frac{\mathbb{E}[Z(\theta,\eta;(X,C,Y))]}{\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^p\right)^{1/q}}.\end{aligned}$$

Since $(\theta,\eta) \mapsto Z(\theta,\eta;(X,C,Y))$ is pM-Lispchitz, we again get from a standard bracketing number argument for uniform concentration of Lipschitz functions [68, Theorem 2.7.11]

$$\sup_{\eta\in[0,M]}\left|\mathbb{E}[Z(\theta,\eta;(X,C,Y))] - \mathbb{E}_{\widehat{P}_n}[Z(\theta,\eta;(X,C,Y))]\right| \le c_1 M^2\sqrt{\frac{\log\frac{1}{\gamma}}{n}} \tag{40}$$

with probability at least $1-\gamma$, where c_1 is some universal constant. Hence, with probability at least $1-\gamma$, for all $\theta\in\Theta, \eta\in[0,M]$

$$\begin{aligned}\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^p\right)^{1/p} \le \left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^p\right)^{-1/q}\mathbb{E}_{\widehat{P}_n}[Z(\theta,\eta;(X,C,Y))] \\ + \left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^p\right)^{-1/q} c_1 M^2\sqrt{\frac{\log\frac{1}{\gamma}}{n}}.\end{aligned} \tag{41}$$

Next, we upper bound $\mathbb{E}_{\widehat{P}_n}[Z(\theta,\eta;(X,C,Y))]$ by our empirical objective $\widehat{R}_{p,\epsilon,L,\delta}(\theta,\eta)$. To this end, uniform concentration of Lipschitz functions [68, Theorem 2.7.11] again yields

$$\mathbb{E}_{\widehat{P}_n}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^{p-1} \le \mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C]-\eta\right)_+^{p-1} + c_2 M^2\sqrt{\frac{\log\frac{1}{\gamma}}{n}} \tag{42}$$

for all $\theta\in\Theta, \eta\in[0,M]$, with probability at least $1-\gamma$. Define the functions

$$\begin{aligned}h_\eta^\star(x) &:= \frac{\left(e(x)-\eta\right)_+^{p-1}}{\left(\mathbb{E}\left(e(X,C)-\eta\right)_+^p\right)^{1/q}}, \\ f^\star(x,c) &:= \frac{1}{\left(\mathbb{E}\left(e(X,C)-\eta\right)_+^p\right)^{1/q}}\left(\left(e(x,c)-\eta\right)_+^{p-1} - \left(e(x)-\eta\right)_+^{p-1}\right),\end{aligned}$$

and note that

$$\mathbb{E}_{\widehat{P}_n}[(h_\eta^\star(X) + f^\star(X,C))^q] \le 1 + c_2 M^2 \left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C] - \eta\right)_+^p\right)^{-1/q} \sqrt{\frac{\log\frac{1}{\gamma}}{n}}. \quad (43)$$

with probability at least $1-\gamma$.

Since our desired bound holds trivially if $\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C] - \eta\right)_+^p\right)^{1/q} \le \epsilon$, we now assume that $\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C] - \eta\right)_+^p\right)^{1/q} \ge \epsilon$. Since $\epsilon h_\eta^\star(x) \in \mathcal{H}_{L,p}$, we have

$$\begin{aligned}
&\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C] - \eta\right)_+^p\right)^{-1/q} \mathbb{E}_{\widehat{P}_n}[Z(\theta,\eta;(X,C,Y))] \\
&= (1+\tau_n(\gamma,\epsilon))^{1/q} \mathbb{E}_{\widehat{P}_n}\left[\frac{h_\eta^\star(X) + f^\star(X,C)}{(1+\tau_n(\gamma,\epsilon))^{1/q}}(\ell(\theta;(X,Y)) - \eta)\right] \\
&\le (1+\tau_n(\gamma,\epsilon))^{1/q} \sup_{h\in\mathcal{H}_{L_n(\gamma),n}, f\in\mathcal{F}_{\delta_n(\gamma),p,n}} \left\{\mathbb{E}_{\widehat{P}_n}\left[\frac{h(X)+f(X,C)}{\epsilon}(\ell(\theta;(X,Y)) - \eta)\right] \;\middle|\; \right. \\
&\qquad\qquad \left. \mathbb{E}_{\widehat{P}_n}[(h(X)+f(X,C))^q] \le \epsilon^q \right\}
\end{aligned}$$

with probability at least $1-\gamma$, where we used the bound (43) in the second inequality. Combining the preceeding display with the bound (41), with probability at least $1-2\gamma$,

$$\left(\mathbb{E}\left(\mathbb{E}[\ell(\theta;(X,Y)) \mid X,C] - \eta\right)_+^p\right)^{1/p} \le (1+\tau_n(\gamma,\epsilon))^{1/q}\ \widehat{R}_{p,\epsilon,L_n(\gamma),\delta_n(\gamma)}(\theta,\eta) + \frac{c_1 M^2}{\epsilon^{q-1}}\sqrt{\frac{\log\frac{1}{\gamma}}{n}}.$$

for all $\theta \in \Theta, \eta \in [0,M]$.

A.11 Proof of Lemma 5.2

We take the dual of the optimization problem

$$\begin{aligned}
&\underset{h,f\in\mathbb{R}^n}{\text{maximize}}\ \frac{1}{n}\sum_{i=1}^n \frac{h_i+f_i}{\epsilon}(\ell(\theta;(X_i,Y_i)) - \eta) \\
&\text{subject to}\quad h_i + f_i \ge 0 \ \text{ for all } \ i\in[n],\quad \frac{1}{n}\sum_{i=1}^n (h_i+f_i)^q \le \epsilon^q, \\
&\qquad\qquad h_i - h_j \le L^{p-1}\|X_i - X_j\|^{p-1} \ \text{ for all } \ i,j\in[n], \\
&\qquad\qquad |f_i| \le \delta^{p-1} \ \text{ for all } \ i\in[n]
\end{aligned}$$

where $h_i := h(X_i)$ and $f_i = f(X_i,C_i)$. To ease notation, we do a change of variables $h_i \leftarrow \frac{h_i}{\epsilon}$, $f_i \leftarrow \frac{f_i}{\epsilon}$ and $q_i \leftarrow h_i + f_i$ which gives

$$\begin{aligned}
&\underset{q,h\in\mathbb{R}^n}{\text{maximize}}\ \frac{1}{n}\sum_{i=1}^n q_i(\ell(\theta;(X_i,Y_i)) - \eta) \qquad (44) \\
&\text{subject to}\quad q_i \ge 0 \ \text{ for all } \ i\in[n],\quad \frac{1}{n}\sum_{i=1}^n q_i^q \le 1, \\
&\qquad\qquad h_i - h_j \le \frac{L^{p-1}}{\epsilon}\|X_i - X_j\|^{p-1} \ \text{ for all } \ i,j\in[n], \\
&\qquad\qquad |q_i - h_i| \le \frac{\delta^{p-1}}{\epsilon} \ \text{ for all } \ i\in[n]. \qquad (45)
\end{aligned}$$

For $\gamma \in \mathbb{R}^n_+$, $\lambda \geq 0$, $B \in \mathbb{R}^{n\times n}_+$, $\alpha^+, \alpha^- \in \mathbb{R}^n_+$, the associated Lagrangian is

$$\begin{aligned}\mathcal{L}(q, h, \gamma, \lambda, B, \alpha^+, \alpha^-) := \frac{1}{n}\sum_{i=1}^n q_i(\ell(\theta; (X_i, Y_i)) - \eta) + \frac{1}{n}\gamma^\top q + \frac{\lambda}{2}\left(1 - \frac{1}{n}\sum_{i=1}^n q_i^q\right) \\ + \frac{1}{n^2}\left(\frac{L^{p-1}}{\epsilon}\operatorname{tr}(B^\top D) - h^\top(B\mathbb{1} - B^\top\mathbb{1})\right) \\ + \frac{\alpha^{+\top}}{n}\left(\frac{\delta^{p-1}}{\epsilon}\mathbb{1} - (q-h)\right) + \frac{\alpha^{-\top}}{n}\left(\frac{\delta^{p-1}}{\epsilon}\mathbb{1} + (q-h)\right)\end{aligned}$$

where $D \in \mathbb{R}^{n\times n}$ is a matrix with entries $D_{ij} = \|X_i - X_j\|^{p-1}$. From strong duality, the primal optimal value (44) is $\inf_{\gamma\in\mathbb{R}^n_+, \lambda\geq 0, B\in\mathbb{R}^{n\times n}_+, \alpha^+, \alpha^-\in\mathbb{R}^n_+} \sup_{q,h} \mathcal{L}(q, h, \gamma, \lambda, B, \alpha^+, \alpha^-)$.

The first order conditions for the inner supremum give

$$\begin{aligned}q_i^\star &:= \frac{1}{n\lambda}\left(\ell(\theta; (X_i, Y_i)) - \eta + n\gamma - \alpha_i^+ + \alpha_i^-\right) \\ \frac{1}{n}(B\mathbb{1} - B^\top\mathbb{1}) &= \left(\alpha^+ - \alpha^-\right).\end{aligned}$$

By nonnegativity of B and α^+, α^-, the second equality implies that $\alpha^+ = \frac{1}{n}B\mathbb{1}$ and $\alpha^- = \frac{1}{n}B^\top\mathbb{1}$. Substituting these values and infimizing out $\lambda, \gamma \geq 0$ as in Lemma 4.2, we obtain

$$\begin{aligned}\inf_{\lambda\geq 0, \gamma\in\mathbb{R}^n_+} \sup_{q,h} \mathcal{L}(q, , h, \gamma, \lambda, B, \alpha^+, \alpha^-) = \left(\frac{p-1}{n}\sum_{i=1}^n \left(\ell(\theta; (X_i, Y_i)) - \frac{1}{n}\sum_{j=1}^n (B_{ij} - B_{ji}) - \eta\right)_+^p\right)^{1/p} \\ + \frac{L^{p-1}}{\epsilon n^2}\sum_{i,j=1}^n \|X_i - X_j\| B_{ij} + \frac{2\delta^{p-1}}{\epsilon n^2}\sum_{i,j=1}^n |B_{ij}|.\end{aligned}$$

Taking the infimum with respect to $B \in \mathbb{R}^{n\times n}_+$ gives the lemma.