Distributionally robust risk evaluation with an isotonic constraint

Yu Gui, Rina Foygel Barber, and Cong Ma

Department of Statistics, University of Chicago

December 19, 2024

Abstract

Statistical learning under distribution shift is challenging when neither prior knowledge nor data from the target distribution is available. Distributionally robust learning (DRL) aims to control the worst-case statistical performance within a set of candidate distributions, but how to properly specify the set remains challenging. To enable distributional robustness without being overly conservative, in this paper we propose a shape-constrained approach to DRL, which incorporates prior information about the way in which the unknown target distribution differs from its estimate—specifically, we assume the unknown density ratio between the target distribution and its estimate is isotonic with respect to some partial order. At the population level, we provide a solution to the shape-constrained optimization problem that can be solved without the challenge of an explicit isotonic constraint. At the sample level, we provide consistency results for an empirical estimator of the target in a range of different settings. Empirical studies on both synthetic and real data demonstrate the improved efficiency of the proposed shape-constrained approach.

1 Introduction

Evaluating the performance of an estimator is of significant importance in statistics. To give several motivating examples, we first consider supervised learning settings, where our observations consist of features $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and a response $Y \in \mathcal{Y} \subseteq \mathbb{R}$:

- Given a fitted model $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$, we may want to estimate the expected value of the squared error $(Y - \widehat{\mu}(X))^2$ with respect to a target distribution on (X, Y).
- Or, in predictive inference, suppose we have constructed a prediction band $\widehat{C}_{1-\alpha}$, where $\widehat{C}_{1-\alpha}(X) \subseteq \mathbb{R}$ is a confidence region for the response Y given features X, and $1 - \alpha$ denotes the target coverage level. Then to determine whether $\widehat{C}_{1-\alpha}$ does in fact achieve coverage at level $1 - \alpha$ for data points drawn from some target distribution, we would like to estimate the expected value of $\mathbb{1}\{Y \notin \widehat{C}_{1-\alpha}(X)\}$ with respect to this target distribution. This is the probability that our interval *fails* to cover the response.

We can also consider unsupervised learning settings, where observations consist only of features $X \in \mathcal{X} \subseteq \mathbb{R}^d$:

- In principal component analysis (PCA), suppose we have obtained a set of pre-fitted principal components $\widehat{\mathcal{V}}_K = \{\widehat{v}_1, \ldots, \widehat{v}_K\}$ which forms an orthonormal basis for a K-dimensional subspace of $\mathbb{R}^d$. To evaluate how well the variance in X is explained by the top K principal components, it would be of interest to analyze the expected value of the reconstruction error $\|X - \sum_{k=1}^K (X^\top \widehat{v}_k)\widehat{v}_k\|^2$ with respect to the distribution of X.

- Another example is density estimation. In this case, given a density estimate P_θ learned from data, we may want to evaluate its performance using the expected log-likelihood $-\log \mathsf{d}P_\theta(X)$ over a target distribution P_{target}. In fact, $\mathbb{E}_{P_{\text{target}}}[-\log \mathsf{d}P_\theta(X)]$ is the cross-entropy of P_θ relative to P_{target}.

A key challenge for any of these problems is that the target distribution (say, the distribution of the general population) may be unknown, and our available data (say, individuals who participate in our study) may be drawn from a different distribution.

1.1 Problem formulation

To make the problem more concrete, and unify the examples mentioned above, here we introduce some notation to formulate the question at hand.

The unsupervised setting. Let $R : \mathcal{X} \to \mathbb{R}_+$ denote a *risk function*, where our goal is to evaluate the expected value $\mathbb{E}_{P_{\text{target}}}[R(X)]$ with respect to some target distribution P_{target} over $\mathcal{X}$. However, the available data only provides information about P, a potentially different distribution. Using a calibration data set comprised of samples $X_1, \ldots, X_n$ drawn from P, we can estimate $\mathbb{E}_P[R(X)]$ with the empirical mean, $n^{-1}\sum_{i=1}^n R(X_i))$. Our aim, though, is to provide a bound on the risk $\mathbb{E}_{P_{\text{target}}}[R(X)]$—or, at least, to bound the difference in risks (often called the *excess risk*), $\mathbb{E}_{P_{\text{target}}}[R(X)] - \mathbb{E}_P[R(X)]$.

If we assume that the unknown distribution P_{target} lies in some class $\mathcal{Q}$ (to be specified later on), then defining the *worst-case excess risk*

$$\Delta(R; \mathcal{Q}) = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[R(X)] - \mathbb{E}_P[R(X)], \tag{1}$$

we can bound the risk under distribution P_{target} by $\mathbb{E}_{P_{\text{target}}}[R(X)] \le \mathbb{E}_P[R(X)] + \Delta(R; \mathcal{Q})$.

The supervised setting: covariate shift assumption. In the supervised learning setting, the data contains both features X and a response Y. Here we will consider a loss function $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, for instance, $r(x, y) = (y - \widehat{\mu}(x))^2$ for the squared error in a regression, or $r(x, y) = \mathbb{1}\{y \notin \widehat{C}_{1-\alpha}(x)\}$ for characterizing the (mis)coverage of a prediction interval in predictive inference.

Throughout this paper, for the supervised learning setting, we will assume the *covariate shift* setting, where the distribution of the available data and the target distribution may differ in the marginal distribution of the covariates X, but share the same conditional distribution $Y \mid X$. To make this concrete, if our calibration data consists of n data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from $\widetilde{P}$, while our goal is to control the expected loss with respect to the target distribution $\widetilde{P}_{\text{target}}$ on (X, Y), we will assume that we can write

$$\text{data distribution: } \widetilde{P} = P \times P_{Y|X}, \quad \text{target distribution: } \widetilde{P}_{\text{target}} = P_{\text{target}} \times P_{Y|X},$$

so that $\widetilde{P}$ and $\widetilde{P}_{\text{target}}$ share the same conditional distribution $P_{Y|X}$ for $Y \mid X$.

In fact, under covariate shift, this supervised setting can be unified with the unsupervised one by defining the risk $R(X) = \mathbb{E}[r(X,Y) \mid X]$, which is the conditional expectation of $r(X,Y)$ under *either* $\widetilde{P}$ or $\widetilde{P}_{\text{target}}$. The quantity of interest is then given by $\mathbb{E}_{P_{\text{target}}}[R(X)] = \mathbb{E}_{\widetilde{P}_{\text{target}}}[r(X,Y)]$, but our calibration data, which is sampled from P, instead enables us to estimate $\mathbb{E}_P[R(X)] = \mathbb{E}_{\widetilde{P}}[r(X,Y)]$. If we again assume that $P_{\text{target}} \in \mathcal{Q}$, then $\Delta(R;\mathcal{Q})$ again allows us to bound the risk of our estimator under the target distribution:

$$\mathbb{E}_{\widetilde{P}_{\text{target}}}[r(X,Y)] \leq \mathbb{E}_{\widetilde{P}}[r(X,Y)] + \Delta(R;\mathcal{Q}).$$

Estimating the risk or tuning the model? In this paper, we consider the setting where our estimator—say, a prediction band $\widehat{C}_{1-\alpha}$—is *pretrained*, meaning that we have available calibration data sampled from P (in the unsupervised setting) or $\widetilde{P}$ (in the supervised setting) that is independent of the fitted estimator. Consequently, our available calibration data provides us with an unbiased estimate of $\mathbb{E}_P[R(X)]$ (or, equivalently in the supervised setting, $\mathbb{E}_{\widetilde{P}}[r(X,Y)]$); given a constraint set $\mathcal{Q}$, we can then use this estimate to bound $\mathbb{E}_{P_{\text{target}}}[R(X)]$ (or, in the supervised setting, $\mathbb{E}_{\widetilde{P}_{\text{target}}}[r(X,Y)]$).

In some settings, the goal may be to estimate the risk of each estimator within a family of (pretrained) options, in order to select a good estimator. Returning again to the example of a prediction band, suppose, we actually are given a nested family of prediction bands, $\{\widehat{C}_{1-a} : a \in [0,1]\}$, where $1-a$ denotes the confidence level. Choosing $R_a(X) = \mathbb{P}_{P_{Y|X}}(Y \notin \widehat{C}_{1-a}(X))$ or accordingly, $r_a(X,Y) = \mathbb{1}\{Y \notin \widehat{C}_{1-a}(X)\}$, then, if we can compute a bound on the miscoverage rate $\mathbb{E}_{P_{\text{target}}}[R_a(X)]$ for each a, then we can choose a value of a that achieves some desired level of coverage. More generally, we may do the same in other settings as well—that is, given a family of candidate estimators, bounding the risk of each one under the distribution P_{target} provides an intermediate step towards choosing the tuning parameter.

Throughout this paper, then, we will primarily discuss the question of estimating the expected risk. Later on, in our experiments, we will turn to the aim of using these estimates to tune a procedure for achieving a desired bound on the error.

1.2 Prior work: distributionally robust learning

Our work builds upon the distributionally robust learning (DRL) literature (Ben-Tal and Nemirovski, 1998; El Ghaoui et al., 1998; Lam, 2016; Duchi and Namkoong, 2018), which is a well-established framework for risk evaluation under distribution shift. In this framework, the target distribution P_{target} is assumed to lie in some neighborhood around the distribution P of the available data—for instance, we might assume that $D_{\text{KL}}(P_{\text{target}} \| P) \leq \rho$, where D_{KL} denotes the Kullbeck–Leibler (KL) divergence. DRL takes a conservative approach and evaluate the performance on P_{target} via its upper bound, i.e., the worst-case performance over all distributions within the specified neighborhood of P,

$$\mathbb{E}_{P_{\text{target}}}[R(X)] \leq \sup\left\{\mathbb{E}_Q[R(X)] : D_{\text{KL}}(Q\|P) \leq \rho\right\}, \tag{2}$$

or, equivalently, $\mathbb{E}_{P_{\text{target}}}[R(X)] \leq \mathbb{E}_P[R(X)] + \Delta(R; \mathcal{Q}_{\text{KL}}(\rho))$, where $\Delta(R; \mathcal{Q}_{\text{KL}}(\rho))$ is defined as in (1) with $\mathcal{Q} = \mathcal{Q}_{\text{KL}}(\rho) = \{Q : D_{\text{KL}}(Q \| P) \leq \rho\}$. More generally, we can consider divergence measures beyond the KL distance, as we will describe in more detail below.

1.3 Our proposal: iso-DRL

If the assumption $D_{\text{KL}}(P_{\text{target}} \| P) \leq \rho$ is correct, then the upper bound (2) is valid. However, since this bound uses only the KL divergence to define the constraint $P_{\text{target}} \in \mathcal{Q}$ on the target distribution, it could be quite conservative. In many practical settings, additional side information or prior knowledge on the structure of the distribution shift may allow for a tighter bound, which would be less conservative than the worst-case excess risk of DRL (2). This raises the following key question:

Can we use side information on the distribution shift between the distribution P and the target distribution P_{target}, to improve the worst-case excess risk of DRL in risk evaluation?

In this paper, we study one specific example of this type of setting: we assume that the density ratio $(\mathsf{d}P_{\text{target}}/\mathsf{d}P)(\cdot)$ between the target distribution and the data distribution is (approximately) isotonic (i.e., monotone) with respect to some order or partial order on $\mathcal{X}$.

Motivation: recalibration of an estimated density ratio. To motivate the use of such side information, consider a practical supervised setting where we have an initial estimate w_0 for the density ratio:

$$w_0(x) \approx \frac{\mathsf{d}P_{\text{target}}}{\mathsf{d}P}(x).$$

This ratio is possible to estimate in settings where, in addition to labeled data (i.e., (X, Y) pairs) sampled from the data distribution $P \times P_{Y|X}$, we also have access to unlabeled (i.e., X only) data from the target population P_{target}. We may use these two data sets to train w_0. Although there is no guarantee that the estimate w_0 is accurate, the shape or relative magnitude of w_0 may provide us with useful side information: large values of w_0 can identify portions of the target population that are *underrepresented* under the data distribution P. This motivates us to recalibrate w_0 within the set of density ratios that are isotonic in w_0.

To express this scenario in the notation of the problem formulation above, we assume that the target distribution P_{target} satisfies an isotonicity constraint, $P_{\text{target}} \in \mathcal{Q}_{\text{iso}}(w_0)$, where

$$\mathcal{Q}_{\text{iso}}(w_0) = \left\{ Q : \frac{\mathsf{d}Q}{\mathsf{d}P}(x) \text{ is a monotonically nondecreasing function of } w_0(x) \right\}.$$

If we assume as before that the target distribution P_{target} satisfies $D_{\text{KL}}(P_{\text{target}} \| P) \leq \rho$, then we can bound

$$\mathbb{E}_{P_{\text{target}}}[R(X)] \leq \mathbb{E}_P[R(X)] + \Delta(R; \mathcal{Q}_{\text{KL}}(\rho) \cap \mathcal{Q}_{\text{iso}}(w_0)). \tag{3}$$

The benefits of iso-DRL. What are the benefits of iso-DRL, as compared to the existing DRL framework? Of course, thus far the idea is quite straightforward—if we have stronger constraints on P_{target}, then we can place a tighter bound on the excess risk $\mathbb{E}_{P_{\text{target}}}[R(X)] - \mathbb{E}_P[R(X)]$. But as we will see below, adding the isotonic constraint plays a crucial role in enabling DRL to provide bounds that are useful in practical scenarios. Specifically, consider a practical setting where the bound ρ on the distribution shift is a positive constant. As we will see below, the existing worst-case excess risk $\Delta(R; \mathcal{Q}_{\mathrm{KL}}(\rho))$ of DRL is often quite large, leading to extremely conservative statistical conclusions; in contrast, the worst-case excess risk $\Delta(R; \mathcal{Q}_{\mathrm{KL}}(\rho) \cap \mathcal{Q}_{\mathrm{iso}}(w_0))$ given by iso-DRL is often vanishingly small, leading to much more informative conclusions. Moreover, surprisingly, this improvement in the bound does not incur any additional computational challenges—even though the constraint set $\mathcal{Q}_{\mathrm{KL}}(\rho) \cap \mathcal{Q}_{\mathrm{iso}}(w_0)$ appears more complex than the original set $\mathcal{Q}_{\mathrm{KL}}(\rho)$, we will see that $\Delta(R; \mathcal{Q}_{\mathrm{KL}}(\rho) \cap \mathcal{Q}_{\mathrm{iso}}(w_0))$ can be computed nearly *as easily as* the original quantity $\Delta(R; \mathcal{Q}_{\mathrm{KL}}(\rho))$. In addition, we further show in Appendix Section C.6 that the worst-case excess risk of iso-DRL can be consistently estimated with noisy observations of $R(X)$, while the estimation of the worst-case excess risk of DRL can be challenging even with bounded risks.

Empirical example: predictive inference for the `wine quality` dataset. To illustrate the advantage of the proposed approach, Figure 1 presents a numerical example for a predictive inference problem on the `wine quality` dataset (Cortez et al., 2009).[1] (See Section 5.2 for full details of this experiment.)

We are given a pretrained family of prediction bands $\widehat{C}_{1-a}$, indexed by the target coverage level $1-a$. At each value $a \in [0, 1]$, we define $R_a(X) = \mathbb{P}(Y \notin \widehat{C}_{1-a}(X) \mid X)$, the probability of the prediction band failing to cover the true response value Y given features X. Our goal is to return a prediction band with 90% coverage—that is, we would like to choose a value of a such that the expected risk $\mathbb{E}_{P_{\mathrm{target}}}[R_a(X)] = \mathbb{P}_{\widetilde{P}_{\mathrm{target}}}(Y \notin \widehat{C}_{1-a}(X))$ is bounded by $0.1 = 1 - 90\%$. In our experiment, the available data is given by all samples that are white wines (with distribution $\widetilde{P}$), while the target population is comprised of the samples that are red wines (with a different distribution $\widetilde{P}_{\mathrm{target}}$). In Figure 1, we compare four methods (see Section 5.2 for details):

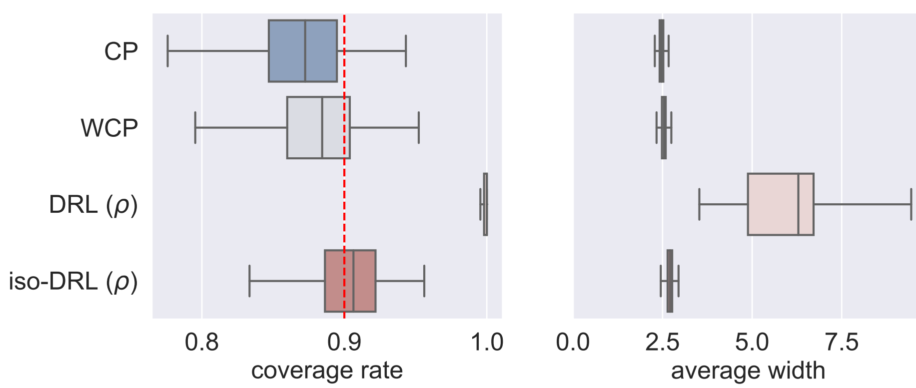


Figure 1: Coverage rate and average width of intervals for the `wine quality` dataset. The red dashed line (in the left-hand plot) marks the nominal coverage level, $1 - \alpha = 90\%$.

- An uncorrected interval—using conformal prediction (CP) (Vovk et al., 2005): the value a is chosen by tuning on the calibration data set (i.e., we choose a to satisfy $\mathbb{E}_P[R_a(X)] \le 0.1$), without correcting for the distribution shift.

[1] Available at https://archive.ics.uci.edu/dataset/186/wine+quality

- A corrected interval—using weighted conformal prediction (WCP) (Tibshirani et al., 2019): the value a is chosen by tuning on the calibration data set using an estimated density ratio w_0 to correct for the covariate shift between distributions $\widetilde{P}$ and $\widetilde{P}_{\text{target}}$. Since w_0 is estimated from data, this correction is imperfect.

- The DRL interval: we choose a to satisfy $\mathbb{E}_P[R_a(X)]+\Delta(R_a; \mathcal{Q}_{\text{KL}}(\rho)) \leq 0.1$, where $\mathbb{E}_P[R_a(X)]$ and $\Delta(R_a; \mathcal{Q}_{\text{KL}}(\rho))$ are estimated using the calibration data.

- The iso-DRL interval: we choose a to satisfy $\mathbb{E}_P[R_a(X)] + \Delta(R_a; \mathcal{Q}_{\text{KL}}(\rho) \cap \mathcal{Q}_{\text{iso}}(w_0)) \leq 0.1$, where $\mathbb{E}_P[R_a(X)]$ and $\Delta(R_a; \mathcal{Q}_{\text{KL}}(\rho) \cap \mathcal{Q}_{\text{iso}}(w_0))$ are estimated using the calibration data.[2]

As we can see in Figure 1, the CP and WCP intervals both undercover—for CP, this is because the method does not correct for distribution shift, while for WCP, this is because the ratio w_0 that corrects for distribution shift is imperfectly estimated. At the other extreme, DRL shows substantial overcoverage with extremely wide prediction intervals due to the worst-case nature of the bound $\Delta(R_a; \mathcal{Q}_{\text{KL}}(\rho))$. In contrast, our proposed method, iso-DRL, achieves the target coverage rate 90% without excessive increase in the size of the prediction interval, showing the benefit of adding the isotonic constraint to the DRL framework.

The motivating example demonstrates that, when we have access to meaningful—but imperfect—side information (e.g., in the form of the density ratio w_0), adding the isotonic constraint to iso-DRL can provide an estimate of the risk that is *more reliable* than a non-distributionally-robust approach, but *less conservative* than the original DRL approach.

1.4 Organization of paper

Section 2 introduces a general class of uncertainty sets for candidate distributions and further studies the property of the worst-case excess risk defined in (1) for generic DRL. For the worst-case excess risk with the isotonic constraint, we prove that it is equivalent to the worst-case excess risk for a projected risk function without the isotonic constraint in Section 3. In Section 4, we propose an estimator of the worst-case excess risk with the isotonic constraint and establish the estimation error bounds. Numerical results for both synthetic and real data are shown in Section 5 and additional related work is summarized in Section 6. We defer technical proofs and additional simulations to the Appendix.

Notation. Before proceeding, we introduce useful notation for theoretical developments later on. To begin with, we write $(a)_+$ as the positive part of $a \in \mathbb{R}$. We denote by $L_p(P)$ $(1 \leq p \leq \infty)$ the L_p function space under the probability measure P, i.e., when $p \neq \infty$, $L_p(P) = \{f : \|f\|_p = (\int_{\mathcal{X}} |f(x)|^p \mathsf{d}P(x))^{1/p} < \infty\}$. When $p = \infty$, the set $L_\infty(P)$ consists of measurable functions that are bounded almost surely under P. In addition, for a measurable function w defined on $\mathcal{X}$ and a measure P on $\mathcal{X}$, the pushforward measure $w_\# P$ denotes the measure satisfiying that $(w_\# P)(B) = P(w^{-1}(B))$ for any measurable set B, where $w^{-1}(B) = \{x \in \mathcal{X} : w(x) \in B\}$ denotes the preimage of B under w. In other words, if $X \sim P$, then $w(X)$ follows the distribution $w_\# P$. We say a function h is A_h-bounded if $\sup_x |h(x)| \leq A_h$. Fix a partial (pre)order $\preceq$ on $\mathcal{X} \subseteq \mathbb{R}^d$. A function g is isotonic

[2]For both the DRL and iso-DRL methods, the parameter ρ is an estimate of the actual KL distance $D_{\text{KL}}(P_{\text{target}} \| P)$—see Section 5.2 and Appendix D.2 for details

if $g(x_1) \leq g(x_2)$ for any $x_1 \preceq x_2$. Correspondingly, we define the cone of isotonic functions by $\mathcal{C}^{\text{iso}}_{\preceq} = \{w : w \text{ is isotonic w.r.t. partial order } \preceq\}$. Lastly, to compare two probability distributions Q and P, the convex ordering $\overset{cvx}{\preceq}$ is defined as $Q' \overset{cvx}{\preceq} Q$ if and only if $\mathbb{E}_{Q'}[\psi(X)] \leq \mathbb{E}_Q[\psi(X)]$ for all convex functions ψ.

2 The distributional robustness framework

As we have explained in Section 1.1, both the unsupervised and supervised setting under covariate shift can be unified. Therefore, throughout this section, to develop our theoretical results we will use the notation of the unsupervised setting with the risk function $R(X)$, with the understanding that this also covers the supervised setting under covariate shift.

Recall that $\mathcal{X}$ is the feature domain. We consider a bounded risk function $R : \mathcal{X} \to [0, B_R]$ with $0 < B_R < \infty$. The goal is to evaluate (or bound) the target risk $\mathbb{E}_{P_{\text{target}}}[R(X)]$ using samples from P, by assuming that the target distribution P_{target} is in some sense similar to the available distribution P—more concretely, by assuming that the target distribution P_{target} lies in some neighborhood $\mathcal{Q}$ around the distribution P of the available data.

Reformulating the neighborhood. To unify the different examples of constraints described in Section 1, we will start by considering settings where we can express the constraint $Q \in \mathcal{Q}$ using conditions on the density ratio $w = \mathsf{d}Q/\mathsf{d}P$. This type of framework includes the sensitivity analysis setting via bounds on w (Cornfield et al., 1959; Rosenbaum, 1987; Tan, 2006; Ding and VanderWeele, 2016; Zhao et al., 2019; Yadlowsky et al., 2018; Jin et al., 2022), and f-divergence constraints such as a bound on the KL divergence (Duchi et al., 2021; Namkoong and Duchi, 2017; Duchi and Namkoong, 2018; Cauchois et al., 2020).

Concretely, we can reparameterize the distribution Q using the density ratio $w(x) = (\mathsf{d}Q/\mathsf{d}P)(x)$. Then we can reformulate the constraint $Q \in \mathcal{Q}$ into a constraint on this density ratio, i.e., $Q \in \mathcal{Q} \iff w_\# P \in \mathcal{B}$, where $\mathcal{B}$ is a set of distributions, and where $w_\# P$ denotes the pushforward measure (as defined in Section 1.4). Let us now consider the two examples mentioned above.

Example 1: bound-constrained distribution shift. In sensitivity analysis, it is common to assume that the likelihood ratio $\mathsf{d}P_{\text{target}}/\mathsf{d}P$ is bounded from above and below. This corresponds to a constraint set of the form $\mathcal{Q} = \{Q : a \leq (\mathsf{d}Q/\mathsf{d}P)(X) \leq b \ P\text{-almost surely}\}$, for some constants $0 \leq a < 1 < b < +\infty$. In particular, when $a = \Gamma^{-1}$ and $b = \Gamma$ for some $\Gamma > 1$, this constraint set represents the marginal Γ-selection model for the density ratio in sensitivity analysis (Rosenbaum, 1987; Tan, 2006). By defining

$$\mathcal{B} = \mathcal{B}_{a,b} = \left\{ \widetilde{Q} : \ \mathbb{E}_{Z\sim\widetilde{Q}}[Z] = 1, \ \mathbb{P}_{Z\sim\widetilde{Q}}(a \leq Z \leq b) = 1 \right\},$$

we can verify that $Q \in \mathcal{Q} \iff w_\# P \in \mathcal{B}_{a,b}$ with $w(x) = (\mathsf{d}Q/\mathsf{d}P)(x)$.

Example 2: f-divergence constrained distribution shift. The f-divergence is a generalized way of measuring the distance between distributions, which includes common metrics such as KL divergence or chi-squared divergence as special cases. For a convex function $f : [0, \infty) \to \mathbb{R}$ satisfying

$f(1) = 0$, the f-divergence (Ali and Silvey, 1966; Rényi, 1961) of Q from P is defined as $D_f(Q||P) = \mathbb{E}_P[f((\mathsf{d}Q/\mathsf{d}P)(X))]$. In this example, we consider a constraint set $\mathcal{Q}$ defined via a bound on the f-divergence: $\mathcal{Q} = \{Q : D_f(Q||P) \le \rho\}$. For instance, if we take $\mathcal{Q} = \mathcal{Q}_{\mathrm{KL}}(\rho) = \{Q : D_{\mathrm{KL}}(Q\|P) \le \rho\}$, this corresponds to choosing $f(x) = x\log(x)$. Choosing

$$\mathcal{B} = \mathcal{B}_{f,\rho} = \{\widetilde{Q} :\ \mathbb{E}_{Z\sim\widetilde{Q}}[Z] = 1,\ \mathbb{E}_{Z\sim\widetilde{Q}}[f(Z)] \le \rho,\ \mathbb{P}_{Z\sim\widetilde{Q}}(Z \ge 0) = 1\},$$

we can verify that $Q \in \mathcal{Q} \iff w_{\#}P \in \mathcal{B}_{f,\rho}$ with $w(x) = (\mathsf{d}Q/\mathsf{d}P)(x)$.

2.1 Worst-case excess risk with DRL

In this section, we explore some properties of the generic DRL, without the isotonic constraint. Building this framework will help us to introduce the isotonic constraint as follows.

Based on the equivalence of $\mathcal{Q}$ and $\mathcal{B}$ in representing the uncertainty set, we focus on the following equivalent representation of $\Delta(R;\mathcal{Q})$:

$$\begin{aligned} \Delta(R;\mathcal{B}) \quad = \quad & \sup_{w\ge 0} \quad \mathbb{E}_P\left[w(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad w_{\#}P \in \mathcal{B}, \end{aligned} \tag{4}$$

where abusing notation we now write $\Delta(\cdot;\mathcal{B})$ to express that $\mathcal{B}$ is a constraint on the distribution of the density ratio $w(X) = (\mathsf{d}Q/\mathsf{d}P)(X)$, where previously we instead wrote $\Delta(\cdot;\mathcal{Q})$. We will say that $\Delta(R;\mathcal{B})$ is *attainable* if this supremum is attained by some w^* in the constraint set. Throughout the paper, we assume that the set of distributions $\mathcal{B}$ satisfies the following condition.

Condition 2.1. The set $\mathcal{B}$ contains the point mass on the value 1. Moreover, $\mathcal{B}$ is closed under convex ordering, that is, if $Q \in \mathcal{B}$, then for any $Q' \overset{cvx}{\preceq} Q$, it holds that $Q' \in \mathcal{B}$.

This condition enables the following reformulation of the quantity of interest, $\Delta(R;\mathcal{B})$:

Proposition 2.2. *Assume Condition 2.1 holds. Then $\Delta(R;\mathcal{B})$ can be written as*

$$\begin{aligned} \Delta(R;\mathcal{B}) \quad = \quad & \sup_{\phi:\mathbb{R}\to\mathbb{R}_+} \quad \mathbb{E}_P\left[(\phi\circ R)(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad (\phi\circ R)_{\#}P \in \mathcal{B}, \qquad \phi \textit{ is nondecreasing.} \end{aligned}$$

Moreover, if $\Delta(R;\mathcal{B})$ is attainable (i.e., the supremum is attained by some w^* satisfying the constraints), then this equivalent formulation is attainable as well (i.e., the supremum is attained by some ϕ^* satisfying the constraints), and it then holds that $w^*(X) = \phi^*(R(X))$ P-almost surely.

In words, this proposition shows that the excess risk is maximized by considering functions $w(x)$ that are monotonically nondecreasing with respect to $R(x)$. This is intuitive, since maximizing the expected value of $w(X)R(X)$ implies that we should choose a function w that is large when R is large. Most importantly, Proposition 2.2 implies that for a class of constraint sets $\mathcal{B}$, the optimal value in the constrained optimization problem (4) only depends on the distribution of X through the distribution of $R(X)$. We note that, in the special case when $\mathcal{B}$ is specified in terms of an f-divergence (as in Example 2 above), the conclusion of Proposition 2.2 is established by Donsker and Varadhan, 1976; Lam, 2016; Namkoong et al., 2022. We verify that this result holds in aforementioned settings as follows.

Returning to Example 1: bound-constrained distribution shift. Recall that in this example, we take the constraint set $\mathcal{B} = \mathcal{B}_{a,b} = \{\widetilde{Q} : \mathbb{E}_{Z\sim\widetilde{Q}}[Z] = 1, \mathbb{P}_{Z\sim\widetilde{Q}}(a \leq Z \leq b) = 1\}$, for some $0 \leq a < 1 < b < +\infty$. It is straightforward to verify that $\mathcal{B}_{a,b}$ satisfies Condition 2.1, implying that Proposition 2.2 can be applied.

Moreover, we can actually calculate the maximizing density ratio $w^*(x)$ explicitly. The worst-case density ratio that attains the worst-case excess risk takes the form

$$w^*(x) = a \cdot \mathbb{1}\left\{R(x) < q_R\left(\tfrac{b-1}{b-a}\right)\right\} + b \cdot \mathbb{1}\left\{R(x) > q_R\left(\tfrac{b-1}{b-a}\right)\right\} + c \cdot \mathbb{1}\left\{R(x) = q_R\left(\tfrac{b-1}{b-a}\right)\right\},$$

where $q_R(t)$ is the t-quantile of the distribution of $R(X)$ under $X \sim P$ and $c \in [a, b]$ is defined as the unique value ensuring that $\mathbb{E}[w^*(X)] = 1$, namely,

$$c = a + \frac{(b-a)t^* - (b-1)}{\mathbb{P}\left\{R(X) = q_R\left(\tfrac{b-1}{b-a}\right)\right\}} \qquad \text{with} \qquad t^* = \mathbb{P}\left\{R(X) \leq q_R\left(\tfrac{b-1}{b-a}\right)\right\} \geq \frac{b-1}{b-a}.$$

We can see that $w^*(x)$ is nondecreasing in $R(x)$, validating the conclusion of Proposition 2.2.

Returning to Example 2: f-divergence constrained distribution shift. Recall that for an f-divergence constraint, we define $\mathcal{B} = \mathcal{B}_{f,\rho} = \{\widetilde{Q} : \mathbb{E}_{Z\sim\widetilde{Q}}[Z] = 1, \mathbb{E}_{Z\sim\widetilde{Q}}[f(Z)] \leq \rho, Z \geq 0\}$. Since f is convex, this immediately implies that $\mathcal{B}_{f,\rho}$ satisfies Condition 2.1. If we further assume that f is differentiable, by the results of Shapiro (2017); Donsker and Varadhan (1976); Lam (2016), the worst-case excess risk $\Delta_\rho(R; \mathcal{B}_{f,\rho})$ is attained at

$$w^*(x) = w(x; \lambda^*, \nu^*) = \left\{(f')^{-1}\left(\frac{R(x) - \nu^*}{\lambda^*}\right)\right\}_+,$$

where λ^*, ν^* are the solutions to the dual problem

$$\inf_{\lambda\geq 0,\, \nu}\left\{\lambda\rho + \nu + \mathbb{E}_P\Big[w(X;\lambda,\nu)(R(X) - \nu) - \lambda f(w(X;\lambda,\nu))\Big]\right\}. \tag{5}$$

Since f is convex, its inverse derivative $(f')^{-1}$ is nondecreasing, meaning that $w^*(x)$ is nondecreasing in $R(x)$, which again validates the result in Proposition 2.2.

3 Worst-case excess risk with an isotonic constraint

In this section, we will now formally introduce our iso-DRL method, adding an isotonic constraint to the DRL framework developed in Section 2 above. As in Section 2, throughout this section we use the notation of the unsupervised learning setting with risk $R(X)$, since the supervised case can also be reduced to this setting.

Recall the cone of isotonic functions $\mathcal{C}^{\text{iso}}_{\preceq} = \{w : \mathcal{X} \to \mathbb{R} : w \text{ is isotonic w.r.t. } \preceq\}$. In this paper, we actually allow $\preceq$ to be a partial *preorder* rather than a partial order, meaning that it may be the case that both $x \preceq x'$ and $x' \preceq x$, even when $x \neq x'$. As an example, we denote $\mathcal{C}^{\text{iso}}_{w_0} = \{w : w(x) \text{ is a monotonically nondecreasing function of } w_0(x)\}$—this is obtained by the (pre)order given by $x \preceq x'$ whenever $w_0(x) \leq w_0(x')$.

Our focus is the worst-case excess risk with the isotonic constraint:

$$\begin{aligned} \Delta^{\text{iso}}(R;\mathcal{B}) \quad &= \quad \sup_{w\geq 0} \quad \mathbb{E}_P\left[w(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ &\quad\ \text{subject to} \quad w_{\#}P \in \mathcal{B}, \quad w \in \mathcal{C}^{\text{iso}}_{\preceq}. \end{aligned} \tag{6}$$

To make this more concrete with a specific example, in the bound (3), this example corresponds to choosing $\mathcal{B} = \mathcal{B}_{f,\rho}$ for the f-divergence $f(x) = x\log x$, as for the KL distance constraint. In particular, the bound (3) assumed two constraints on the distribution P_{target}—first, $D_{\text{KL}}(P_{\text{target}}\|P) \leq \rho$ (which corresponds to assuming $(\mathsf{d}P_{\text{target}}/\mathsf{d}P)_{\#}P \in \mathcal{B}_{f,\rho}$, in our new notation), and second, $P_{\text{target}} \in \mathcal{Q}_{\text{iso}}(w_0)$ (which is expressed by assuming $w \in \mathcal{C}^{\text{iso}}_{\preceq}$ when we take the partial (pre)order defined as $x \preceq x'$ whenever $w_0(x) \leq w_0(x')$—or equivalently, we can write this as $w \in \mathcal{C}^{\text{iso}}_{w_0}$).

3.1 Equivalent formulation

Optimization problems with isotonic constraints may be difficult to tackle both theoretically and computationally, since the isotonic cone, despite being convex, may be challenging to optimize over when working with an infinite-dimensional object such as the density ratio. In this section, we will show that the maximization problem (6) can equivalently be reformulated as an optimization problem *without* an isotonic constraint, by drawing a connection to the original (not isotonic) DRL maximization problem (4).

Given the probability measure P, we will define π as the projection to the isotonic cone $\mathcal{C}^{\text{iso}}_{\preceq}$ with respect to $L_2(P)$, i.e., $\pi(a) = \operatorname*{argmin}_{b\in\mathcal{C}^{\text{iso}}_{\preceq}} \int (a(x) - b(x))^2 \mathsf{d}P(x)$. As $L_2(P)$ is reflexive and strictly convex, the projection $\pi(a)$ exists and is unique (up to sets of measure zero) for all $a \in L_2(P)$ (Megginson, 2012). Then, with the projection π in place, we are ready to state our main equivalence result.

Theorem 3.1. *For any $\mathcal{B}$ and any partial (pre)order $\preceq$ on $\mathcal{X}$, it holds that $\Delta^{\text{iso}}(R;\mathcal{B}) \leq \Delta(\pi(R);\mathcal{B})$. If in addition Condition 2.1 holds, then we have*

$$\Delta^{\text{iso}}(R;\mathcal{B}) = \Delta(\pi(R);\mathcal{B}),$$

and moreover, $\Delta^{\text{iso}}(R;\mathcal{B})$ is attainable if and only if $\Delta(\pi(R);\mathcal{B})$ is attainable.

To interpret this theorem, recall from the definition (4) that we have

$$\begin{aligned} \Delta(\pi(R);\mathcal{B}) \quad &= \quad \sup_{w\geq 0} \ \mathbb{E}_P\big[w(X)[\pi(R)](X)\big] - \mathbb{E}_P\big[[\pi(R)](X)\big] \\ &\quad\ \text{subject to} \quad w_{\#}P \in \mathcal{B}. \end{aligned} \tag{7}$$

Compared with the formulation (6) that defines the isotonic worst-case risk $\Delta^{\text{iso}}(R;\mathcal{B})$, we see that this equivalent formulation removes the constraint $w \in \mathcal{C}^{\text{iso}}_{\preceq}$ by replacing R with its isotonic projection $\pi(R)$. This brings computational benefits. The equivalent formulation (7) separates two constraints $w_{\#}P \in \mathcal{B}$ and $w \in \mathcal{C}^{\text{iso}}_{\preceq}$, allowing us to first project the risk function R onto $\mathcal{C}^{\text{iso}}_{\preceq}$, and then solve a problem that is as simple as the problem stated earlier in (4). More concretely, as seen in Examples 1 and 2, for many common choices of $\mathcal{B}$, we have closed-form solutions to (7) in terms of the projected risk $\pi(R)$.

3.2 Setting: iso-DRL with estimated density ratio

We now return to the scenario described in (3) in Section 1.3, where we would like to recalibrate a pretrained density ratio w_0 that estimates $\mathsf{d}P_{\text{target}}/\mathsf{d}P$. As the shape or relative magnitude of w_0 could contain useful information about the true density ratio, we consider candidate distributions with the density ratio as an isotonic function of w_0, which is equivalent to considering the partial (pre)order $x \preceq x' \iff w_0(x) \leq w_0(x')$. We will denote the specific isotonic cone under this partial (pre)order as $\mathcal{C}^{\text{iso}}_{w_0}$ and its isotonic projection as π_{w_0}, and abusing notation, we write $\Delta^{\text{iso}}(R; \mathcal{B}, w_0)$ to denote the excess risk for this particular setting, to emphasize the role of w_0. Under the condition of Theorem 3.1, we have the equivalence $\Delta^{\text{iso}}(R; \mathcal{B}, w_0) = \Delta(\pi_{w_0}(R); \mathcal{B})$. To understand the projection onto the cone $\mathcal{C}^{\text{iso}}_{w_0}$ in a more straightforward way, we can derive a further simplification. Write π_1 to denote the isotonic projection of functions $\mathbb{R} \to \mathbb{R}$ under the measure $(w_0)_\# P$, and define a function $\widetilde{R} : \mathbb{R} \to \mathbb{R}$ to satisfy $\widetilde{R}(w_0(X)) = \mathbb{E}_P[R(X) \mid w_0(X)]$ P-almost surely. We then have the following simplified equivalence:

Proposition 3.2. *Assume Condition 2.1 holds. We have the equivalence* $\Delta^{\text{iso}}(R; \mathcal{B}, w_0) = \Delta(\pi_1(\widetilde{R}) \circ w_0; \mathcal{B}, w_0)$*, where we define*

$$\begin{aligned} \Delta(R; \mathcal{B}, w_0) \quad = \quad & \sup_{h:\ h \circ w_0 \geq 0} \quad \mathbb{E}_P\left[(h \circ w_0)(X) R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad (h \circ w_0)_\# P \in \mathcal{B}. \end{aligned} \tag{8}$$

In comparison to the equivalence $\Delta^{\text{iso}}(R; \mathcal{B}, w_0) = \Delta(\pi_{w_0}(R); \mathcal{B})$, the equivalence in the proposition relies on an isotonic projection with respect to the canonical order on the real line (i.e., the projection π_1).

Moreover, when the true distribution shift does not obey the isotonic constraint exactly, in Appendix Section B.4, we can nonetheless provide a bound on the worst-case excess risk, which is tighter than the (non-iso) DRL bound whenever the isotonic constraint provides a reasonable approximation.

4 Estimation of worst-case excess risk with isotonic constraint

So far, our focus has been on the population level problem, namely, we have assumed full access to the data distribution P and the risk function R. In practice, however, we may only be able to access the data distribution P via samples, and we may only be able to learn about the risk function R via noisy evaluations of $R(X)$ on each sampled point X in the unsupervised setting. Or, in the supervised setting, we can only access $\widetilde{P}$ via samples of labeled data points drawn from this distribution, and can learn about r only through evaluating $r(X, Y)$ on these sampled data points.

In this section, we propose a fully data dependent estimator for the worst-case excess risk $\Delta^{\text{iso}}(R; \mathcal{B})$. Moreover, we characterize the estimation error for different choices of $\mathcal{B}$, including the bounds constraint and the f-divergence constraint for the distribution shift.

4.1 Plug-in estimators

We start with presenting the plug-in estimators of the worst-case excess risk under the isotonic constraint in both the unsupervised and supervised settings.

The unsupervised setting. We have n i.i.d. observations $\{X_i\}_{i\leq n}$ from a distribution P. Given a risk function $R:\mathcal{X}\to\mathbb{R}_+$, and the uncertainty set $\mathcal{B}$, we estimate the worst-case excess risk $\Delta^{\mathrm{iso}}(R;\mathcal{B})$ (cf. Equation (6)) via

$$\begin{aligned}\widehat{\Delta}^{\mathrm{iso}}(R;\mathcal{B}) \quad := \quad & \max_{w\geq 0} \quad \frac{1}{n}\sum_{i\leq n} w(X_i)R(X_i) - \frac{1}{n}\sum_{i\leq n} R(X_i) \\ & \text{subject to} \quad w_{\#}\widehat{P}_n \in \mathcal{B}, \quad w \in \mathcal{C}^{\mathrm{iso}}_{\preceq}. \end{aligned} \tag{9}$$

Here, $\widehat{P}_n$ denotes the empirical distribution of the sample $\{X_i\}_{i\leq n}$ drawn from P.

The supervised setting. In this case, we have $\{(X_i, Y_i)\}_{i\leq n}$ drawn i.i.d. from $\widetilde{P} = P \times P_{Y|X}$. Given a risk function r, and the uncertainty set $\mathcal{B}$, we propose to estimate the worst-case excess risk $\widehat{\Delta}^{\mathrm{iso}}(r;\mathcal{B})$ by replacing $R(X_i)$ with $r(X_i, Y_i)$ in (9).

4.1.1 Adding a boundedness constraint

When calculating the excess risk at the population level, the constraint set $\mathcal{B}$ may not require w to be bounded—specifically, while $\mathcal{B}_{a,b}$ imposes an upper bound on w, the f-divergence constraint set $\mathcal{B}_{f,\rho}$ does not. In the empirical setting, however, a boundedness constraint is more crucial: we want to avoid degenerate scenarios, such as $w(X_i)$ taking an arbitrarily large value for a single i, and being zero for the remaining $n-1$ data points. To this end, we will assume from this point on that $\mathcal{B}$ includes a boundedness constraint:

Condition 4.1. There exists Ω such that any distribution $Q \in \mathcal{B}$ is supported on $[0, \Omega]$.

This is trivially true for $\mathcal{B} = \mathcal{B}_{a,b}$ with $\Omega = b$. But this constraint actually allows us to work with the f-divergence example, as well, as established by the following result.

Proposition 4.2. *Assume the convex function f is differentiable on $\mathbb{R}_+$. The worst-case excess risk $\Delta^{\mathrm{iso}}(R;\mathcal{B}_{f,\rho})$ is attained at some $w^{*\mathrm{iso}}_{f,\rho} \in \mathcal{C}^{\mathrm{iso}}_{\preceq}$ with $\|w^{*\mathrm{iso}}_{f,\rho}\|_\infty < \infty$.*

In particular, defining $\mathcal{B}_{f,\rho,\Omega} = \{\widetilde{Q} :\ \mathbb{E}_{Z\sim\widetilde{Q}}[Z] = 1,\ \mathbb{E}_{Z\sim\widetilde{Q}}[f(Z)] \leq \rho,\ \mathbb{P}_{Z\sim\widetilde{Q}}(0 \leq Z \leq \Omega) = 1\}$, which adds a boundedness requirement in addition to the f-divergence constraint, we can see that for sufficiently large Ω (namely, $\Omega \geq \|w^{*\mathrm{iso}}_{f,\rho}\|_\infty$), even though $\mathcal{B}_{f,\rho,\Omega} \subsetneq \mathcal{B}_{f,\rho}$, it nonetheless holds that $\Delta^{\mathrm{iso}}(R;\mathcal{B}_{f,\rho,\Omega}) = \Delta^{\mathrm{iso}}(R;\mathcal{B}_{f,\rho})$. Therefore, by working with the constraint set $\mathcal{B}_{f,\rho,\Omega}$, we are estimating the *same* excess risk, but Condition 4.1 nonetheless holds. (Of course, in practice, the value of $\|w^{*\mathrm{iso}}_{f,\rho}\|_\infty$ is unknown and so we can simply set Ω to be a large constant.)

4.2 Computation: estimation after projection

Before moving onto the statistical performance of the two estimators $\widehat{\Delta}^{\text{iso}}(R;\mathcal{B})$ and $\widehat{\Delta}^{\text{iso}}(r;\mathcal{B})$, we pause to discuss fast computational methods for these. The key is Theorem 3.1—we may accelerate the computation of both estimators via an equivalent optimization problem without the isotonic constraint.

To be more specific, we begin by considering the supervised setting. Denote $r^{\text{iso}} = (r_i^{\text{iso}})_{i\leq n} \in \mathbb{R}^n$ as the isotonic projection of $(r(X_i,Y_i))_{i\leq n}$ with respect to the empirical distribution $\widehat{P}_n$ under the partial order $\preceq$. Then, consider the optimization problem

$$\begin{aligned} \widehat{\Delta}(r^{\text{iso}};\mathcal{B}) \quad &:= \quad \max_{w\geq 0} \quad \frac{1}{n}\sum_{i\leq n} w(X_i) r_i^{\text{iso}} - \frac{1}{n}\sum_{i\leq n} r_i^{\text{iso}} \\ &\quad\quad \text{subject to} \quad w_{\#}\widehat{P}_n \in \mathcal{B}. \end{aligned} \tag{10}$$

By Theorem 3.1 (applied with $\widehat{P}_n$ in place of P), we have $\widehat{\Delta}(r^{\text{iso}};\mathcal{B}) = \widehat{\Delta}^{\text{iso}}(r;\mathcal{B})$. Analogously, in the unsupervised setting, we instead have $\widehat{\Delta}^{\text{iso}}(R;\mathcal{B}) = \widehat{\Delta}(R^{\text{iso}};\mathcal{B})$, where $R^{\text{iso}} = (R_i^{\text{iso}})_{i\leq n} \in \mathbb{R}^n$ as the isotonic projection of $(R(X_i))_{i\leq n}$ with respect to the empirical distribution $\widehat{P}_n$ under the partial order $\preceq$.

Note that in iso-DRL with estimated density ratio in Section 3.2, we can simply apply the isotonic regression for $(r(X_i,Y_i))_{i\leq n}$ on $(w_0(X_i))_{i\leq n}$ to obtain the projected risk. We can now see concretely that this equivalence allows for a much more efficient calculation. For example, in the case $\mathcal{X} = \mathbb{R}$, this isotonic projection can be computed in $\mathcal{O}(n)$ time (e.g., via the PAVA algorithm, which provides an exact calculation of isotonic projection in $\mathbb{R}^n$ (Grotzinger and Witzgall, 1984)), this leads to a very simple implementation for computing $\widehat{\Delta}(r^{\text{iso}};\mathcal{B})$—in particular, once the vector r^{iso} has been computed, the remaining optimization problem is simple since there is no remaining isotonic constraint.

4.3 Performance guarantees for plug-in estimators

In this section, we present the performance guarantees for plug-in estimators for a general constraint set $\mathcal{B}$. To jump ahead to the conclusion, we will see that our results imply the following consistency properties for the settings $\mathcal{B} = \mathcal{B}_{a,b}$ and $\mathcal{B} = \mathcal{B}_{f,\rho,\Omega}$. Here we define the Rademacher complexity of a function class $\mathcal{G}$ by $\mathcal{R}_n(\mathcal{G}) = \mathbb{E}[\sup_{g\in\mathcal{G}} |n^{-1}\sum_{i\leq n}\sigma_i g(Z_i)|]$, where $\{Z_i\}_{i\leq n}$ is a sample of size n from P and $\{\sigma_i\}_{i\leq n}$ are independent random variables drawn uniformly from $\{+1,-1\}$.

Proposition 4.3 (Informal result for examples)**.** *For both $\mathcal{B} = \mathcal{B}_{a,b}$ and $\mathcal{B} = \mathcal{B}_{f,\rho,\Omega}$, and for both supervised ($\widehat{\Delta}^{\text{iso}}(\mathcal{B}) = \widehat{\Delta}^{\text{iso}}(r;\mathcal{B})$) and unsupervised ($\widehat{\Delta}^{\text{iso}}(\mathcal{B}) = \widehat{\Delta}^{\text{iso}}(R;\mathcal{B})$) learning, and under some mild additional conditions specified below, it holds with probability $\geq 1 - 3n^{-1}$ that*

$$\left|\widehat{\Delta}^{\text{iso}}(\mathcal{B}) - \Delta^{\text{iso}}(R;\mathcal{B})\right| \leq C\left(\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}) + \sqrt{\frac{\log n}{n}}\right),$$

where $\mathcal{C}^{\text{iso}}_{\preceq,\Omega} = \{w \in \mathcal{C}^{\text{iso}}_{\preceq} : \ 0 \leq w \leq \Omega\}$ is the bounded isotonic cone, where the constant C will be defined in the theorems below, and where we set $\Omega = b$ for the case $\mathcal{B} = \mathcal{B}_{a,b}$.

Our bounds rely on the Rademacher complexity term $\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega})$, which will naturally depend on the properties of the (pre)ordering $\preceq$ that defines this isotonic cone. To provide further intuition, we now give two concrete examples to make apparent the dependence of the Rademacher complexity on the sample size.

(1) When $d = 1$, e.g., in the setting of density ratio recalibration in Section 3.2, similar to the results of Chatterjee and Lafferty (2019), one can show by Dudley's theorem (Dudley, 1967) that $\mathcal{R}_n\left(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}\right) \lesssim n^{-1/2}$ up to logarithmic factors.

(2) For $\mathbb{R}^d$ with a fixed dimension $d \geq 2$ and a bounded domain $\mathcal{X}$ equipped with the componentwise order (Han et al., 2019; Deng and Zhang, 2020; Gao and Wellner, 2007), i.e., $x \preceq z$ if and only if $x_j \leq z_j$ for all $j \in [d]$, by Han et al. (2019), if the density $\mathsf{d}P(x)$ is bounded below (away from zero) and above, then we have $\mathcal{R}_n\left(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}\right) \lesssim n^{-1/d}$ up to logarithmic factors.

4.3.1 Formal results

Now we turn to developing these results formally, in a general framework. We will begin with a deterministic result, which shows that, if certain concentration inequalities hold, then $\widehat{\Delta}^{\text{iso}}(\mathcal{B}_{a,b})$ is an accurate estimate of $\Delta^{\text{iso}}(R;\mathcal{B}_{a,b})$. Then we will show that the concentration results hold with high probability, in both of our two settings, $\mathcal{B} = \mathcal{B}_{a,b}$ and $\mathcal{B} = \mathcal{B}_{f,\rho,\Omega}$.

We first need a few definitions. For any distributions P_0, P_1, if $w_\# P_0 \in \mathcal{B}$, we define

$$\varepsilon_{\mathcal{B}}\left(w; P_0, P_1\right) = \inf\left\{s \geq 0 \ : \ \exists\, t \geq 0,\ \left((1-s)\cdot w + t\cdot \mathbf{1}\right)_\# P_1 \in \mathcal{B}\right\}.$$

In other words, if weight function w satisfies the constraints relative to distribution P_0, we need to find constants s, t such that the modified weight function $(1-s)\cdot w + t \cdot \mathbf{1}$ satisfies the constraints relative to distribution P_1. (Note that we must have $\varepsilon_{\mathcal{B}}\left(w; P_0, P_1\right) \leq 1$, since choosing $s = t = 1$ will always be feasible, because $\mathbf{1}_\# P_1$ is the point mass on the value 1, and therefore satisfies the constraints of $\mathcal{B}$, by assumption.) Of importance is the quantity

$$\varepsilon_{\mathcal{B}} = \sup_{w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}} \max\left\{\varepsilon_{\mathcal{B}}\left(w; P, \widehat{P}_n\right), \varepsilon_{\mathcal{B}}\left(w; \widehat{P}_n, P\right)\right\},$$

which characterizes the feasibility gap between the population and sample problems. In addition, we define

$$\varepsilon_R = \sup_{w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}} \left|\mathbb{E}_{\widehat{P}_n}[(w(X)-1)r(X,Y)] - \mathbb{E}_P[(w(X)-1)R(X)]\right|$$

in the case of unsupervised learning, or

$$\varepsilon_R = \sup_{w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}} \left|\mathbb{E}_{\widehat{P}_n}[(w(X)-1)R(X)] - \mathbb{E}_P[(w(X)-1)R(X)]\right|$$

in the case of supervised learning. The value of ε_R measures the concentration between the empirical risk and the population one. With these definitions in place, we are ready to state the generic performance guarantee of the plug-in estimators.

Theorem 4.4. *Suppose that the risk is B_R-bounded (i.e., R or r, in the unsupervised or supervised case, respectively), and that the constraint set $\mathcal{B}$ satisfies Condition 4.1. Then, it holds for both supervised ($\widehat{\Delta}^{\text{iso}}(\mathcal{B}) = \widehat{\Delta}^{\text{iso}}(r;\mathcal{B})$) and unsupervised ($\widehat{\Delta}^{\text{iso}}(\mathcal{B}) = \widehat{\Delta}^{\text{iso}}(R;\mathcal{B})$) learning that*

$$\left|\widehat{\Delta}^{\text{iso}}(\mathcal{B}) - \Delta^{\text{iso}}(R;\mathcal{B})\right| \leq \varepsilon_R + 2B_R\Omega\cdot\varepsilon_{\mathcal{B}}.$$

Of course, in order for this result to be meaningful, we need to ensure that ε_R and $\varepsilon_{\mathcal{B}}$ are likely to be small, with high probability, We now turn to the question of establishing such concentration results. First we bound ε_R.

Lemma 4.5. *Suppose that the risk is B_R-bounded (i.e., R or r, in the unsupervised or supervised case, respectively). Then, with probability at least $1-n^{-1}$, it holds that $\varepsilon_R \leq 4B_R\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}) + B_R\Omega\sqrt{\log n/(2n)}$.*

Next we turn to bounding $\varepsilon_{\mathcal{B}}$, which we will do separately for our two examples.

Lemma 4.6. *Let $\mathcal{B} = \mathcal{B}_{a,b}$, where $a < 1 < b$. Then, with probability at least $1-n^{-1}$, it holds that $\varepsilon_{\mathcal{B}} \leq C(\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}) + \Omega\sqrt{\log n/(2n)})$, where we take $\Omega = b$ and C depends only on a, b.*

Finally, to complete this section, we turn to the f-divergence constraint, $\mathcal{B}_{f,\rho}$.

Lemma 4.7. *Let $\mathcal{B} = \mathcal{B}_{f,\rho,\Omega}$, where we take any $\Omega \geq \|w^{*\text{iso}}_{f,\rho}\|_\infty$ for $w^{*\text{iso}}_{f,\rho}$ defined as in Proposition 4.2. Assume also that f is L_Ω-Lipschitz on $[0,\Omega]$. Then, with probability at least $1-2n^{-1}$, it holds that $\varepsilon_{\mathcal{B}} \leq C(\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}) + \sqrt{\log n/(2n)})$, where C depends only on Ω, L_Ω, and ρ.*

4.4 The role of the isotonic constraint

The consistency bounds developed above show that, under appropriate conditions, the error in estimating $\Delta^{\text{iso}}(R;\mathcal{B})$ can be controlled whenever the appropriate Rademacher complexity terms are small. In the Appendix Section C.6, we will construct an example with $\mathcal{B} = \mathcal{B}_{a,b}$ such that, without an isotonic constraint, $\Delta(R;\mathcal{B}_{a,b}) = 0$ but $\widehat{\Delta}(r;\mathcal{B}_{a,b}) > c$ with high probability, where $c > 0$ doesn't vanish with n, i.e., this empirical estimate is *not* a consistent estimator of the true excess risk. This suggests that the isotonic constraint plays an important role: essentially, the isotonic constraint induces a form of regularization, ensuring that we work with a low-complexity class of functions.

5 Numerical experiments

In this section, we demonstrate the benefits of iso-DRL in calibrating prediction sets under covariate shift with empirical examples, as previewed in Section 1.3. Throughout all experiments, we have a training data set $\mathcal{D}_{\text{train}}$ containing data points (X_i, Y_i) drawn from the data distribution $\widetilde{P}$, and a test set $\mathcal{D}_{\text{test}}$ containing data points $(\widetilde{X}_i, \widetilde{Y}_i)$ drawn from $\widetilde{P}_{\text{target}}$. We consider both synthetic and real datasets. Code to reproduce all experiments is available at https://github.com/yugjerry/iso-DRL.

Background. When covariate shift is present, Tibshirani et al. (2019) proposes the weighted conformal prediction (WCP) method, which produces a prediction set $C^{w_0}_{1-\alpha}(X)$ with an estimated density ratio w_0, which is only valid for the shifted covariate distribution $\widehat{P}$ defined by $\mathsf{d}\widehat{P} \propto w_0 \cdot \mathsf{d}P$. The validity for the target distribution $\widetilde{P}_{\text{target}}$ is only guaranteed up to a coverage gap due to the estimation error or potential misspecification in w_0 (Lei and Candès, 2020; Candès et al., 2023; Gui et al., 2024, 2023).

Dataset partition. The datasets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ are partitioned as follows: (1) first, we use a subset $\mathcal{D}_1 \subseteq \mathcal{D}_{\text{train}}$ of the training data of size $|\mathcal{D}_1| = n_{\text{pre}}$, and a subset $\mathcal{D}_{\text{test},1} \subseteq \mathcal{D}_{\text{test}}$ of the test data of size $|\mathcal{D}_{\text{test},1}| = n_{\text{pre}}$, to train the function w_0; (2) next, we use a subset $\mathcal{D}_2 \subseteq \mathcal{D}_{\text{train}} \backslash \mathcal{D}_1$ of the training data of size $|\mathcal{D}_2| = n_{\text{train}}$ to train CP or WCP prediction intervals; (3) then, $\mathcal{D}_3 = \mathcal{D}_{\text{train}} \backslash (\mathcal{D}_1 \cup \mathcal{D}_2)$ is used to for estimating upper bounds on the excess risk for the DRL and iso-DRL methods. We further define $n = |\mathcal{D}_3|$ to ease notations; (4) finally, $\mathcal{D}_{\text{test},0} = \mathcal{D}_{\text{test}} \backslash \mathcal{D}_{\text{test},1}$ is used for estimating the actual performance of each method relative to the target distribution: defining $n_{\text{test}} = |\mathcal{D}_{\text{test},0}|$, we compute $\texttt{Coverage rate}(C, \alpha) = n_{\text{test}}^{-1} \sum_{i \in \mathcal{D}_{\text{test},0}} \mathbb{1}\{\widetilde{Y}_i \in C(\widetilde{X}_i)\}$. We next turn to the details of how these steps are carried out.

Initial density ratio estimation. Using data from $\mathcal{D}_1$ and $\mathcal{D}_{\text{test},1}$, we construct a data set comprised of the covariate X and a binary label $L \in \{0, 1\}$ (0 for the training points, 1 for the test points). We then fit a logistic regression model and obtain the estimated probability $\widehat{p}(x)$ for $\mathbb{P}(L = 1 \mid X = x)$, with which we define $w_0(x) = \widehat{p}(x)/(1 - \widehat{p}(x))$.

Split conformal prediction and weighted split conformal prediction. With data from $\mathcal{D}_2$, we use Ordinary Least Squares (OLS) as the base algorithm, where we denote $\widehat{\mu}$ as the fitted regression model, and apply split conformal prediction with the nonconformity score $V(x, y) = |y - \widehat{\mu}(x)|$ to obtain the following prediction intervals for comparison: (1) CP: conformal prediction interval $C_{1-\alpha}$ without adjusting for covariate shift; (2) WCP-oracle: weighted conformal prediction interval $C^{w^*}_{1-\alpha}$ with true density ratio $w^* = \mathsf{d}P_{\text{target}}/\mathsf{d}P$; (3) WCP: weighted conformal prediction interval $C^{w_0}_{1-\alpha}$ with estimated density ratio w_0.

DRL methods: estimation of worst-case excess risks. We then consider two distributionally robust methods. Using the subset $\mathcal{D}_3$ of the training data, the observed risks can be calculated by $r_i = \mathbb{1}\{Y_i \notin C_{1-\alpha}(X_i)\}$, $i \in \mathcal{D}_3$. We adopt the KL divergence constraint $D_{\text{KL}}(Q \| P) \leq \rho$ to measure the magnitude of distribution shift. Then, we obtain

$$\widehat{\Delta}(\alpha) = \max_{\|w\|_\infty \leq \Omega} \quad \frac{1}{n} \sum_{i \in \mathcal{D}_3} (w_i - 1) r_i \quad \text{s.t.} \quad \frac{1}{n} \sum_{i \in \mathcal{D}_3} w_i = 1, \ \frac{1}{n} \sum_{i \in \mathcal{D}_3} w_i \log w_i \leq \rho, \tag{11}$$

with the upper bound set as $\Omega = 100$ throughout the experiments. Next, given the estimated density ratio w_0, we run isotonic regression for $(r_i)_{i \leq n}$ on $(w_0(X^{(3)}_i))_{i \leq n}$ to obtain the projected risk $(r^{\text{iso}}_i)_{i \in \mathcal{D}_3}$, with which we can calculate the worst-case excess risk

$$\widehat{\Delta}^{\text{iso}}(\alpha) = \max_{\|w\|_\infty \leq \Omega} \quad \frac{1}{n} \sum_{i \in \mathcal{D}_3} (w_i - 1) r^{\text{iso}}_i \quad \text{s.t.} \quad \frac{1}{n} \sum_{i \in \mathcal{D}_3} w_i = 1, \ \frac{1}{n} \sum_{i \in \mathcal{D}_3} w_i \log w_i \leq \rho. \tag{12}$$

Given these estimates of the worst-case excess risks, we compare the following methods: (1) DRL: CP interval $C_{1-\widetilde{\alpha}}$, where $\widetilde{\alpha} = \max\{0, \alpha - \widehat{\Delta}(\alpha)\}$.[3]; (2) iso-DRL-$w_0$: CP interval $C_{1-\alpha_{\text{iso}}}$, where $\alpha_{\text{iso}} = \max\{0, \alpha - \widehat{\Delta}^{\text{iso}}(\alpha)\}$.

5.1 Synthetic dataset

We start with a synthetic example, in which we fix $n_{\text{train}} = n = n_{\text{test}} = 500$ and will vary n_{pre} to see how will the initial density ratio estimation w_0 affect the result. We will consider two settings—the "well-specified" and "misspecified" settings. Specifically, for the marginal distributions of X, we set the well-specified setting with $P : X \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and $P_{\text{target}} : X \sim \mathcal{N}(\mu, \mathbf{I}_d)$, and the misspecified setting with $P : X \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and $P_{\text{target}} : X \sim \mathcal{N}(\mu, \mathbf{I}_d + \frac{\zeta}{d}\mathbf{1}_d\mathbf{1}_d^\top)$, where $d = 20$, $\mu = (2/\sqrt{d}) \cdot (1, \cdots, 1)^\top$, and $\zeta = 6$. Since the estimate w_0 for the density ratio will be fitted via logistic regression as described above, the first setting is indeed well-specified since, due to the fact that P and P_{target} have the same covariance, the logistic model is correct for the distribution shift from P to P_{target}. In contrast, the second setting is misspecified since, due to the change in covariance matrix, the underlying log-density ratio is no longer a linear function of $\mu^\top X$, and therefore cannot be characterized exactly by a logistic regression model. Finally, for the conditional distribution of $Y \mid X$, we set $Y \mid X \sim 0.2 \cdot \mathcal{N}(X^\top\beta + \sin(X_1) + 0.4X_3^3 + 0.2X_4^2, 1)$ for both training and target distributions, where $\beta \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$.

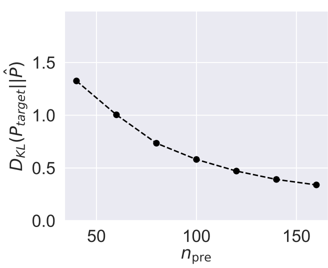
5.1.1 Results with varying sample size n_{pre} for estimating w_0

We first consider the scenario with an estimated density ratio w_0. Recall that we use the subsets $\mathcal{D}_1 \subset \mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test},1} \subset \mathcal{D}_{\text{test}}$ with $|\mathcal{D}_1| = |\mathcal{D}_{\text{test},1}| = n_{\text{pre}}$ for estimating w_0; consequently, for larger values of n_{pre}, we will expect a more accurate w_0. By varying n_{pre} in $\{40, 60, 80, 100, 120, 140, 160\}$, we aim to investigate the robustness of WCP and iso-DRL with respect to the accuracy in w_0, where we fix $\rho = \rho^* := D_{\text{KL}}(P_{\text{target}} \| P)$.

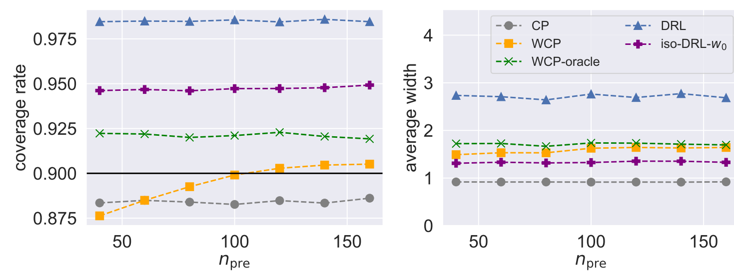
Well-specified setting. In Figure 2b, where the solid horizontal line (in the middle plot) marks the nominal coverage level, $1-\alpha = 90\%$, we can see that the uncorrected CP exhibits undercoverage due to the mismatch between P_{target} and P, while the coverage of WCP using w_0 increases to 90% as n_{pre} increases, since w_0 becomes more accurate with larger n_{pre} (cf. Figure 2a). The generic DRL, even with $\rho = \rho^*$, tends to be conservative and has the widest interval. In comparison, iso-DRL-w_0 has coverage very close to the target level.

Misspecified w_0. In Figure 3b, we show results for the misspecified setting. Since w_0 is estimated from a model class that does not contain the true density ratio, consequently $D_{\text{KL}}(P_{\text{target}} \| \widehat{P})$ does

[3] To explain this construction, recall from Section 1 that we can use the excess risk estimate to choose a tuning parameter that achieves a desired bound on risk. Specifically, for any value of $\widetilde{\alpha}$, we can bound the risk (i.e., the miscoverage) for the CP interval $C_{1-\widetilde{\alpha}}$ as $\mathbb{E}_{\widetilde{P}_{\text{target}}}[Y \notin C_{1-\widetilde{\alpha}}(X)] \le \mathbb{E}_{\widetilde{P}}[Y \notin C_{1-\widetilde{\alpha}}(X)] + \Delta(R_{\widetilde{\alpha}}; \mathcal{B}_{f,\rho}) \le \widetilde{\alpha} + \Delta(R_{\widetilde{\alpha}}; \mathcal{B}_{f,\rho})$ (where $R_{\widetilde{\alpha}}$ is the risk defined by the CP interval $C_{1-\widetilde{\alpha}}$, for any value of $\widetilde{\alpha}$). Since $a \mapsto R_a$ is nondecreasing, this also implies that $a \mapsto \Delta(R_a; \mathcal{B}_{f,\rho})$ is nondecreasing (recall from Section 2.1 that $\Delta(R; \mathcal{B})$ is monotone in R, as a corollary of Proposition 2.2). Thus, for $\widetilde{\alpha} \le \alpha$ we have $\mathbb{E}_{\widetilde{P}_{\text{target}}}[Y \notin C_{1-\widetilde{\alpha}}(X)] \le \widetilde{\alpha} + \Delta(R_\alpha; \mathcal{B}_{f,\rho}) \approx \widetilde{\alpha} + \widehat{\Delta}(\alpha)$. Consequently, the above choice of $\widetilde{\alpha}$ ensures that miscoverage will be (approximately) bounded by α. A similar argument also holds for iso-DRL-w_0.

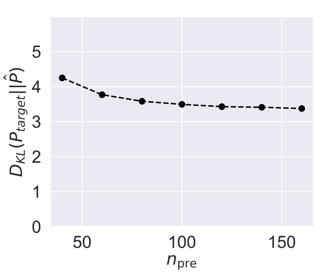


(a) $D_{\text{KL}}(P_{\text{target}}\|\widehat{P})$ vs. n_{pre}.

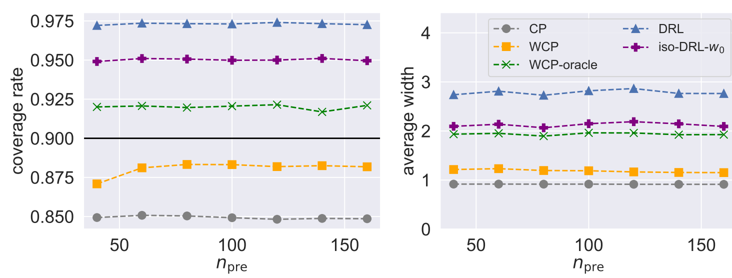


(b) Comparison of all methods with varying n_{pre}.

Figure 2: Results in the well-specified setting.



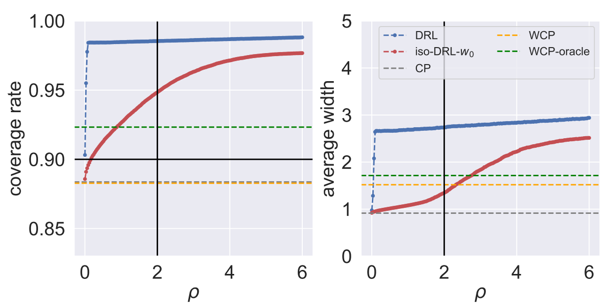
(a) $D_{\text{KL}}(P_{\text{target}}\|\widehat{P})$ vs. n_{pre}.



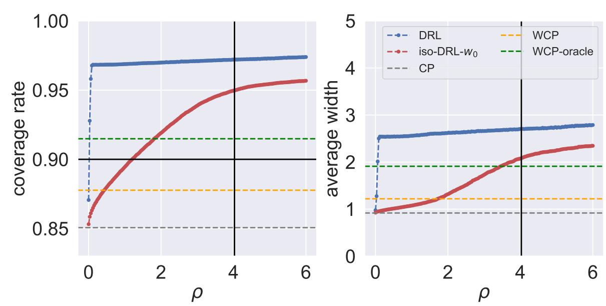
(b) Comparison of all methods with varying n_{pre}.

Figure 3: Results in the misspecified setting.

not converge to zero as n_{pre} increases (cf. Figure 3a). As a result, both uncorrected CP and WCP (which is weighted with the misspecified w_0) exhibit undercoverage. The proposed iso-DRL-w_0 method has coverage slightly above 90% but has interval width close to that of WCP-oracle, while DRL is overly conservative.



(a) Results in the well-specified setting.



(b) Results in the misspecified setting.

Figure 4: Results with varying ρ.

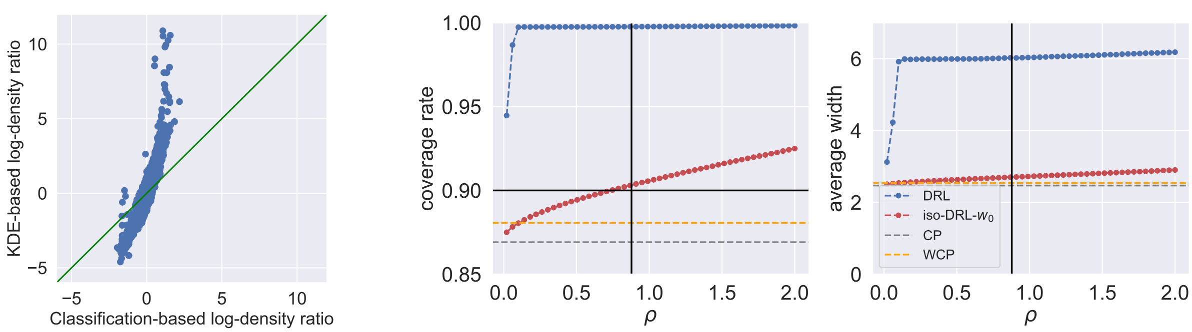
5.1.2 Results with varying ρ

In this section, we investigate the sensitivity of each approach (DRL and iso-DRL-w_0) to the choice of ρ. Fixing $n_{\text{pre}} = 50$, we vary ρ in $[0.002, 6]$. The solid vertical line in each plot denotes the true KL divergence, $\rho^* = D_{\text{KL}}(P_{\text{target}}\|P)$. Other methods that do not depend on ρ behave in the same way as shown in the previous section.

We can see from both plots that the prediction intervals produced by DRL are quite conservative and much wider than the oracle interval across nearly the entire range of ρ, even values ρ much smaller than the true distribution shift magnitude $\rho^* = D_{\mathrm{KL}}(P_{\mathrm{target}} \| P)$. In comparison, for iso-DRL-w_0, when $\rho = \rho^*$, the width of intervals is comparable to the oracle interval in both cases, and the coverage and width vary slowly as we change the value of ρ. From this we can see that the isotonic constraint offers a significant gain in accuracy if we have a reasonable estimate of ρ^*.

5.2 Real data: `wine quality` dataset

We next consider a real dataset: the **`wine quality`** dataset (Cortez et al., 2009)[4]. The dataset includes 12 variables that measure the physicochemical properties of wine and we treat the variable `quality` as the response of interest. The entire dataset consists of two groups: the white and red variants of the Portuguese "Vinho Verde" wine (1599 data points for the red wine and 4898 data points for the white wine). The subset of red wine is treated as the test dataset and that of white wine is viewed as the training set. All variables are nonnegative and we scale each variable by its largest value such that the entries are bounded by 1. Similar to the dataset partition in synthetic simulation, we fix $n_{\mathrm{pre}} = 50$, $n_{\mathrm{train}} = n = 1900$, and $n_{\mathrm{test}} = 1000$. We first fit a kernel density estimator (Gaussian kernel with a bandwidth suggested by cross-validation) using the entire dataset as a proxy of the oracle density ratio. Figure 5a plots this against the log-density ratio obtained from logistic regression fitted on n_{pre} samples from each group. It can be seen that the two density ratios exhibit an approximately isotonic trend, which motivates us to consider the isotonic constraint with respect to w_0.



(a) Log-density ratio estimation: KDE versus logistic regression.

(b) Comparison with varying ρ (sample size $n_{\mathrm{pre}} = 50$).

Figure 5: Results for **`wine quality`** dataset.

To assess the performance of the proposed approach, we estimate w_0 using the same procedure as for the simulated data, with sample size $n_{\mathrm{pre}} = 50$ for each group. We consider the uncertainty set defined by KL-divergence and choose ρ from 50 uniformly located grid points in $[0.02, 2]$ in Figure 5b. The solid vertical line (in the middle and right plots) denotes an estimate $\widehat{\rho}$ of the KL divergence $D_{\mathrm{KL}}(P_{\mathrm{target}} \| P)$ to ensure that we are considering a reasonable range of values of ρ (see

[4]Available at `https://archive.ics.uci.edu/dataset/186/wine+quality`

Appendix D.2 for details on this estimate). In Figure 5b, similar to the performance in Section 5.1 for simulated data, DRL tends to be conservative: the coverage rate quickly approaches 1 while ρ is still below 0.1 and the intervals tend to be wide. In the meantime, iso-DRL-w_0 captures the approximate isotonic trend in Figure 5a and achieves valid coverage by recalibrating the weighted approach. The key message is that in the real data case, even when there is no oracle information for selecting ρ and the isotonic trend is not exact, the proposed iso-DRL-w_0 with the isotonic constraint with respect to the pre-fitted density ratio is less sensitive to the selection of ρ.

6 Additional related work

In this section, we discuss some additional literature in several related areas, including transfer learning, DRL, sensitivity analysis, shape-constrained learning, and conformal prediction.

Transfer learning. Transfer learning, in which data from one distribution is used to improve performance on a related but different distribution, is usually categorized into domain adaptation and inductive transfer learning (Redko et al., 2020).

Domain adaptation focuses on the scenario with covariate shift. From the theoretical side, the performance of machine learning models is analyzed in Ben-David et al. (2010); Ben-David and Urner (2012); Pathak et al. (2022); Pathak and Ma (2024); Hanneke and Kpotufe (2019). To implement efficient predictions, weighted methods are adopted as the first trial to draw P closer to Q (Cortes et al., 2008; Gretton et al., 2009; Ma et al., 2023; Ge et al., 2023). Another scenario requires a small number of labeled target samples, which can be feasible in reality and related works include Chen et al. (2011); Chattopadhyay et al. (2013); Yang et al. (2012), etc. Inductive transfer learning, on the other hand, assumes that the marginal distributions of X are invariant for training and target distributions and is studied in difference statistical settings (Bastani, 2021; Cai and Wei, 2019; Li et al., 2021; Tian and Feng, 2021).

Distributionally robust learning (DRL). Our work is directly related to DRL (Ben-Tal and Nemirovski, 1998; El Ghaoui and Lebret, 1997), which aims to control certain statistical risks uniformly over a set of candidate distributions for the target distribution. Different classes of the uncertainty set are studied in the literature, such as the optimal transport discrepancy (Shafieezadeh Abadeh et al., 2015; Blanchet and Murthy, 2019; Blanchet et al., 2019; Esfahani and Kuhn, 2015) and f-divergence (Duchi et al., 2021; Duchi and Namkoong, 2018; Weiss et al., 2023). Further constraints on the uncertainty set as the improvement of DRL are explored by Duchi et al. (2019); Setlur et al. (2023); Esteban-Pérez and Morales (2022); Liu et al. (2023); Popescu (2007); Shapiro and Pichler (2023). The recent work of Wang et al. (2023) considers the constraint that the unseen target distribution is a mixture of data distributions from multiple sources.

Sensitivity analysis. Sensitivity analysis is closely related to DRL and is widely studied in the field of causal inference (Cornfield et al., 1959; Rosenbaum, 1987; Tan, 2006; Ding and VanderWeele, 2016; Zhao et al., 2019; De Bartolomeis et al., 2023) with the goal of evaluating the effect of unmeasured confounders and relaxing untestable assumptions. Sensitivity models can be viewed as a specific example of constraints on distribution shift. For example, the marginal Γ-selection model

(Tan, 2006) with a binary treatment T imposes a bound constraint on the distribution shift from the data distribution $P_{Y(1)|X,T=1}$ to the counterfactual $P_{Y(1)|X,T=0}$. Recent works also investigate the sensitivity model from the perspective of DRL, such as Yadlowsky et al. (2018); Jin et al. (2022, 2023); Sahoo et al. (2022). Sensitivity analysis that incorporates more informative constraints is explored in Huang and Pimentel (2024); Nie et al. (2021).

Statistical learning with shape constraints. Our work also borrows ideas from shape-constrained learning, which have been studied across various applications (Grenander, 1956; Matzkin, 1991). The isotonic constraint is the most common one among these. Since Rao (1969), the properties of isotonic regression are well studied in the literature (Brunk et al., 1957, 1972; Zhang, 2002; Han et al., 2019; Yang and Barber, 2019; Durot and Lopuhaä, 2018). Moreover, the isotonic constraint is also widely applied to calibration for distributions in regression and classification settings (Zadrozny and Elkan, 2002; Niculescu-Mizil and Caruana, 2012; van der Laan et al., 2023; Henzi et al., 2021; Berta et al., 2024).

Conformal prediction. One important application of iso-DRL is to recalibrate conformal prediction intervals. Conformal prediction (Vovk et al., 2005; Shafer and Vovk, 2008) provides a framework for distribution-free uncertainty quantification, which constructs prediction intervals that are valid with exchangeable data from any underlying distribution and with any "black-box" algorithm. As the validity of WCP (Tibshirani et al., 2019) with the estimated density ratio only holds up to a coverage gap due to the error the estimate w_0 (Lei and Candès, 2020; Candès et al., 2023; Gui et al., 2024), the work Jin et al. (2023) further establish a robust guarantee via sensitivity analysis. Besides the weighted approaches, there are other solutions in the literature: Cauchois et al. (2020); Ai and Ren (2024) address the issue of joint distribution shift via the DRL; Qiu et al. (2023); Yang et al. (2024); Chen and Lei (2024) formulate the covariate shift problem within the semiparametric/nonparametric framework and utilize the doubly-robust theory to correct the distributional bias.

7 Discussion

In this paper, we focus on distributionally robust risk evaluation with the isotonic constraint on the density ratio as the regularization, which aims to avoid over-pessimistic candidate distributions. This is similar in flavor to many tools in high-dimensional statistical learning, where regularization/inductive bias is introduced to improve generalization. We provide an efficient approach to solve the shape-constrained optimization problem via an equivalent reformulation, for which estimation error bounds for the worst-case excess risk are also provided. To conclude, we provide further discussions on the proposed iso-DRL framework and highlight several open questions.

Stability against distribution shift. Excess risk can also be interpreted from the perspective of stability against distribution shift (Lam, 2016; Namkoong et al., 2022; Rothenhäusler and Bühlmann, 2023). With a fixed budget $\varepsilon \ll 1$ for the excess risk, the largest tolerance of distribution shift such that the excess risk is under control is of interest. Taking the f-divergence constraint as an example, to ensure $\Delta_\rho(R; \mathcal{B}_{f,\rho}) \leq \varepsilon \ll 1$, then ρ needs to obey $\rho \leq (2\mathrm{Var}(R(X)))^{-1} f''(1) \cdot \varepsilon^2 +$

$o(\varepsilon^2)$ (Lam, 2016; Duchi and Namkoong, 2018; Blanchet and Shapiro, 2023). However, with the additional isotonic constraint on the density ratio, we can tolerate larger distribution shift: $\rho \leq (2\mathrm{Var}([\pi(R)](X)))^{-1} f''(1) \cdot \varepsilon^2 + o(\varepsilon^2)$. This improvement implies that when side information of the underlying distribution shift is provided, risk evaluation will be less sensitive to the hyperparameters describing the uncertainty set (e.g., ρ), thus is more robust with the presence of distribution shift.

From risk evaluation to distributionally robust optimization. Different from risk evaluation, distributionally robust optimization (DRO) focuses on the optimization problem with a loss function $\ell_\theta(x)$, i.e., $\widehat{\theta} \in \underset{\theta\in\Theta}{\mathrm{argmin}} \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q\, \ell_\theta(X)$. Under smoothness conditions on ℓ_θ, asymptotic normality for $\widehat{\theta}$ is established in the literature (Duchi and Namkoong, 2018). The DRO framework is shown to regularize $\widehat{\theta}$ in terms of variance penalization (Lam, 2016; Duchi and Namkoong, 2018) or explicit norm regularization (Blanchet and Murthy, 2019). It is interesting to incorporate the isotonic constraint into DRO and to understand the effect of the isotonic constraint in the asymptotics of $\widehat{\theta}^{\mathrm{iso}}$.

Acknowledgements

R.F.B. was supported by the Office of Naval Research via grant N00014-20-1-2337, and by the National Science Foundation via grant DMS-2023109. C.M. was partially supported by the National Science Foundation via grant DMS-2311127.

Appendix

A Proofs of results in Section 2

A.1 Proof of Proposition 2.2

It is straightforward to check that $\Delta(R;\mathcal{B})$ is always an upper bound of the new formulation stated in Proposition 2.2, simply by taking $w = \phi \circ R$. Therefore, it remains to show the converse: under Condition 2.1, $\Delta(R;\mathcal{B})$ is also a *lower* bound of the new formulation stated in Proposition 2.2.

To this end, it suffices to prove that for any $w_\# P \in \mathcal{B}$, there exists a nondecreasing function ϕ such that $(\phi \circ R)_\# P \in \mathcal{B}$, and

$$\mathbb{E}_P\left[w(X)R(X)\right] \leq \mathbb{E}_P\left[\phi(R(X))R(X)\right].$$

We construct such a function ϕ in two steps.

Step 1: Conditioning. For any w such that $w_\# P \in \mathcal{B}$, we define g as a measurable function satisfying

$$g(R(X)) = \mathbb{E}\left[w(X) \mid R(X)\right], \quad P\text{-almost surely.}$$

(Note that g is not necessarily a monotone function.) As a result, by the tower law, we have

$$\mathbb{E}_P[w(X)R(X)] = \mathbb{E}_P[g(R(X))R(X)]. \tag{13}$$

Since $w_{\#}P \in \mathcal{B}$, by Jensen's inequality, for any convex function ψ, we have

$$\mathbb{E}_P\left[\psi\left(g(R(X))\right)\right] = \mathbb{E}_P\left[\psi\left(\mathbb{E}\left[w(X) \mid R(X)\right]\right)\right] \le \mathbb{E}_P\left[\psi(w(X))\right],$$

which implies $(g \circ R)_{\#}P \in \mathcal{B}$ by Condition 2.1.

Step 2: Rearrangement. Denote F_1 and F_2 as the cumulative distribution functions of $g(R(X))$ and $R(X)$, respectively. Let $U \sim \mathrm{Unif}([0,1])$. Then, we have $F_1^{-1}(U) \overset{d}{=} g(R(X))$ and $F_2^{-1}(U) \overset{d}{=} R(X)$, where F_k^{-1} is the generalized inverse of F_k for $k = 1, 2$, and where $\overset{d}{=}$ denotes equality in distribution. Moreover, F_1^{-1} is nondecreasing and

$$g(F_2^{-1}(U)) \overset{d}{=} g(R(X)) \overset{d}{=} F_1^{-1}(U),$$

which implies that F_1^{-1} is the monotone rearrangement of $g \circ F_2^{-1}$. By (Hardy et al., 1952, eqn. (378)), we have

$$\mathbb{E}_P\left[g(R(X))R(X)\right] = \mathbb{E}\left[g(F_2^{-1}(U))F_2^{-1}(U)\right] \le \mathbb{E}\left[F_1^{-1}(U)F_2^{-1}(U)\right]. \tag{14}$$

Next, let ϕ be a measurable function satisfying

$$\phi(F_2^{-1}(U)) = \mathbb{E}\left[F_1^{-1}(U) \mid F_2^{-1}(U)\right],$$

almost surely with respect to the distribution $U \sim \mathrm{Unif}([0,1])$. Since F_k^{-1} is the generalized inverse of a CDF F_k, for each $k = 1, 2$, it is therefore monotone nondecreasing. Therefore, we can choose ϕ to be a monotone nondecreasing function. Moreover, to verify that $(\phi \circ R)_{\#}P \in \mathcal{B}$, we will check that $\phi(R(X)) \overset{cvx}{\preceq} g(R(X))$ (and use Condition 2.1, along with the fact that $(g \circ R)_{\#}P \in \mathcal{B}$ as established above): for any convex function ψ, we have

$$\begin{aligned}\mathbb{E}_P[\psi(\phi(R(X)))] &\overset{d}{=} \mathbb{E}[\psi(\phi(F_2^{-1}(U)))] \\ &= \mathbb{E}[\psi(\mathbb{E}\left[F_1^{-1}(U) \mid F_2^{-1}(U)\right])] \le \mathbb{E}[\psi(F_1^{-1}(U))] = \mathbb{E}_P[\psi(g(R(X)))],\end{aligned}$$

where the inequality holds by Jensen's inequality.

We then have

$$\begin{aligned}\mathbb{E}\left[F_1^{-1}(U)F_2^{-1}(U)\right] &= \mathbb{E}\left[\mathbb{E}\left[F_1^{-1}(U) \mid F_2^{-1}(U)\right] F_2^{-1}(U)\right] \\ &= \mathbb{E}\left[\phi(F_2^{-1}(U))F_2^{-1}(U)\right] = \mathbb{E}_P\left[\phi(R(X))R(X)\right].\end{aligned}$$

This equality, combined with (13) and (14), yields the desired outcome: $\mathbb{E}_P\left[w(X)R(X)\right] \le \mathbb{E}_P\left[\phi(R(X))R(X)\right]$. We hence complete the proof.

B Proofs of results in Section 3

B.1 Preliminaries

Before we present the proof, we begin with some preliminaries: we introduce some notation, definitions, and facts that will aid in the proof below.

B.1.1 Adding an L_2 constraint

First, we will define a version of our optimization problem that defines $\Delta(R;\mathcal{B})$, by adding an L_2 constraint:

$$\begin{aligned} \Delta_2(R;\mathcal{B}) \quad = \quad & \sup_{w\geq 0,\ w\in L_2(P)} \quad \mathbb{E}_P\left[w(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ & \text{subject to} \quad w_{\#}P \in \mathcal{B}. \end{aligned} \tag{15}$$

We can observe that, by construction,

$$\Delta_2(R;\mathcal{B}) = \Delta(R;\mathcal{B}\cap\mathcal{B}_{L_2}),$$

where $\mathcal{B}_{L_2}$ is the set of all distributions with finite second moment. The following result verifies that adding the L_2 constraint does not change the outcome of the optimization problem:

Proposition B.1. *Under the notation and definitions above, it holds that $\Delta(R;\mathcal{B}) = \Delta_2(R;\mathcal{B})$.*

We defer the proof of this proposition to Section B.5.

B.1.2 The isotonic projection

We next review some facts regarding the isotonic projection operator π. To ease notation, we denote $\langle a,b\rangle_P = \int_{\mathcal{X}} a(x)b(x)\mathsf{d}P(x)$ for any functions $a,b\in L_2(P)$.

The first property relates to the isotonic projection as a projection to a convex cone (Bauschke and Combettes (2019), Theorem 3.14; Edwards (2012), Proposition 1.12.4):

$$\text{For any } w\in L_2(P) \text{ and any } v\in \mathcal{C}^{\text{iso}}_{\preceq}\cap L_2(P),\ \langle v, w-\pi(w)\rangle_P \leq 0. \tag{16}$$

Moreover, it holds that (Brunk (1963), Theorem 1; Brunk (1965), Corollary 3.1):

$$\text{For any } w\in L_2(P) \text{ and any } h:\mathbb{R}\to\mathbb{R},\ \langle h\circ\pi(w), w-\pi(w)\rangle_P = 0. \tag{17}$$

In particular, by chosing $h(t)\equiv 1$, we can see that isotonic projection preserves the mean,

$$\text{For any } w\in L_2(P),\ \ \mathbb{E}_P[w(X)] = \mathbb{E}_P[[\pi(w)](X)]. \tag{18}$$

Finally, we relate the isotonic projection to the convex ordering:

$$\text{For any } w\in L_2(P),\ \pi(w) \overset{cvx}{\preceq} w. \tag{19}$$

To see (19), for any convex function ψ, by the nonnegativity of Bregman divergence (Bregman, 1967), it holds that

$$\langle \psi(w)-\psi(\pi(w)), 1\rangle_P \geq \langle \psi'\circ\pi(w), w-\pi(w)\rangle_P.$$

According to the property (17), we further obtain $\langle \psi'\circ\pi(w), w-\pi(w)\rangle_P$, which implies that $\pi(w) \overset{cvx}{\preceq} w$ by definition.

B.2 Proof of Theorem 3.1

We split the proof into three steps:

1. prove that $\Delta^{\text{iso}}(R;\mathcal{B}) \leq \Delta(\pi(R);\mathcal{B})$;
2. prove that $\Delta^{\text{iso}}(R;\mathcal{B}) = \Delta(\pi(R);\mathcal{B})$ provided that Condition 2.1 holds;
3. prove the claim on attainability of minimizers provided that Condition 2.1 holds.

Step 1: Prove $\Delta^{\text{iso}}(R;\mathcal{B}) \leq \Delta(\pi(R);\mathcal{B})$. By the definition of $\Delta^{\text{iso}}(R;\mathcal{B})$ as the supremum in the optimization problem (6), for any $\varepsilon > 0$, there exists a feasible w_ε such that

$$\mathbb{E}_P[w_\varepsilon(X) \cdot R(X)] - \mathbb{E}_P[R(X)] \geq \Delta^{\text{iso}}(R;\mathcal{B}) - \varepsilon. \tag{20}$$

Next, define a sequence of truncated functions, $w_{\varepsilon,n}(x) = \min\{w_\varepsilon(x), n\}$. Since $w_\varepsilon \in \mathcal{C}^{\text{iso}}_{\preceq}$, it holds that $w_{\varepsilon,n} \in \mathcal{C}^{\text{iso}}_{\preceq}$ as well, and moreover since the truncated function is bounded we also have $w_{\varepsilon,n} \in L_2(P)$. By fact (16), it therefore holds that

$$\mathbb{E}_P[w_{\varepsilon,n}(X) \cdot (R - [\pi(R)])(X)] = \langle w_{\varepsilon,n}, R - \pi(R)\rangle_P \leq 0,$$

for each $n \geq 1$. Then, by the dominated convergence theorem, taking a limit as $n \to \infty$ we obtain

$$\mathbb{E}_P[w_\varepsilon(X) \cdot (R - [\pi(R)])(X)] \leq 0. \tag{21}$$

Moreover, $\mathbb{E}_P[[\pi(R)](X)] = \mathbb{E}_P[R(X)]$ by (18). Combining everything, then,

$$\begin{aligned}\Delta^{\text{iso}}(R;\mathcal{B}) - \varepsilon &\leq \mathbb{E}_P[w_\varepsilon(X) \cdot R(X)] - \mathbb{E}_P[R(X)] \\ &\leq \mathbb{E}_P[w_\varepsilon(X) \cdot [\pi(R)](X)] - \mathbb{E}_P[[\pi(R)](X)] \leq \Delta(\pi(R);\mathcal{B}),\end{aligned}$$

where the last step holds since, because w_ε is feasible for the optimization problem (6) that defines $\Delta^{\text{iso}}(R;\mathcal{B})$, it is also feasible for $\Delta(\pi(R);\mathcal{B})$ (i.e., $w \geq 0$ and $w_\# P \in \mathcal{B}$). Since $\varepsilon > 0$ is arbitrary, we obtain the desired result $\Delta^{\text{iso}}(R;\mathcal{B}) \leq \Delta(\pi(R);\mathcal{B})$.

Step 2: Prove $\Delta^{\text{iso}}(R;\mathcal{B}) = \Delta(\pi(R);\mathcal{B})$ under Condition 2.1. By Proposition B.1, we have $\Delta(\pi(R);\mathcal{B}) = \Delta_2(\pi(R);\mathcal{B}) = \Delta(\pi(R);\mathcal{B} \cap \mathcal{B}_{L_2})$. Next, note that if Condition 2.1 holds for $\mathcal{B}$, then this condition holds for $\mathcal{B} \cap \mathcal{B}_{L_2}$ as well (because for any $Q' \overset{cvx}{\preceq} Q$, we have $\mathbb{E}_{Q'}[X^2] \leq \mathbb{E}_Q[X^2]$ by definition of the convex ordering—and so if $Q \in \mathcal{B}_{L_2}$ then $Q' \in \mathcal{B}_{L_2}$ as well.) Therefore, we can apply Proposition 2.2 to the term $\Delta(\pi(R);\mathcal{B} \cap \mathcal{B}_{L_2})$, which yields the following equivalent formulation:

$$\begin{aligned}\Delta(\pi(R);\mathcal{B} \cap \mathcal{B}_{L_2}) \quad = \quad &\sup_{\phi:\mathbb{R}\to\mathbb{R}_+} \quad \mathbb{E}_P\left[(\phi \circ \pi(R))(X) \cdot [\pi(R)](X)\right] - \mathbb{E}_P[[\pi(R)](X)] \\ &\text{subject to} \quad (\phi \circ \pi(R))_\# P \in \mathcal{B}, \quad \phi \circ \pi(R) \in L_2(P), \quad \phi \text{ is nondecreasing.}\end{aligned} \tag{22}$$

Then, for any $\varepsilon > 0$, there exists some ϕ_ε satisfying the above constraints so that

$$\mathbb{E}_P\left[(\phi_\varepsilon \circ \pi(R))(X) \cdot [\pi(R)](X)\right] - \mathbb{E}_P[[\pi(R)](X)] \geq \Delta(\pi(R);\mathcal{B} \cap \mathcal{B}_{L_2}) - \varepsilon = \Delta(\pi(R);\mathcal{B}) - \varepsilon.$$

Now define $\widetilde{w}_\varepsilon = \phi_\varepsilon \circ \pi(R)$, i.e., we have

$$\mathbb{E}_P\left[\widetilde{w}_\varepsilon(X) \cdot [\pi(R)](X)\right] - \mathbb{E}_P[[\pi(R)](X)] \geq \Delta(\pi(R); \mathcal{B}) - \varepsilon,$$

where $(\widetilde{w}_\varepsilon)_\# P \in \mathcal{B}$ and $\widetilde{w}_\varepsilon \in L_2(P)$, and also $\widetilde{w}_\varepsilon \in \mathcal{C}^{\text{iso}}_{\preceq}$, by construction and by feasibility of ϕ_ε. Moreover, by the facts (18) and (17),

$$\mathbb{E}_P[[\pi(R)](X)] = \mathbb{E}_P[R(X)], \quad \langle \widetilde{w}_\varepsilon, R - \pi(R)\rangle_P = \langle \phi_\varepsilon \circ \pi(R), R - \pi(R)\rangle_P = 0,$$

and therefore,

$$\mathbb{E}_P\left[\widetilde{w}_\varepsilon(X) \cdot R(X)\right] - \mathbb{E}_P[R(X)] \geq \Delta(\pi(R); \mathcal{B}) - \varepsilon.$$

But we have verified above that $\widetilde{w}_\varepsilon$ is feasible for the optimization problem (6) defining $\Delta^{\text{iso}}(R; \mathcal{B})$, i.e.,

$$\mathbb{E}_P\left[\widetilde{w}_\varepsilon(X) \cdot R(X)\right] - \mathbb{E}_P[R(X)] \leq \Delta^{\text{iso}}(R; \mathcal{B}).$$

Since $\varepsilon > 0$ is arbitrary, this verifies that $\Delta(\pi(R); \mathcal{B}) \leq \Delta^{\text{iso}}(R; \mathcal{B})$, and thus completes this step.

Step 3: attainability of minimizers under Condition 2.1. Suppose $\Delta(\pi(R); \mathcal{B})$ is attained at $\widetilde{w}$, i.e.,

$$\mathbb{E}_P[\widetilde{w}(X) \cdot [\pi(R)](X)] - \mathbb{E}_P[[\pi(R)](X)] = \Delta(\pi(R); \mathcal{B}).$$

By Proposition 2.2, we can construct some nondecreasing function $\widetilde{\phi}$, with $(\widetilde{\phi} \circ \pi(R))_\# P \in \mathcal{B}$, such that

$$\Delta(\pi(R); \mathcal{B}) = \mathbb{E}_P[\widetilde{\phi}([\pi(R)](X)) \cdot [\pi(R)](X)] - \mathbb{E}_P[[\pi(R)](X)].$$

Recalling that $\mathbb{E}_P[[\pi(R)](X)] = \mathbb{E}_P[R(X)]$ by (18), and $\Delta(\pi(R); \mathcal{B}) = \Delta^{\text{iso}}(R; \mathcal{B})$ by Steps 1 and 2, we now have

$$\Delta^{\text{iso}}(R; \mathcal{B}) = \mathbb{E}_P[\widetilde{\phi}([\pi(R)](X)) \cdot [\pi(R)](X)] - \mathbb{E}_P[R(X)].$$

Next, by fact (17),

$$\mathbb{E}_P[\widetilde{\phi}([\pi(R)](X)) \cdot (R(X) - [\pi(R)](X))] = \langle \widetilde{\phi} \circ \pi(R), R - \pi(R)\rangle_P = 0,$$

and so

$$\Delta^{\text{iso}}(R; \mathcal{B}) = \mathbb{E}_P[\widetilde{\phi}([\pi(R)](X)) \cdot R(X)] - \mathbb{E}_P[R(X)].$$

Therefore, $\Delta^{\text{iso}}(R; \mathcal{B})$ is attained at $\widetilde{\phi} \circ \pi(R)$ (which, by construction, satisfies $\widetilde{\phi} \circ \pi(R) \in \mathcal{C}^{\text{iso}}_{\preceq}$, as well as $(\phi \circ \pi(R))_\# P \in \mathcal{B}$ as above, and is therefore feasible).

Conversely, suppose that $\Delta^{\text{iso}}(R; \mathcal{B})$ is attained at $\widetilde{w}$, i.e.,

$$\mathbb{E}_P[\widetilde{w}(X) \cdot R(X)] - \mathbb{E}_P[R(X)] = \Delta^{\text{iso}}(R; \mathcal{B}).$$

Again applying (18), and the fact that $\Delta(\pi(R); \mathcal{B}) = \Delta^{\text{iso}}(R; \mathcal{B})$ by Steps 1 and 2,

$$\Delta(\pi(R); \mathcal{B}) = \mathbb{E}_P[\widetilde{w}(X) \cdot R(X)] - \mathbb{E}_P[[\pi(R)](X)] \leq \mathbb{E}_P[\widetilde{w}(X) \cdot [\pi(R)](X)] - \mathbb{E}_P[[\pi(R)](X)],$$

where for the last step, since $\widetilde{w} \in \mathcal{C}^{\text{iso}}_{\preceq}$ (because it is feasible for $\Delta^{\text{iso}}(R; \mathcal{B})$), we have

$$\mathbb{E}_P[\widetilde{w}(X) \cdot (R(X) - [\pi(R)](X))] = \langle \widetilde{w}, R - \pi(R)\rangle_P \leq 0,$$

by (16). But $\widetilde{w}$ is feasible for $\Delta(\pi(R); \mathcal{B})$ (since it is feasible for $\Delta^{\text{iso}}(R; \mathcal{B})$), and therefore, we also have

$$\Delta(\pi(R); \mathcal{B}) = \Delta^{\text{iso}}(R; \mathcal{B}) \leq \mathbb{E}_P[\widetilde{w}(X) \cdot [\pi(R)](X)] - \mathbb{E}_P[[\pi(R)](X)].$$

In other words, $\Delta(\pi(R); \mathcal{B})$ is attained at $\widetilde{w}$, which completes the proof.

B.3 Proof of Proposition 3.2

We formally define $\Delta^{\text{iso}}(R; \mathcal{B}, w_0)$ as follows:

$$\begin{aligned} \Delta^{\text{iso}}(R; \mathcal{B}, w_0) \quad &= \quad \sup_{w \geq 0} \quad \mathbb{E}_P\left[w(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ &\text{subject to} \quad w_{\#}P \in \mathcal{B}, \quad w \in \mathcal{C}^{\text{iso}}_{w_0}. \end{aligned} \tag{23}$$

For comparison, we also consider the following optimization problem

$$\begin{aligned} \widetilde{\Delta}^{\text{iso}}(R; \mathcal{B}, w_0) \quad &= \quad \sup_{h:\ h \circ w_0 \geq 0} \quad \mathbb{E}_P\left[(h \circ w_0)(X)R(X)\right] - \mathbb{E}_P[R(X)] \\ &\text{subject to} \quad (h \circ w_0)_{\#}\, P \in \mathcal{B}, \quad h \in \mathcal{C}^{\text{iso}}_1, \end{aligned} \tag{24}$$

where $\mathcal{C}^{\text{iso}}_1$ denotes the cone of isotonic functions defined on $\mathbb{R}$ equipped with the natural ordering. In fact, since $\mathcal{C}^{\text{iso}}_{w_0} = \{h \circ w_0 : h \in \mathcal{C}^{\text{iso}}_1\}$ by definition, we therefore have $\Delta^{\text{iso}}(R; \mathcal{B}, w_0) = \widetilde{\Delta}^{\text{iso}}(R; \mathcal{B}, w_0)$.

In addition, recalling that $\widetilde{R}(w_0(X)) = \mathbb{E}_P[R(X) \mid w_0(X)]$, by the change of measure, the optimization problem in (24) can be further rewritten as

$$\begin{aligned} \Delta^{\text{iso}}(R; \mathcal{B}, w_0) \quad &= \quad \sup_{h:\ h \geq 0} \quad \mathbb{E}_{(w_0)_{\#}P}\left[h(U)\widetilde{R}(U)\right] - \mathbb{E}_{(w_0)_{\#}P}\left[\widetilde{R}(U)\right] \\ &\text{subject to} \quad (h \circ w_0)_{\#}P \in \mathcal{B}, \quad h \in \mathcal{C}^{\text{iso}}_1. \end{aligned} \tag{25}$$

We observe that (25) has the same form with the definition of $\Delta^{\text{iso}}(R; \mathcal{B})$ in (6), where we consider the probability measure $(w_0)_{\#}P$ instead of P and $\widetilde{R}$ in place of R, and with the specific isotonic cone $\mathcal{C}^{\text{iso}}_1$ on $\mathbb{R}$.

Applying Theorem 3.1 to (25) yields

$$\begin{aligned} \Delta^{\text{iso}}(R; \mathcal{B}, w_0) \quad &= \quad \sup_{h:\ h \geq 0} \quad \mathbb{E}_{(w_0)_{\#}P}\left[h(U)[\pi_1(\widetilde{R})](U)\right] - \mathbb{E}_{(w_0)_{\#}P}\left[\widetilde{R}(U)\right] \\ &\text{subject to} \quad (h \circ w_0)_{\#}P \in \mathcal{B}, \end{aligned} \tag{26}$$

where π_1 is the projection onto $\mathcal{C}^{\text{iso}}_1$ under the measure $(w_0)_{\#}P$. By definition of $\widetilde{R}$, we can rewrite this as

$$\begin{aligned} \Delta^{\text{iso}}(R; \mathcal{B}, w_0) \quad &= \quad \sup_{h:\ h \geq 0} \quad \mathbb{E}_P\left[h(w_0(X))[\pi_1(\widetilde{R})](w_0(X))\right] - \mathbb{E}_P\left[\widetilde{R}(w_0(X))\right] \\ &\text{subject to} \quad (h \circ w_0)_{\#}P \in \mathcal{B}, \end{aligned}$$

which is equal to $\Delta(\pi_1(\widetilde{R}) \circ w_0; \mathcal{B}, w_0)$ as defined in (8) since we also have $\mathbb{E}_P\left[\widetilde{R}(w_0(X))\right] = \mathbb{E}_P\left[[\pi_1(\widetilde{R})](w_0(X))\right]$ by (18). We herein complete the proof.

B.4 A misspecified isotonic constraint

When the true distribution shift does not obey the isotonic constraint exactly, we can nonetheless provide a bound on the worst-case excess risk, which is tighter than the (non-iso) DRL bound whenever the isotonic constraint provides a reasonable approximation.

Denote $\widetilde{w}^*$ as the underlying density ratio $\mathsf{d}P_{\text{target}}/\mathsf{d}P$ and $\Delta^*(R) = \mathbb{E}_P[\widetilde{w}^*(X)R(X)] - \mathbb{E}_P[R(X)]$ as the true excess risk. Then, we have the following connections between $\Delta^*(R)$ and $\Delta^{\text{iso}}(R; \mathcal{B})$.

Proposition B.2. *Assume Condition 2.1 holds. If $\widetilde{w}^*_{\#}P \in \mathcal{B}$ and $\widetilde{w}^* \in L_2(P)$, then we have*

$$\Delta^*(R) \le \Delta^{\rm iso}(R;\mathcal{B}) + \mathbb{E}_P\bigg[[\widetilde{w}^* - \pi(\widetilde{w}^*)](X)\cdot[R-\pi(R)](X)\bigg].$$

In particular, if either $\widetilde{w}^* \in \mathcal{C}^{\rm iso}_{\preceq}$ or $R \in \mathcal{C}^{\rm iso}_{\preceq}$, then $\Delta^*(R) \le \Delta^{\rm iso}(R;\mathcal{B})$.

The result states that when the isotonic constraint is violated, the worst-case excess risk of iso-DRL will be no worse than the true excess risk plus a gap which can be controlled by the correlation between $[\widetilde{w}^* - \pi(\widetilde{w}^*)](X)$ and $[R-\pi(R)](X)$. In particular, if *either* the risk or the true density ratio is itself isotonic (or approximately isotonic), then the gap term must be zero (or approximately zero)—and so the excess risk calculation $\Delta^{\rm iso}(R;\mathcal{B})$, which is tighter than the non-iso DRL bound $\Delta(R;\mathcal{B})$, will never underestimate the true risk $\Delta^*(R)$ (or will only be a mild underestimate).

B.4.1 Proof of Proposition B.2

Recall that $\widetilde{w}^*$ is the underlying density ratio $\mathsf{d}P_{\rm target}/\mathsf{d}P$. Since $\pi(\widetilde{w}^*) \overset{cvx}{\preceq} w^*$ by (19), and $\mathcal{B}$ is closed under the convex ordering by Condition 2.1, we have $\pi(\widetilde{w}^*)_{\#}P \in \mathcal{B}$; of course, we also have $\pi(\widetilde{w}^*) \in \mathcal{C}^{\rm iso}_{\preceq}$ by definition. Therefore, $\pi(\widetilde{w}^*)$ is feasible for the optimization problem (6), and so we have

$$\Delta^{\rm iso}(R;\mathcal{B}) \ge \mathbb{E}_P\bigg[[\pi(\widetilde{w}^*)](X)R(X)\bigg] - \mathbb{E}_P[R(X)].$$

We therefore have

$$\Delta^*(R) = \mathbb{E}_P[\widetilde{w}^*(X)R(X)] - \mathbb{E}_P[R(X)] \le \Delta^{\rm iso}(R;\mathcal{B}) + \mathbb{E}_P\left[\Big(\widetilde{w}^*(X) - [\pi(\widetilde{w}^*)](X)\Big)R(X)\right].$$

Moreover,

$$\mathbb{E}_P\left[\Big(\widetilde{w}^*(X) - [\pi(\widetilde{w}^*)](X)\Big)\cdot[\pi(R)](X)\right] = \langle \pi(R), \widetilde{w}^* - \pi(\widetilde{w}^*)\rangle_P \le 0$$

by (16), and so we have

$$\mathbb{E}_P\left[\Big(\widetilde{w}^*(X) - [\pi(\widetilde{w}^*)](X)\Big)R(X)\right] \le \mathbb{E}_P\left[\Big(\widetilde{w}^*(X) - [\pi(\widetilde{w}^*)](X)\Big)\cdot\big(R(X) - [\pi(R)](X)\big)\right].$$

This completes the proof.

B.5 Proof of Proposition B.1

For any $w \ge 0$, we define the sequence of truncated functions $\{w_n\}_{n\in\mathbb{N}}$ via

$$w_n(x) = w(x)\cdot \mathbb{1}\{w(x) \le n\} + L_n \cdot \mathbb{1}\{w(x) > n\},$$

where $L_n = \mathbb{E}[w(X) \mid w(X) > n]$. By construction for each n, $\mathbb{E}_P[w_n(X)] = 1$ and, since $\max\{n, L_n\} = L_n < \infty$, $w_n \in L_2(P)$ for each $n \ge 1$.

Step 1: Feasibility of w_n. We first prove the feasibility of w_n. To see this, as $\mathbb{E}_P[w_n(X)] = 1$ by construction, we need to show that $(w_n)_{\#}P \in \mathcal{B}$. By Condition 2.1, since $\mathcal{B}$ is closed under the convex ordering, it suffices to show that

$$\mathbb{E}_P\left[\psi(w_n(X))\right] \le \mathbb{E}_P\left[\psi(w(X))\right] \qquad \text{for any convex function } \psi.$$

This is true by Jensen's inequality, since, by construction, $\mathbb{E}_P[w(X) \mid w_n(X)] = w_n(X)$.

Step 2: Convergence of $\mathbb{E}_P[w_n(X)R(X)]$. To verify the convergence of $\mathbb{E}_P[w_n(X)R(X)]$, consider

$$\begin{aligned}
&\left|\mathbb{E}_P[w_n(X)R(X)] - \mathbb{E}_P[w(X)R(X)]\right| \\
&= \left| \int_{w(x)>n} (L_n - w(x))R(x)\mathsf{d}P(x) \right| \\
&\le B_R \int_{w(x)>n} \left|L_n - w(x)\right| \mathsf{d}P(x) \\
&\le B_R \left(\int_{w(x)>n} w(x)\mathsf{d}P(x) + L_n \mathbb{P}(w(X) > n) \right) \\
&= 2\mathbb{E}_P\left[w(X) \cdot \mathbb{1}\{w(X) > n\}\right].
\end{aligned}$$

Finally, since $\mathbb{E}_P[w(X)] = 1$ (i.e., we know that $w \in L_1(P)$), this means that

$$\lim_{n\to\infty} \mathbb{E}_P\left[w(X) \cdot \mathbb{1}\{w(X) > n\}\right] = 0.$$

Conclusion. For any $\varepsilon > 0$, there exists $w \ge 0$ such that $\mathbb{E}_P[w(X)] = 1$, $w_\# P \in \mathcal{B}$, and

$$\mathbb{E}_P[w(X)R(X)] - \mathbb{E}_P[R(X)] \ge \Delta(R; \mathcal{B}) - \varepsilon/2$$

Then, based on Step 2, for sufficiently large n it holds that $\mathbb{E}_P[w_n(X)R(X)] \ge \mathbb{E}_P[w(X)R(X)] - \varepsilon/2$. From Step 1, we know that w_n is feasible for $\Delta_2(R; \mathcal{B})$, i.e.,

$$\Delta_2(R; \mathcal{B}) \ge \mathbb{E}_P[w_n(X)R(X)] - \mathbb{E}_P[R(X)] \ge (\mathbb{E}_P[w(X)R(X)] - \varepsilon/2) - \mathbb{E}_P[R(X)] \ge \Delta(R; \mathcal{B}) - \varepsilon.$$

Since ε is arbitrary this verifies that $\Delta_2(R; \mathcal{B}) \ge \Delta(R; \mathcal{B})$, and clearly we must have $\Delta_2(R; \mathcal{B}) \le \Delta(R; \mathcal{B})$ by construction, which completes the proof.

C Proofs of results in Section 4

C.1 Proof of Proposition 4.2

To prove the proposition, it suffices to show that $\|w^{*\text{iso}}_{f,\rho}\|_\infty < \infty$. Recall the dual formulation. There exists a pair (λ^*, ν^*) such that

$$w^{*\text{iso}}_{f,\rho}(x) = \mathcal{P}_{[0,+\infty)} \left\{ (f')^{-1} \left(\frac{[\pi(R)](x) - \nu^*}{\lambda^*} \right) \right\}.$$

Note that ν^* is the parameter for standardization, thus to guarantee $\mathbb{E}_P[w^{*\text{iso}}_{f,\rho}(X)] = 1$, we have

$$(f')^{-1} \left(\frac{B_R - \nu^*}{\lambda^*} \right) \ge \sup_{x\in\mathcal{X}} \; w^{*\text{iso}}_{f,\rho}(x) \ge 1.$$

Moreover, it holds that $(f')^{-1}(-\nu^*/\lambda^*) \le \min_{x\in\mathcal{X}} \; w^{*\text{iso}}_{f,\rho}(x) \le 1$. Then, combining the inequalities yields

$$-\lambda^* f'(1) \le \nu^* \le B_R - \lambda^* f'(1). \tag{27}$$

If we further have $\lambda^* \geq \underline{\lambda} > 0$, it holds that

$$\|w_{f,\rho}^{*\text{iso}}\|_\infty \leq (f')^{-1}\left(\frac{B_R + \lambda^* f'(1)}{\lambda^*}\right) \leq (f')^{-1}\left(f'(1) + \frac{B_R}{\underline{\lambda}}\right) < \infty.$$

Then, it remains to prove that $\lambda^* \neq 0$. To see this, consider the KKT condition:

$$\begin{aligned} -[\pi(R)](x) + \lambda^* f'(w_{f,\rho}^{*\text{iso}}(x)) + \nu^* &= 0, \\ \lambda^* \left(\mathbb{E}_P[f(w_{f,\rho}^{*\text{iso}}(X))] - \rho\right) &= 0, \\ \nu^* \left(\mathbb{E}_P[w_{f,\rho}^{*\text{iso}}(X)] - 1\right) &= 0. \end{aligned}$$

If $\lambda^* = 0$, we have $[\pi(R)](X) = \nu^*$ P-almost surely, which implies that $w_{f,\rho}^{*\text{iso}}(X) = 1$ P-almost surely, in which case $w_{f,\rho}^{*\text{iso}}$ is also bounded. Combining pieces above, we have shown that $\|w_{f,\rho}^{*\text{iso}}\|_\infty < \infty$.

C.2 Proof of Theorem 4.4

We first fix any $w \in \mathcal{C}_{\preceq}^{\text{iso}}$ with $w_{\#}P \in \mathcal{B}$. By Condition 4.1, it holds that $w(X) \leq \Omega$ P-almost surely, and therefore without loss of generality we can assume $w \in \mathcal{C}_{\preceq,\Omega}^{\text{iso}}$. Then, by definition of $\varepsilon_{\mathcal{B}}$, for any $\delta > 0$, we can find some $s, t \geq 0$ with $s + t \leq \varepsilon_{\mathcal{B}} + \delta$, such that we have $w'_{\#}\widehat{P}_n \in \mathcal{B}$ by defining $w' = (1-s)\cdot w + t \cdot \mathbf{1}$.

Moreover, by construction, we must have $w' \in \mathcal{C}_{\preceq}^{\text{iso}}$. Therefore, by optimality, we have

$$\begin{aligned} \widehat{\Delta}^{\text{iso}}(\mathcal{B}) &\geq \frac{1}{n}\sum_{i=1}^n w'(X_i) r(X_i, Y_i) - \frac{1}{n}\sum_{i=1}^n r(X_i, Y_i) \\ &= \mathbb{E}_{\widehat{P}_n}[(w'(X) - 1) r(X, Y)] \\ &\geq \mathbb{E}_P[(w'(X) - 1) R(X)] - \varepsilon_R, \end{aligned}$$

where the last inequality is by the definition of ε_R. Plugging in the definition of w', we obtain that

$$\begin{aligned} \widehat{\Delta}^{\text{iso}}(\mathcal{B}) &\geq \mathbb{E}_P[((1-s)w(X) + t) - 1)\, R(X)] - \varepsilon_R \\ &= (1-s)\mathbb{E}_P[(w(X) - 1)R(X)] + (t - s)\mathbb{E}_P[R(X)] - \varepsilon_R \\ &= \mathbb{E}_P[(w(X) - 1)R(X)] - (s+t)\mathbb{E}_P[w(X)R(X)] + t\mathbb{E}_P[(w(X) + 1)R(X)] - \varepsilon_R \\ &\geq \mathbb{E}_P[(w(X) - 1)R(X)] - 2B_R\Omega \cdot \varepsilon_{\mathcal{B}} - \varepsilon_R, \end{aligned}$$

where the last inequality is by the fact that $\|w\|_\infty \leq \Omega$ and R is B_R-bounded, and $\Omega \geq 1$. Since this holds for every $w \in \mathcal{C}_{\preceq}^{\text{iso}}$ with $w_{\#}P \in \mathcal{B}$, by definition of $\Delta^{\text{iso}}(R; \mathcal{B})$, we therefore have

$$\widehat{\Delta}^{\text{iso}}(\mathcal{B}) \geq \Delta^{\text{iso}}(R; \mathcal{B}) - \varepsilon_R - 2B_R\Omega \cdot \varepsilon_{\mathcal{B}}.$$

By identical arguments, with the roles of P and $\widehat{P}_n$ reversed, we can also show that

$$\Delta^{\text{iso}}(R; \mathcal{B}) \geq \widehat{\Delta}^{\text{iso}}(\mathcal{B}) - \varepsilon_R - 2B_R\Omega \cdot \varepsilon_{\mathcal{B}},$$

which completes the proof.

C.3 Proof of Lemma 4.5

Throughout this proof we will use the notation of supervised learning, since unsupervised learning can be viewed as a special case.

In the first step, we will bound $\mathbb{E}[\varepsilon_R]$. By symmetrization (Wellner et al. (2013) Theorem 2.3.1), we have

$$\mathbb{E}[\varepsilon_R] = \mathbb{E}\left[\sup_{w\in\mathcal{C}^{\text{iso}}_{\preceq,\Omega}}\left|\mathbb{E}_{\widehat{P}_n}[(w(X)-1)r(X,Y)] - \mathbb{E}_P[(w(X)-1)R(X)]\right|\right]$$
$$\leq 2\mathbb{E}\left[\sup_{w\in\mathcal{C}^{\text{iso}}_{\preceq,\Omega}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i(w(X_i)-1)r(X_i,Y_i)\right|\right],$$

where σ_i's are independent Unif$\{\pm 1\}$ random variables. Since risk is B_R-bounded, by the Ledoux-Talagrand contraction lemma (Ledoux and Talagrand (2013) Theorem 4.12) applied with functions $\phi_i(t) = (t-1)\cdot r(X_i,Y_i)$, we further have

$$\mathbb{E}\left[\sup_{w\in\mathcal{C}^{\text{iso}}_{\preceq,\Omega}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i(w(X_i)-1)r(X_i,Y_i)\right|\right] \leq 2B_R\mathbb{E}\left[\sup_{w\in\mathcal{C}^{\text{iso}}_{\preceq,\Omega}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i w(X_i)\right|\right] = 2B_R\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}).$$

Now we bound ε_R with high probability. Since risk is B_R-bounded, and any function $w\in\mathcal{C}^{\text{iso}}_{\preceq,\Omega}$ is Ω-bounded, we have $(w(X)-1)r(X,Y)\in[-B_R,(\Omega-1)B_R]$, and so resampling one data point can perturb ε_R by at most $\Omega B_R/n$. Therefore, by McDiarmid's inequality (McDiarmid et al., 1989), with probability at least $1-n^{-1}$, it holds that

$$\varepsilon_R \leq \mathbb{E}[\varepsilon_R] + B_R\Omega\sqrt{\frac{\log n}{2n}}.$$

Combining all these calculations yields the desired bound.

C.4 Proof of Lemma 4.6

Recall that

$$\varepsilon_{\mathcal{B}} = \sup_{w\in\mathcal{C}^{\text{iso}}_{\preceq,\Omega}} \max\left\{\varepsilon_{\mathcal{B}}\left(w; P, \widehat{P}_n\right), \varepsilon_{\mathcal{B}}\left(w; \widehat{P}_n, P\right)\right\},$$

where

$$\varepsilon_{\mathcal{B}}(w; P_0, P_1) = \inf\left\{s\geq 0 \;:\; \exists\, t\geq 0,\ \left((1-s)\cdot w + t\cdot\mathbf{1}\right)_{\#}P_1\in\mathcal{B}\right\}.$$

First, following the exact same steps as in the proof of Lemma 4.5, with the notation $\delta_w = \mathbb{E}_P[w(X)] - \mathbb{E}_{\widehat{P}_n}[w(X)]$, we have

$$\sup_{w\in\mathcal{C}^{\text{iso}}_{\preceq,\Omega}} |\delta_w| \leq 4\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}) + \Omega\sqrt{\frac{\log n}{2n}} =: \varepsilon' \tag{28}$$

with probability at least $1-n^{-1}$.

Assume the event (28) holds. Fix any $w\in\mathcal{C}^{\text{iso}}_{\preceq,\Omega}$ with $w_{\#}P\in\mathcal{B}_{a,b}$, and define

$$w' = (1-s)\cdot w + t\cdot\mathbf{1},$$

where $s, t \geq 0$ are chosen such that $\mathbb{E}_{\widehat{P}_n}[w'(X)] = 1$, indicating that $t = s + (1-s)\delta_w$.

If $\varepsilon' = 4\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}) + \Omega\sqrt{\frac{\log n}{2n}} > \frac{1}{2}\min\{1-a, b-1\}$, then since $\varepsilon_{\mathcal{B}} \leq 1$ holds by definition, the result of the lemma must hold trivially. Therefore we can restrict our attention to the case that

$$\varepsilon' \leq \frac{1}{2}\min\{1-a, b-1\}.$$

We can further choose

$$s = 2\max\left\{\frac{\varepsilon'}{b-1}, \frac{\varepsilon'}{1-a}\right\} \geq \max\left\{\frac{\varepsilon'}{b-1-\varepsilon'}, \frac{\varepsilon'}{1-a-\varepsilon'}\right\},$$

with which, we can verify that

$$w'(X) \leq (1-s)b + t = (1-s)(b+\delta_w) + s \leq (b+\varepsilon') + s(1-b+\varepsilon') \leq b,$$

and similarly, $w'(X) \geq a$. Therefore, we have $w'_{\#}\widehat{P}_n \in \mathcal{B}_{a,b}$.

The same construction holds with the roles of P and $\widehat{P}_n$ reversed. Therefore, we can take $\varepsilon_{\mathcal{B}} = s$, which completes the proof.

C.5 Proof of Lemma 4.7

First, following the same steps (i.e., symmetrization and contraction) as in the proof of Lemma 4.5, we have

$$\sup_{w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}} \left|\mathbb{E}_{\widehat{P}_n}[w(X)] - \mathbb{E}_P[w(X)]\right| \leq 4\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}) + \Omega\sqrt{\frac{\log n}{2n}} =: \varepsilon' \tag{29}$$

with probability at least $1 - n^{-1}$.

Moreover, denote $t^*_f = \operatorname*{argmin}_{t \in [0,\Omega]} f(t)$. We have the decomposition

$$f(t) = f(t) \cdot \mathbb{1}\{f(t) \geq t^*_f\} + f(t) \cdot \mathbb{1}\{f(t) < t^*_f\} =: f_1 + f_2,$$

where both f_1 and $-f_2$ are nondecreasing. Then, for any $g = f \circ w$ with $w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}$, we have the decomposition $g = f_1 \circ w + f_2 \circ w$, where $f_1 \circ w \in \mathcal{C}^{\text{iso}}_{\preceq}$, $-f_2 \circ w \in \mathcal{C}^{\text{iso}}_{\preceq}$, and both functions f_1, f_2 are L_Ω-Lipschitz. Then, by the Ledoux-Talagrand contraction lemma (Ledoux and Talagrand (2013) Theorem 4.12) applied with functions $\phi_i(t) = f(t)$, we have

$$\begin{aligned}\mathcal{R}_n\left(\{g = f \circ w : \ w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}\}\right) &\leq 2\mathcal{R}_n\left(\{g = f \circ w : \ w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}, \ f \text{ is nondecreasing and } L_\Omega\text{-Lipschitz}\}\right) \\ &\leq 8L_\Omega\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}).\end{aligned}$$

Hence, similar to the proof of Lemma 4.5, we have

$$\sup_{w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}} \left|\mathbb{E}_{\widehat{P}_n}[f(w(X))] - \mathbb{E}_P[f(w(X))]\right| \leq 8L_\Omega\mathcal{R}_n(\mathcal{C}^{\text{iso}}_{\preceq,\Omega}) + L_\Omega\Omega\sqrt{\frac{\log n}{2n}} =: \varepsilon'' \tag{30}$$

with probability at least $1 - n^{-1}$.

Assume events (29) and (30) both hold. Fix any $w \in \mathcal{C}^{\text{iso}}_{\preceq,\Omega}$ with $w_{\#}P \in \mathcal{B}_{f,\rho}$, and define

$$w' = (1-s) \cdot w + t \cdot \mathbf{1},$$

where $s, t \in (0,1)$ are chosen such that $\mathbb{E}_{\widehat{P}_n}[w'(X)] = 1$, which implies that $t = s + (1-s)\delta_w$.

Since f is L_Ω-Lipschitz on $[0,\Omega]$,

$$f(w'(x)) \le f\left((1-s) \cdot w(x) + s\right) + L_\Omega \cdot |t-s| \le f\left((1-s) \cdot w(x) + s\right) + L_\Omega(1-s)|\delta_w|.$$

And, since f is convex with $f(1) = 0$,

$$f\left((1-s) \cdot w(x) + s\right) \le (1-s)f(w(x)) + sf(1) = (1-s)f(w(x)).$$

Combining everything, for all x, it holds that

$$f(w'(x)) \le (1-s)f(w(x)) + L_\Omega(1-s)|\delta_w| \le (1-s)\left(f(w(x)) + L_\Omega\varepsilon'\right).$$

Hence, we have

$$\mathbb{E}_{\widehat{P}_n}[f(w'(X))] \le (1-s)\mathbb{E}_{\widehat{P}_n}[f(w(X))] + (1-s)L_\Omega\varepsilon'.$$

And by assumption, $\mathbb{E}_{\widehat{P}_n}[f(w(X))] \le \mathbb{E}_P[f(w(X))] + \varepsilon'' \le \rho + \varepsilon''$, so,

$$\mathbb{E}_{\widehat{P}_n}[f(w'(X))] \le (1-s) \cdot (\rho + \varepsilon'' + L_\Omega\varepsilon') \le \rho,$$

where the last step holds by choosing

$$s = \frac{1}{\rho}(\varepsilon'' + L_\Omega\varepsilon') \ge \frac{\varepsilon'' + L_\Omega\varepsilon'}{\rho + \varepsilon'' + L_\Omega\varepsilon'}.$$

This verifies that $w'_{\#}\widehat{P}_n \in \mathcal{B}_{f,\rho}$.

The same construction holds with the roles of P and $\widehat{P}_n$ reversed. Therefore, we can take $\varepsilon_{\mathcal{B}} = s$, which completes the proof.

C.6 The role of the isotonic constraint

The consistency bounds developed above show that, under appropriate conditions, the error in estimating $\Delta^{\text{iso}}(R;\mathcal{B})$ can be controlled whenever the appropriate Rademacher complexity terms are small. This suggests that the isotonic constraint plays an important role: essentially, the isotonic constraint induces a form of regularization, ensuring that we work with a low-complexity class of functions.

To verify this, we now present an example with the constraint set $\mathcal{B} = \mathcal{B}_{a,b}$, *without* an isotonic constraint, where the estimation error of the (non-iso) DRL risk does not converge to zero.

To make the question more concrete, we will work with the bound constraint $\mathcal{B}_{a,b}$ with $0 \le a \le 1 \le b$, and consider the optimization problem

$$\widehat{\Delta}(r;\mathcal{B}_{a,b}) \quad = \quad \max_{w \ge 0} \quad \frac{1}{n}\sum_{i \le n} w(X_i) r_i - \frac{1}{n}\sum_{i \le n} r_i \quad \text{subject to} \quad w_{\#}\widehat{P}_n \in \mathcal{B}_{a,b},$$

which estimates the excess risk without the isotonic constraint. In other words, using $\widehat{\Delta}(r;\mathcal{B}_{a,b})$ as an empirical estimate of $\Delta(R;\mathcal{B}_{a,b})$, is analogous to using $\widehat{\Delta}^{\text{iso}}(r;\mathcal{B}_{a,b})$ as an empirical estimate of $\Delta^{\text{iso}}(R;\mathcal{B}_{a,b})$ in the presence of an additional isotonic constraint.

The following result shows that, without an isotonic constraint, this empirical estimate is *not* a consistent estimator of the true excess risk.

Proposition C.1. *Assume $R(X) = 1/2$ holds P-almost surely. Then, $\Delta(R;\mathcal{B}_{a,b}) = 0$, but with probability at least $1 - 2e^{-n/24}$, it holds that $\widehat{\Delta}(r;\mathcal{B}_{a,b}) \geq \min\{1-a, b-1\}/16$.*

In other words, $\widehat{\Delta}(r;\mathcal{B}_{a,b})$ is not a consistent estimator of the true excess risk $\Delta(R;\mathcal{B}_{a,b})$, since the error in the estimate is bounded away from zero (as long as $a < 1 < b$). This means that the constraint set $\mathcal{B}_{a,b}$, on its own, is not sufficiently constrained to enable consistent estimation—while in contrast, as we have seen in our theoretical guarantees for estimation for iso-DRL, adding an isotonic constraint enables the excess risk to be estimated consistently with an empirical sample.

C.6.1 Proof of Proposition C.1

By construction, we have $r_i \sim \text{Bern}(R(X_i)) = \text{Bern}(1/2)$ independently for $i = 1, \ldots, n$. According to Section 2, the worst-case weights take the form $w_i = w(X_i) = c_1 \cdot \mathbb{1}\{r_i = 0\} + c_2 \cdot \mathbb{1}\{r_i = 1\}$, where $a \leq c_1 \leq 1 \leq c_2 \leq b$. Moreover, by the KKT condition, at least one of $c_1 = a$ and $c_2 = b$ holds, which implies that $c_2 - c_1 \geq \min\{1-a, b-1\} =: \delta$. Then, the estimated excess risk can be expressed as

$$\widehat{\Delta}(r;\mathcal{B}_{a,b}) = \frac{c_2}{n}\sum_{i\leq n} r_i - \frac{1}{n}\sum_{i\leq n} r_i = \frac{c_2 - 1}{n}\sum_{i\leq n} r_i.$$

Since $n^{-1}\sum_{i\leq n} w_i = 1$, we have

$$\frac{1}{n}\sum_{i\leq n}(1 - r_i) = \frac{c_2 - 1}{c_2 - c_1},$$

which implies

$$c_2 - 1 = \frac{c_2 - c_1}{n}\sum_{i\leq n}(1 - r_i) \geq \frac{\delta}{n}\sum_{i\leq n}(1 - r_i)\,.$$

In the meantime, by Chernoff bounds, with probability at least $1 - 2e^{-n/24}$, it holds that

$$\left|\frac{1}{n}\sum_{i\leq n} r_i - \frac{1}{2}\right| \leq \frac{1}{4}.$$

Then, for the excess risk, with probability at least $1 - 2e^{-n/24}$, it holds that

$$\widehat{\Delta}(r;\mathcal{B}_{a,b}) = \frac{c_2 - 1}{n}\sum_{i\leq n} r_i \geq \delta\left(\frac{1}{n}\sum_{i\leq n}(1 - r_i)\right)\cdot\left(\frac{1}{n}\sum_{i\leq n} r_i\right) \geq \frac{\delta}{16}.$$

D Additional simulation results

D.1 Simulations for iso-DRL under componentwise order

In Section 5, we mainly focused on the partial order with respect to $w_0(x)$. In this section, to demonstrate the effect of various choices of the partial (pre)order, we further consider an alternative choice of the partial (pre)order: the componentwise order where

$$x \preceq x' \qquad \text{if and only if} \qquad x_j \leq x'_j, \text{ for all } j \in [m],$$

where we set $m = 5 < d = 20$. Let iso-DRL-comp denote the CP interval with calibrated target level $\alpha'_{\text{iso}} = \max\{0, \alpha - \widetilde{\Delta}^{\text{iso}}\}$, where

$$\begin{aligned} \widetilde{\Delta}^{\text{iso}} = \max \quad & \frac{1}{n} \sum_{i \in \mathcal{D}_3} w_i \widetilde{r}_i^{\text{iso}} - \frac{1}{n} \sum_{i \in \mathcal{D}_3} r_i \\ \text{subject to} \quad & \frac{1}{n} \sum_{i \in \mathcal{D}_3} w_i = 1, \quad \frac{1}{n} \sum_{i \in \mathcal{D}_3} w_i \log w_i \leq \rho, \quad 0 \leq w_i \leq \Omega, \end{aligned} \tag{31}$$

and $(\widetilde{r}_i)_{i \in \mathcal{D}_3}$ is the isotonic projection of $(r_i)_{i \in \mathcal{D}_3}$ with respect to the componentwise order.

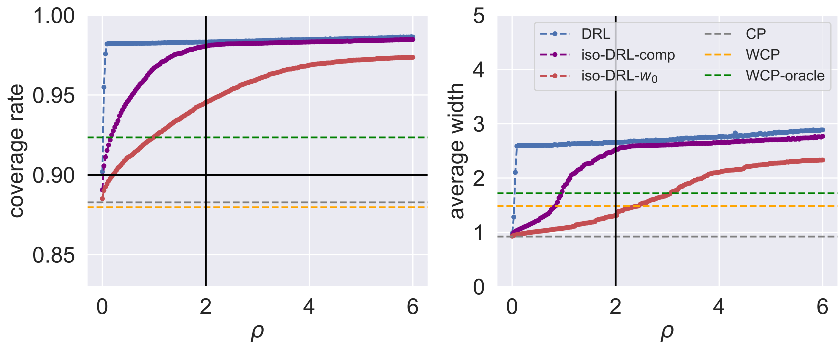


Figure 6: Results with varying ρ in the well-specified setting. The solid vertical line denotes an estimate $\widehat{\rho}$ of the KL divergence, $D_{\text{KL}}(P_{\text{target}} \| P)$ (See Appendix D.2 for details). The solid horizontal line (in the left-hand plot) marks the nominal coverage level, $1 - \alpha = 90\%$.

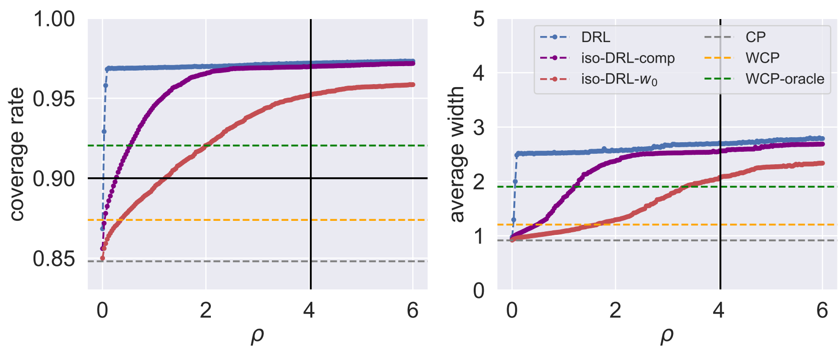


Figure 7: Results with varying ρ in the misspecified setting. The solid vertical line denotes an estimate $\widehat{\rho}$ of the KL divergence, $D_{\text{KL}}(P_{\text{target}} \| P)$ (See Appendix D.2 for details). The solid horizontal line (in the left-hand plot) marks the nominal coverage level, $1 - \alpha = 90\%$.

We follow exactly the same settings with Section 5.1 with $n_{\text{pre}} = 50$ and vary ρ in $[0.002, 6]$. From Figure 6 and 7, each of the coverage rate and average interval width of iso-DRL-comp lies between that of DRL and iso-DRL-w_0, which indicates that additional constraints will relieve the conservativeness of DRL, but only a proper choice of the partial (pre)order will lead to desired performance close to the oracle weighted CP.

D.2 Details for the `wine quality` data set: a proxy of the oracle KL-divergence

In this section, we examine the choice of ρ in the `wine quality` data experiment from Section 5.2. In a real data setting, the true KL divergence, $D_{\mathrm{KL}}\left(P_{\text{target}} \| P\right)$, is of course unknown, so we need to use a data-driven choice of ρ in order to implement a DRL procedure (with or without an isotonic constraint).

As is shown in Section 5.2, we denote $\widehat{w}_{\text{kde}}$ as the density ratio obtained by kernel density estimation (Gaussian kernel with bandwidth 0.125). Accordingly, let $\mathsf{d}\widehat{Q}_{\text{kde}} = \widehat{w}_{\text{kde}} \cdot \mathsf{d}P$ be an estimate of P_{target}. With a subsample $\{X_i\}_{i \leq K}$ drawn the group of white wine (data distribution P), a reasonable value for $\widehat{\rho}$ (i.e., an estimate of the true divergence ρ between the distributions P and P_{target}) can be calculated by

$$\begin{aligned} \widehat{\rho} &= \frac{1}{K} \sum_{i \leq K} \widehat{w}_{\text{kde}}(X_i) \log\left(\widehat{w}_{\text{kde}}(X_i)\right) \\ &\approx \mathbb{E}_P \left\{ \frac{\mathsf{d}\widehat{Q}_{\text{kde}}}{\mathsf{d}P} \log\left(\frac{\mathsf{d}\widehat{Q}_{\text{kde}}}{\mathsf{d}P}\right) \right\} = D_{\mathrm{KL}}\left(\widehat{Q}_{\text{kde}} \| P\right). \end{aligned}$$

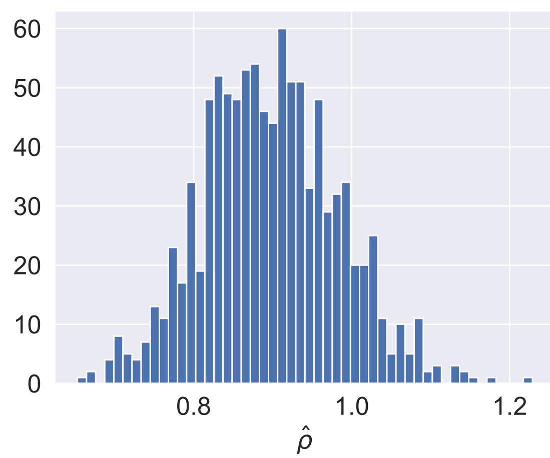


Figure 8: Histogram of $\widehat{\rho}$. (See Appendix D.2 for details.)

To show the range for values of $\widehat{\rho}$, we repeatedly fit KDE on the 80% samples from each group (white and red wine groups respectively). Figure 8 shows the histogram of $\widehat{\rho}$ with 1000 repetitions, of which the median is approximately 0.8950—this is the value of ρ used in our preview of the `wine quality` data experiment, shown in Figure 1.

References

Ai, J. and Ren, Z. (2024). Not all distributional shifts are equal: Fine-grained robust conformal inference. *arXiv preprint arXiv:2402.13042*.

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142.

Bastani, H. (2021). Predicting with proxies: Transfer learning in high dimension. *Manag. Sci.*, 67:2964–2984.

Bauschke, H. and Combettes, P. (2019). Convex analysis and monotone operator theory in hilbert spaces, corrected printing.

Ben-David, S., Lu, T., Luu, T., and Pál, D. (2010). Impossibility theorems for domain adaptation. In *AISTATS*.

Ben-David, S. and Urner, R. (2012). On the hardness of domain adaptation and the utility of unlabeled target samples. In *ALT*.

Ben-Tal, A. and Nemirovski, A. (1998). Robust convex optimization. *Mathematics of operations research*, 23(4):769–805.

Berta, E., Bach, F., and Jordan, M. (2024). Classifier calibration with roc-regularized isotonic regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR.

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

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Blanchet, J. and Shapiro, A. (2023). Statistical limit theorems in distributionally robust optimization. *arXiv preprint arXiv:2303.14867*.

Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217.

Brunk, H. (1963). On an extension of the concept conditional expectation. *Proceedings of the American Mathematical Society*, 14(2):298–304.

Brunk, H. (1965). Conditional expectation given a σ-lattice and applications. *The Annals of Mathematical Statistics*, 36(5):1339–1350.

Brunk, H., Barlow, R. E., Bartholomew, D. J., and Bremner, J. M. (1972). Statistical inference under order restrictions.(the theory and application of isotonic regression). *International Statistical Review*, 41:395.

Brunk, H., Ewing, G., and Utz, W. (1957). Minimizing integrals in certain classes of monotone functions. *Pacific Journal of Mathematics*.

Cai, T. T. and Wei, H. (2019). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *ArXiv*, abs/1906.02903.

Candès, E., Lei, L., and Ren, Z. (2023). Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):24–45.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2020). Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*.

Chatterjee, S. and Lafferty, J. (2019). Adaptive risk bounds in unimodal regression. *Bernoulli*.

Chattopadhyay, R., Fan, W., Davidson, I., Panchanathan, S., and Ye, J. (2013). Joint transfer and batch-mode active learning. In *ICML*.

Chen, M., Weinberger, K. Q., and Blitzer, J. (2011). Co-training for domain adaptation. In *NIPS*.

Chen, Y. and Lei, J. (2024). De-biased two-sample u-statistics with application to conditional distribution testing. *arXiv preprint arXiv:2402.00164*.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203.

Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. (2008). Sample selection bias correction theory. *ArXiv*, abs/0805.2775.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553.

De Bartolomeis, P., Abad, J., Donhauser, K., and Yang, F. (2023). Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. *arXiv preprint arXiv:2312.03871*.

Deng, H. and Zhang, C.-H. (2020). Isotonic regression in multi-dimensional spaces and graphs.

Ding, P. and VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology (Cambridge, Mass.)*, 27(3):368.

Donsker, M. D. and Varadhan, S. S. (1976). Asymptotic evaluation of certain markov process expectations for large time—iii. *Communications on pure and applied Mathematics*, 29(4):389–461.

Duchi, J. and Namkoong, H. (2018). Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*.

Duchi, J. C., Glynn, P. W., and Namkoong, H. (2021). Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969.

Duchi, J. C., Hashimoto, T., and Namkoong, H. (2019). Distributionally robust losses against mixture covariate shifts. *Under review*, 2(1).

Dudley, R. M. (1967). The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330.

Durot, C. and Lopuhaä, H. P. (2018). Limit Theory in Monotone Function Estimation. *Statistical Science*, 33(4):547 – 567.

Edwards, R. E. (2012). *Functional analysis: theory and applications*. Courier Corporation.

El Ghaoui, L. and Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications*, 18(4):1035–1064.

El Ghaoui, L., Oustry, F., and Lebret, H. (1998). Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization*, 9(1):33–52.

Esfahani, P. M. and Kuhn, D. (2015). Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*.

Esteban-Pérez, A. and Morales, J. M. (2022). Partition-based distributionally robust optimization via optimal transport with order cone constraints. *4OR*, 20(3):465–497.

Gao, F. and Wellner, J. A. (2007). Entropy estimate for high-dimensional monotonic functions. *Journal of Multivariate Analysis*, 98(9):1751–1764.

Ge, J., Tang, S., Fan, J., Ma, C., and Jin, C. (2023). Maximum likelihood estimation is all you need for well-specified covariate shift. *arXiv preprint arXiv:2311.15961*.

Grenander, U. (1956). On the theory of mortality measurements. *Skandinavisk Aktuarietidskrift*, 39:1–55.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K. M., Schölkopf, B., Candela, Q., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). Covariate shift by kernel mean matching. In *NIPS 2009*.

Grotzinger, S. J. and Witzgall, C. (1984). Projections onto order simplexes. *Applied mathematics and Optimization*, 12(1):247–270.

Gui, Y., Barber, R., and Ma, C. (2024). Conformalized matrix completion. *Advances in Neural Information Processing Systems*, 36.

Gui, Y., Hore, R., Ren, Z., and Barber, R. F. (2023). Conformalized survival analysis with adaptive cut-offs. *Biometrika*, page asad076.

Han, Q., Wang, T., Chatterjee, S., and Samworth, R. J. (2019). Isotonic regression in general dimensions. *The Annals of Statistics*.

Hanneke, S. and Kpotufe, S. (2019). On the value of target data in transfer learning. In *NeurIPS*.

Hardy, G. H., Littlewood, J. E., and Pólya, G. (1952). *Inequalities*. Cambridge university press.

Henzi, A., Ziegel, J. F., and Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):963–993.

Huang, M. and Pimentel, S. D. (2024). Variance-based sensitivity analysis for weighting estimators results in more informative bounds. *Biometrika*, page asae040.

Jin, Y., Ren, Z., and Candès, E. J. (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120.

Jin, Y., Ren, Z., and Zhou, Z. (2022). Sensitivity analysis under the f-sensitivity models: a distributional robustness perspective. *arXiv preprint arXiv:2203.04373*.

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

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Lei, L. and Candès, E. J. (2020). Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*.

Li, S., Cai, T. T., and Li, H. (2021). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Liu, J., Wu, J., Wang, T., Zou, H., Li, B., and Cui, P. (2023). Geometry-calibrated dro: Combating over-pessimism with free energy implications. *arXiv preprint arXiv:2311.05054*.

Ma, C., Pathak, R., and Wainwright, M. J. (2023). Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761.

Matzkin, R. L. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica: Journal of the Econometric Society*, pages 1315–1327.

McDiarmid, C. et al. (1989). On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188.

Megginson, R. E. (2012). *An introduction to Banach space theory*, volume 183. Springer Science & Business Media.

Namkoong, H. and Duchi, J. C. (2017). Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30.

Namkoong, H., Ma, Y., and Glynn, P. W. (2022). Minimax optimal estimation of stability under distribution shift. *arXiv preprint arXiv:2212.06338*.

Niculescu-Mizil, A. and Caruana, R. A. (2012). Obtaining calibrated probabilities from boosting. *arXiv preprint arXiv:1207.1403*.

Nie, X., Imbens, G., and Wager, S. (2021). Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*.

Pathak, R. and Ma, C. (2024). On the design-dependent suboptimality of the lasso. *arXiv preprint arXiv:2402.00382*.

Pathak, R., Ma, C., and Wainwright, M. (2022). A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pages 17517–17530. PMLR.

Popescu, I. (2007). Robust mean-covariance solutions for stochastic optimization. *Operations Research*, 55(1):98–112.

Qiu, H., Dobriban, E., and Tchetgen Tchetgen, E. (2023). Prediction sets adaptive to unknown covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad069.

Rao, B. P. (1969). Estimation of a unimodal density. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 23–36.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2020). A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv: Learning*.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rothenhäusler, D. and Bühlmann, P. (2023). Distributionally robust and generalizable inference. *Statistical Science*, 38(4):527–542.

Sahoo, R., Lei, L., and Wager, S. (2022). Learning from a biased sample. *arXiv preprint arXiv:2209.01754*.

Setlur, A., Dennis, D., Eysenbach, B., Raghunathan, A., Finn, C., Smith, V., and Levine, S. (2023). Bitrate-constrained dro: Beyond worst case robustness to unknown group shifts. *arXiv preprint arXiv:2302.02931*.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Shafieezadeh Abadeh, S., Mohajerin Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 28.

Shapiro, A. (2017). Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275.

Shapiro, A. and Pichler, A. (2023). Conditional distributionally robust functionals. *Operations Research*.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Tian, Y. and Feng, Y. (2021). Transfer learning under high-dimensional generalized linear models. *ArXiv*, abs/2105.14328.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *NeurIPS*.

van der Laan, L., Ulloa-Pérez, E., Carone, M., and Luedtke, A. (2023). Causal isotonic calibration for heterogeneous treatment effects. *arXiv preprint arXiv:2302.14011*.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Wang, Z., Bühlmann, P., and Guo, Z. (2023). Distributionally robust machine learning with multi-source data. *arXiv preprint arXiv:2309.02211*.

Weiss, A., Lancho, A., Bu, Y., and Wornell, G. W. (2023). A bilateral bound on the mean-square error for estimation in model mismatch. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 2655–2660. IEEE.

Wellner, J. et al. (2013). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. (2018). Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*.

Yang, F. and Barber, R. F. (2019). Contraction and uniform convergence of isotonic regression. *Electronic Journal of Statistics*.

Yang, L., Hanneke, S., and Carbonell, J. G. (2012). A theory of transfer learning with applications to active learning. *Machine Learning*, 90:161–189.

Yang, Y., Kuchibhotla, A. K., and Tchetgen Tchetgen, E. (2024). Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae009.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555.

Zhao, Q., Small, D. S., and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761.