Not All Distributional Shifts Are Equal: Fine-Grained Robust Conformal Inference

Jiahao Ai [1] **Zhimei Ren** [2]

Abstract

We introduce a fine-grained framework for uncertainty quantification of predictive models under distributional shifts. This framework distinguishes the shift in covariate distributions from that in the conditional relationship between the outcome (Y) and the covariates (X). We propose to reweight the training samples to adjust for an identifiable shift in covariate distribution while protecting against the worst-case conditional distribution shift bounded in an f-divergence ball. Based on ideas from conformal inference and distributionally robust learning, we present an algorithm that outputs (approximately) valid and efficient prediction intervals in the presence of distributional shifts. As a use case, we apply the framework to sensitivity analysis of individual treatment effects with hidden confounding. The proposed methods are evaluated in simulations and four real data applications, demonstrating superior robustness and efficiency compared with existing benchmarks.

1. Introduction

It has been widely observed that the performance of predictive models falls short of expectation when generalized to a population whose distribution differs from that of the training data (see e.g., Recht et al. (2019); Miller et al. (2020); Wong et al. (2021); Namkoong et al. (2023); Liu et al. (2023) and the references therein). As predictive models are increasingly employed in high-stakes settings, it is imperative to accompany the predicted outcomes with calibrated uncertainty quantification when deploying them to new environments. A widely adopted approach to uncertainty quantification is to provide a prediction set that contains the true outcome with high probability. The prediction set informs the confidence we have in the predicted outcome.

Among the tools for constructing prediction sets, conformal prediction (CP) (Vovk et al., 2005) is an attractive framework that generates valid prediction sets that are guaranteed to include the true outcome with pre-specified probability. The validity of CP holds for *any* predictive model, as long as the training and test data are exchangeable, e.g., when they are identically independently distributed (i.i.d.). In the presence of distributional shifts, however, the exchangeability/i.i.d. assumption breaks, and CP no longer delivers valid prediction sets. To address this challenge, prior work (Cauchois et al., 2023) proposes a robust CP method that outputs prediction sets that are valid when the target distribution ranges within a neighborhood of the training distribution. To be more specific, let $X \in \mathcal{X}$ denote the covariates and $Y \in \mathcal{Y}$ the outcome/response. Consider a training set of n samples $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ and an independent test unit $(X_{n+1}, Y_{n+1}) \sim Q_{X,Y}$, where we only get to observe X_{n+1} and wish to predict Y_{n+1}. Cauchois et al. (2023) assumes that the f-divergence between $Q_{X,Y}$ and $P_{X,Y}$ is bounded by a parameter ρ, and provides prediction sets that ensure guarantees even for the worst-case $Q_{X,Y}$.

The method of Cauchois et al. (2023) provides robustness against the worst-case *joint* distributional shift of (X, Y), but no distinction is made between the shift in X and that in $Y \mid X$. As pointed out by a recent line of research, different types of distributional shifts appear in different tasks and result in different consequences (Mu et al., 2022; Namkoong et al., 2023; Jin et al., 2023a; Liu et al., 2023). Without separating the sources of distributional shifts and taking specialized treatment, the method of Cauchois et al. (2023) can be overly conservative in practice (to be demonstrated shortly). In this work, we take a closer look at distributional shift, and provide a fine-grained robust predictive inference approach with improved efficiency.

1.1. Decomposing the Distributional Shifts

We decompose the distributional shifts into two types:

(1) *The X shift:* the marginal distribution of X is different in the training and target environment. For example,

[1] School of Mathematical Sciences, Peking University, Beijing, China [2] Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, USA. Correspondence to: Zhimei Ren <zren@wharton.upenn.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

the age/gender structure in the new environment differs from that in the training environment.

(2) *The $Y \mid X$ shift:* the conditional relationship between the outcome and the the covariate is different in the training and target environment. This could happen due to unobserved confounders, or when the training and target data are collected from different periods and the conditional relationship varies over time.

The above two types of distributional shifts are different in nature — for one thing, the former type of distributional shift is *identifiable* but the latter is not. In most cases, the distributional shift is a mixture of the two. Instead of guarding against the worst-case joint distributional shift, we propose to tease apart the two types of shifts, reweighting the training samples according to the estimated X shift and adjusting the confidence level to account for the worst-case $Y \mid X$ shift. Specifically, we assume the $Y \mid X$ shift to be bounded in the f-divergence, i.e., $D_f(Q_{Y \mid X} \,\|\, P_{Y \mid X}) \leq \rho$, but posit no constraints on the X shift. For such distributional shifts, our proposed method aims to construct a prediction interval $\widehat{C}_{f,\rho}(X_{n+1})$ with the training data, such that it covers the true outcome with high probability under the target distribution.

1.2. Our Contributions

This work introduces a new framework for calibrated uncertainty quantification in the presence of distributional shifts. Toward this end, we make the following contributions:

(1) We present *Weighted Robust Conformal Prediction (WRCP)*, which treats the X shift and the $Y \mid X$ shift differently. It (approximately) achieves the desired coverage under the proposed framework, with the miscoverage rate determined by the estimation error of the covariate likelihood ratio $\mathrm{d}Q_X/\mathrm{d}P_X$.

Technically, WRCP builds upon the observation that the "worst-case quantile inflation" does not depend on the observed distribution, so it suffices to use the same adjusted confidence level across different values of X. This observation allows us to generalize prior works.

(2) In the case when estimating the X shift is challenging (e.g., when X is high-dimensional), we propose a debiased variant of WRCP, namely D-WRCP, which enjoys the double-robustness property — its miscoverage rate depends on the product of the estimation error of $\mathrm{d}Q_X/\mathrm{d}P_X$ and that of conditional quantiles of the residuals from predicting the outcomes. This relaxes the requirement for estimating the covariate likelihood ratio (which is typically challenging in high dimensions), and improves upon earlier works (e.g., Tibshirani et al. (2019); Hu and Lei (2023)) that mostly apply to low-dimensional data.

(3) As a special example, we show that our proposed methods can be adapted to conducting sensitivity analysis for individual treatment effects (ITEs) under the f-sensitivity model (Jin et al., 2022).

(4) We empirically evaluate the proposed methods in simulations and four real data applications, demonstrating their validity and improved efficiency.

1.3. Related Literature

Conformal Prediction beyond Exchangeability. With exchangeable/i.i.d. data, there is a long list of works on the theoretical property, efficient implementation and application of conformal prediction (see e.g., Vovk et al. (2005); Papadopoulos et al. (2002); Lei et al. (2018); Romano et al. (2019); Barber et al. (2021); Angelopoulos et al. (2023)).

Beyond exchangeability, Tibshirani et al. (2019); Park et al. (2022) consider the pure covariate shift setting, with the former focusing on the marginal coverage guarantee and the latter the training-conditional guarantee; also under the pure covariate shift setting, Qiu et al. (2022); Yang et al. (2022) builds upon semi-parametric theory to develop more efficient CP methods with asymptotic coverage guarantees. The de-biased version of our proposal draws inspiration from these two works, and we generalize them to the specific distributional shift model under consideration. Podkopaev and Ramdas (2021); Si et al. (2023) tackles the label shift setting, where the marginal distribution of Y is subject to changes but $X \mid Y$ remains invariant in the training and target distribution. The work of Barber et al. (2023) addresses a general form of distribution shift by up-weighting training points whose distribution is closer to that of the target distribution (the weights need to be independent of the data).

As mentioned earlier, (Cauchois et al., 2023) is concerned with robust CP against the worst-case joint shift in (X, Y). Gendler et al. (2021); Ghosh et al. (2023) investigate the robustness of CP under adversarial attacks. Another two closely related works on robust CP are Jin et al. (2023b); Yin et al. (2022), which study sensitivity analysis of ITEs under the marginal Γ-selection model (Tan, 2006); the type of distributional shift (caused by hidden confounding) puts no requirements on the X shift and assumes that the shift in $Y \mid X$ is uniformly bounded by constants, i.e., $1/\Gamma \leq \frac{\mathrm{d}Q_{Y \mid X}}{\mathrm{d}P_{Y \mid X}} \leq \Gamma$. Compared with our model that assumes the $Y \mid X$ shift to be bounded *on average* (the f-divergence takes the expectation over Y), the point-wise bound requires the *maximum* shift to be bounded, which can sometimes be conservative in practice (see more discussion and examples in Jin et al. (2022)).

Distributionally Robust Learning. Distributionally robust learning studies the broad topic of learning from data with guarantees under the worst-case distributional shift within a specified set of distributions. Typical tasks in this field includes parameter estimation (Shafieezadeh Abadeh et al., 2015; Blanchet and Murthy, 2019; Duchi and Namkoong, 2021; Duchi et al., 2023), policy learning (Si et al., 2023; Mu et al., 2022; Zhang et al., 2023), among others. In particular, Mu et al. (2022) proposes learning a robust policy by separately considering X shifts and $Y \mid X$ shifts, echoing the proposal in this paper. Compared with the existing literature, our work takes a different angle by studying the uncertainty quantification problem under distributional shifts.

Sensitivity Analysis. In causal inference, distributional shifts can arise due to unobserved confounders, and sensitivity analysis is a standard tool for assessing the robustness of causal effect estimates under such shifts. Under the aforementioned (marginal) Γ-selection model (Rosenbaum, 1987; Tan, 2006), a line of papers (Zhao et al., 2019; Yadlowsky et al., 2018; Kallus and Zhou, 2020; Sahoo et al., 2022) study the estimation of the average treatment effect (ATE) or the policy values, and the work of Kallus and Zhou (2021); Lei et al. (2023) consider learning the optimal policy. Recently, (Jin et al., 2022) proposes the f-sensitivity model, and discusses how to estimate the ATE under the model. We shall show later in this paper that the distributional shift under the f-sensitivity model fits exactly in our framework, and hence our proposed method can be adopted there for the uncertainty quantification of ITEs.

2. Problem Setup and Background

Consider a training data set $\mathcal{D}_{\text{tr}} = \{(X_i, Y_i)\}_{i=1}^n$, where $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$. For a test unit $(X_{n+1}, Y_{n+1}) \sim Q_{X,Y}$, for which only the covariate is observed, we aim at using $\mathcal{D}_{\text{tr}}$ to construct an interval $\widehat{C}(X_{n+1})$ such that

$$\mathbb{P}_{(X_{n+1},Y_{n+1})\sim Q_{X,Y}}\left(Y_{n+1} \in \widehat{C}(X_{n+1})\right) \geq 1-\alpha, \quad (1)$$

where the probability is taken over the randomness of $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ and $(X_{n+1}, Y_{n+1}) \sim Q_{X,Y}$, and $\alpha \in (0,1)$ is the pre-specified mis-coverage level.

Let $s : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ denote a score function, and we define for each $i \in [n] = \{1, 2, \ldots, n\}$ the nonconformity score $S_i = s(X_i, Y_i)$. For example, when $\widehat{\mu}(x)$ is a fitted function for the conditional mean of $Y \mid X$, one can take $s(x, y) = |y - \widehat{\mu}(x)|$.[1] For other types of nonconformity scores, see also Romano et al. (2019); Chernozhukov et al. (2021); Guan (2023); Gupta et al. (2022).

2.1. Characterizing the Distributional Shifts

As introduced earlier, the distributional shift between $P_{X,Y}$ and $Q_{X,Y}$ can be originating from two sources: (1) the difference between P_X and Q_X and (2) the difference between $P_{Y \mid X}$ and $Q_{Y \mid X}$. The two types of distribution shifts are different in nature: often the X shift is observable and estimable — since we have access to the covariates in the test set — while the $Y \mid X$ shift is not identifiable. Based on this observation, we propose distinct treatments to these two types of distributional shift.

The X shift is represented by the likelihood ratio $w(x) = \frac{dQ_X}{dP_X}(x)$. We do not posit any assumption on $w(x)$ (except that Q_X is absolutely continuous with respect to P_X), and shall use data to estimate this quantity when it is not known a priori. For the conditional distributional shift, we assume that the target distribution $Q_{Y \mid X}$ falls within a "neighborhood ball" of $P_{Y \mid X}$, whose radius is controlled by a parameter ρ. The neighborhood ball is formalized by the f-divergence, defined below.

Definition 2.1 (f-divergence)**.** Let P and Q be two probability distributions over a space Ω such that P is absolutely continuous with respect to Q. For a convex function f such that $f(1) = 0$, the f-divergence of P from Q is defined as $D_f(P \,\|\, Q) = \mathbb{E}_Q[f(dP/dQ)]$, where dP/dQ is the Radon-Nikodym derivative.

Throughout, we assume f to be closed and convex, with $f(1) = 0$ and $f(x) < +\infty$ for $x > 0$. Common choices of f include $f(x) = x \log x$, which yields the Kullback–Leibler (KL) divergence, $f(x) = \frac{1}{2}|x-1|$ that yields the total variation (TV) distance, and $f(x) = (x-1)^2$ that yields the Pearson χ^2-divergence. Cauchois et al. (2023) argues that different choice of f determines the different types of control on the tail performance. The users can choose f based on the target confidence level, and how they want to weigh ρ. We will also evaluate the role of f in our method through simulations (Appendix E).

With the target conditional distribution of $Y \mid X$ satisfying $D_f(Q_{Y \mid X=x} \,\|\, P_{Y \mid X=x}) \leq \rho$, for P_X-almost all x, we can define the set of possible $Q_{X,Y}$ as

$$\mathcal{P}(\rho; P) := \Big\{Q \text{ s.t. } (X_{n+1}, Y_{n+1}) \sim Q : \\ D_f\big(Q_{Y \mid X=x} \,\|\, P_{Y \mid X=x}\big) \leq \rho, \text{ for } P_X\text{-almost all } x\Big\}.$$

Below, we refer to $\mathcal{P}(\rho; P)$ as the identification set.

2.2. Split Conformal Prediction

When $P_{X,Y} = Q_{X,Y}$, the method of conformal inference offers an elegant solution for achiving (1) by leveraging the

[1] Strictly, we should write the score function as $s(x, y; \widehat{\mu})$ as it also depends on $\widehat{\mu}$. For notational simplicity, we suppress the dependence on $\widehat{\mu}$ (or other predictive functions) in the score function when the context is clear.

exchangeability among $\{(X_i, Y_i)\}_{i=1}^{n+1}$. In particular, the split conformal inference (Vovk et al., 2005; Papadopoulos et al., 2002) is a computationally efficient variant of conformal inference that begins by randomly splitting the training data into two folds, $\mathcal{D}_{\text{tr}}^{(0)}$ and $\mathcal{D}_{\text{tr}}^{(1)}$, where $n_0 = |\mathcal{D}_{\text{tr}}^{(0)}|$ and $n_1 = |\mathcal{D}_{\text{tr}}^{(1)}|$. It then uses $\mathcal{D}_{\text{tr}}^{(0)}$ for fitting the prediction function $\widehat{\mu} : \mathcal{X} \mapsto \mathbb{R}$ and $\mathcal{D}_{\text{tr}}^{(1)}$ for obtaining the estimated quantile of S_{n+1}. The prediction interval takes the form

$$\widehat{C}(X_{n+1}) = \Big\{y \in \mathbb{R} : s(X_{n+1}, y) \leq \text{Quantile}\Big(1-\alpha, \{S_i\}_{i \in \mathcal{D}_{\text{tr}}^{(1)}} \cup \{\infty\}\Big)\Big\}, \tag{2}$$

where Quantile$(\beta, \{Z_i\}_{i=1}^n)$ denotes the $\lceil n\beta \rceil$-th smallest element among $Z_1, Z_2, \ldots, Z_n$. The prediction interval (2) guarantees that $\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1})) \geq 1-\alpha$ when $P_{X,Y} = Q_{X,Y}$ without any additional assumptions (Vovk et al., 2005); if the ties among the nonconformity scores happen with probability zero, then the coverage is also tight (Lei et al., 2018), i.e., $\mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1})) \leq 1 - \alpha + 1/n_1$.

3. Methodology

In this section, we describe how to generalize (split) conformal inference to efficiently handle distributional shift. At a high level, we shall reweight the training samples according to the X shift and adjust the confidence level to account for the worst conditional distributional shift. When the X shift is unknown, we further discuss how to estimate it from the data, and provide a debiased version of the method to alleviate the effect of such estimation error.

Throughout, we fix the radius of the identification set $\rho > 0$. As in the standard split conformal prediction, we start by splitting the training set into two folds, $\mathcal{D}_{\text{tr}}^{(0)}$ and $\mathcal{D}_{\text{tr}}^{(1)}$. The fitting fold $\mathcal{D}_{\text{tr}}^{(0)}$ is used for fitting the prediction function $\widehat{\mu}$ (or other functions depending on the type of nonconformity score). The calibration fold $\mathcal{D}_{\text{tr}}^{(1)}$ is devoted to finding the largest quantile of S_{n+1} for $Q \in \mathcal{P}(\rho; P)$. To this end, we follow Cauchois et al. (2023), defining $g_{f,\rho}(\beta) := \inf\big\{z \in [0,1] : \beta f\big(\frac{z}{\beta}\big) + (1-\beta) f\big(\frac{1-z}{1-\beta}\big) \leq \rho\big\}$, and its inverse $g_{f,\rho}^{-1}(\tau) := \sup\{\beta \in [0,1] : g_{f,\rho}(\beta) \leq \tau\}$.

Recall that $w(x) = \frac{dQ_X}{dP_X}(x)$. Our prediction set is

$$\widehat{C}_{f,\rho}(x) = \Big\{y \in \mathbb{R} : s(x,y) \leq \text{Quantile}\Big(g_{f,\rho}^{-1}(1-\alpha), \sum_{i \in \mathcal{D}_{\text{tr}}^{(1)}} p_i(x) \cdot \delta_{S_i} + p_{n+1}(x) \cdot \delta_{\infty}\Big)\Big\}, \tag{3}$$

where $p_i(x) = \frac{w(X_i)}{\sum_{j \in \mathcal{D}_{\text{tr}}^{(1)}} w(X_j) + w(x)}$, and $p_{n+1}(x) = \frac{w(x)}{\sum_{j \in \mathcal{D}_{\text{tr}}^{(1)}} w(X_j) + w(x)}$. In words, we upper bound the $(1-\alpha)$-th quantile of S_{n+1} under Q by a weighted quantile under P at a slightly inflated level. The validity of $\widehat{C}_{f,\rho}(X_{n+1})$ is formalized by Theorem 3.1, whose proof is deferred to Appendix B.1.

Theorem 3.1 (Prediction interval with known X shift). *Assume the training data $\{(X_i, Y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} P_{X,Y}$ and $(X_{n+1}, Y_{n+1}) \sim Q_{X,Y}$ is independent of $\{(X_i, Y_i)\}_{i=1}^n$. Assume that Q is absolutely continuously continuous with respect to P, and denote $w(x) = \frac{dQ_X}{dP_X}(x)$. For $\alpha \in (0,1)$, the prediction set $\widehat{C}_{f,\rho}(X_{n+1})$ defined in (3) satisfies that*

$$\mathbb{P}\big(Y_{n+1} \in \widehat{C}_{f,\rho}(X_{n+1})\big) \geq g_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big).$$

Furthermore, if $g_{f,\rho}(1) \geq 1-\alpha$, then

$$\mathbb{P}\big(Y_{n+1} \in \widehat{C}_{f,\rho}(X_{n+1})\big) \geq 1-\alpha.$$

The condition that $g_{f,\rho}(1) \geq 1-\alpha$ holds for the KL divergence and the χ^2 distance for any choice of $\rho, \alpha > 0$; it holds for the TV distance for $\alpha \geq \rho/2$. Two remarks are in order.

Remark 3.2. As shown in Cauchois et al. (2023, Lemma A.1), $g_{f,\rho}(\beta)$ is non-decreasing in β, which allows for efficient computation of $g_{f,\rho}^{-1}(\tau)$. For example, by binary search, we can get an estimate of $g_{f,\rho}^{-1}(\tau)$ with error ε within $O(\log((1-\tau)/\epsilon))$ runs.

Remark 3.3. When there is no distributional shift in $Y \mid X$, i.e., $\rho = 0$, our method recovers split weighted conformal prediction (Tibshirani et al., 2019); when there is no X shift, i.e., $w(x) \equiv 1$, it recovers the method of Cauchois et al. (2023). Our procedure is therefore a generalization of both methods.

We now have a general recipe for handling distributional shifts in $\mathcal{P}(\rho; P)$. So far the recipe requires that the X shift $w(x)$ to be specified a priori — this may be the case where the X shift is induced by a covariate-based selection rule that is known to the experimenter — but more often, we do not know the exact form of $w(x)$. The following section discusses how to estimate $w(x)$ with data and how the coverage depends on the estimation quality.

3.1. Estimating the X Shift

Consider a common scenario in prediction tasks: there are multiple test units denoted by $\mathcal{D}_{\text{test}} = \{(X_{n+j}, Y_{n+j})\}_{j=1}^m$, where $(X_{n+j}, Y_{n+j}) \overset{\text{i.i.d.}}{\sim} Q_{X,Y}$. For each $j \in [m]$, we aim to construct a prediction interval $\widehat{C}_{f,\rho,n+j}(X_{n+j})$ satisfying (1). The multiple test units allow us to estimate $w(x)$. In particular, we adopt the estimation approach introduced in Tibshirani et al. (2019), where we first randomly split $\mathcal{D}_{\text{test}}$ into two folds: $\mathcal{D}_{\text{test}}^{(0)}$ and $\mathcal{D}_{\text{test}}^{(1)}$, indexed by $\mathcal{I}_{\text{test}}^{(0)}$ and $\mathcal{I}_{\text{test}}^{(1)}$, respectively. Without loss of generality, assume that $n + j \in \mathcal{I}_{\text{test}}^{(1)}$. Recall that the training set $\mathcal{D}_{\text{tr}}$ is also divided into $\mathcal{D}_{\text{tr}}^{(0)}$ and $\mathcal{D}_{\text{tr}}^{(1)}$. We set aside $\mathcal{D}_{\text{tr}}^{(0)} \cup \mathcal{D}_{\text{test}}^{(0)}$ for estimating $w(\cdot)$. Let A be a binary variable indicating whether the sample is from the training set or the test set, i.e., $A_i = 0$ for $i \in \mathcal{I}_{\text{tr}}^{(0)}$ and $A_i = 1$ for $i \in \mathcal{I}_{\text{test}}^{(0)}$. For $i \in \mathcal{I}_{\text{tr}}^{(0)} \cup \mathcal{I}_{\text{test}}^{(0)}$, by Bayes' rule,

$$\frac{\mathbb{P}(A_i = 1 \,|\, X_i = x)}{\mathbb{P}(A_i = 0 \,|\, X_i = x)} = \frac{dQ_X}{dP_X}(x) \cdot \frac{\mathbb{P}(A_i = 1)}{\mathbb{P}(A_i = 0)} \propto w(x).$$

The above tells us that the likelihood ratio $w(x)$ can be estimated by training a classifier on $\mathcal{D}_{\text{tr}}^{(0)} \cup \mathcal{D}_{\text{test}}^{(0)}$: once we obtain $\widehat{\mathbb{P}}(A = 1 \,|\, X = x)$, we can let $\widehat{w}(x) = \frac{\widehat{\mathbb{P}}(A=1 \,|\, X=x)}{1-\widehat{\mathbb{P}}(A=1 \,|\, X=x)}$ — this is an estimator for $w(x)$ (up to constants). We then construct the prediction interval by replacing $w(x)$ with $\widehat{w}(x)$ in (3). The complete procedure is summarized in Algorithm 1 of Appendix **??**, and the following theorem provides the coverage guarantee when the estimated $w(x)$ is used.

Theorem 3.4. *Assume the same conditions of Theorem 3.1. Let $\widehat{w}^{(k)}$ be the estimated weight function in Algorithm 1, and suppose that $\mathbb{E}_{X \sim P_X}\big[\widehat{w}^{(k)}(X)\big] < \infty$, for $k \in \{0, 1\}$. Then for any $k \in \{0,1\}$ and any $n+j \in \mathcal{I}_{\text{test}}^{(k)}$, the prediction set of Algorithm 1 satisfies*

$$\begin{aligned}&\mathbb{P}\big(Y_{n+j} \in \widehat{C}_{f,\rho,n+j}(X_{n+j})\big) \geq g_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big) \\ &- \frac{1}{2} g'_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big) \cdot \mathbb{E}_{X \sim P_X}\left[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\right],\end{aligned}$$

where $g'_{f,\rho}$ is the left derivative of $g_{f,\rho}$. Furthermore, if $g_{f,\rho}(1) \geq 1 - \alpha$, then

$$\begin{aligned}&\mathbb{P}\big(Y_{n+j} \in \widehat{C}_{f,\rho,n+j}(X_{n+j})\big) \geq 1 - \alpha \\ &- \frac{1}{2} g'_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big) \cdot \mathbb{E}_{X \sim P_X}\left[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\right].\end{aligned}$$

The proof of Theorem 3.4 is based on the coupling technique used in Lei and Candès (2021), and is in Appendix B.2.

Remark 3.5. If the number of test units m is small, one can replace $\mathcal{D}_{\text{test}}^{(1-k)}$ with $\mathcal{D}_{\text{test}} \backslash \{X_{n+j}\}$ when estimating $w(x)$, i.e., train the classifier on $\mathcal{D}_{\text{tr}}^{(0)} \cup \mathcal{D}_{\text{test}} \backslash \{X_{n+j}\}$. This approach can improve the accuracy of the classifier but may be computationally intensive when m is large, so we present the sample-splitting version for simplicity.

Remark 3.6. Theorem 3.4 only assumes the estimated weight to be bounded, but it implicitly requires the overlap condition between P and Q, since otherwise there is no hope that the estimated weight is close to w.

With estimated $w(x)$, Theorem 3.4 suggests that the miscoverage rate inflation depends on the estimation error of $\widehat{w}(x)$. In general, when x is low-dimensional, we can obtain a relatively accurate estimator of $w(x)$, and the resulting prediction interval is approximately valid. In other situations where high-dimensional covariates are present, estimating $w(x)$ can be challenging. To handle this issue, we propose an alternative method that leverages the debiasing technique to construct efficient prediction intervals. We present it in detail in the following section.

3.2. Doubly Robust Prediction Sets

Continue focusing on the test unit $n + j \in \mathcal{I}_{\text{test}}^{(1)}$. Recall that we fit $\widehat{w}^{(1)}(x)$ on $\mathcal{D}_{\text{tr}}^{(0)} \cup \mathcal{D}_{\text{test}}^{(0)}$; we now reuse $\mathcal{D}_{\text{tr}}^{(0)}$ to fit the function $x \mapsto \mathbb{E}\big[\mathbb{1}\{S(X,Y) \leq t\} \,\big|\, X = x\big]$, denoting the estimator by $\widehat{m}^{(1)}(x;t)$. Since our estimand is the conditional cumulative distribution function (CDF), we assume the estimator $\widehat{m}(x;t)$ to be bounded in $[0,1]$, non-decreasing in t, and right-continuous without loss of generality.

To motivate the doubly robust prediction set, let us take another look at the coverage probability under a pure covariate shift at a fixed threshold t, which can be written as

$$\begin{aligned}&\mathbb{P}_{(X,Y) \sim Q_X \times P_{Y \,|\, X}}\big(S(X,Y) \leq t\big) \\ &= \frac{\mathbb{E}_{(X,Y) \sim P_{X,Y}}\big[w(X) \cdot \big(\mathbb{1}\{S(X,Y) \leq t\} - \widehat{m}^{(1)}(X;t)\big)\big]}{\mathbb{E}_{X \sim P_X}[w(X)]} \\ &\quad + \mathbb{E}_{X \sim Q_X}\big[\widehat{m}^{(1)}(X;t)\big].\end{aligned}$$

In the above decomposition, the first term can be estimated with the training data, and the second term with the test data. We therefore modify the coverage probability estimator at threshold t to be

$$\begin{aligned}\widehat{p}^{(1)}(t) = &\frac{\sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \widehat{w}^{(1)}(X_i) \cdot \big(\mathbb{1}\{S_i \leq t\} - \widehat{m}^{(1)}(X_i;t)\big)}{\sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \widehat{w}^{(1)}(X_i)} \\ &+ \frac{1}{|\mathcal{I}_{\text{test},j}^{(1)}|} \sum_{i \in \mathcal{I}_{\text{test},j}^{(1)}} \widehat{m}^{(1)}(X_j;t),\end{aligned}$$

where $\mathcal{I}_{\text{test},j}^{(1)} = \mathcal{I}_{\text{test}}^{(1)} \backslash \{j\}$. Note that $\widehat{p}^{(1)}(t)$ is no longer monotone in t; to obtain the quantile, we consider a "monotonized" version of $\widehat{p}^{(1)}(t)$. The specific prediction interval is then constructed as

$$\begin{aligned}&\widehat{C}_{f,\rho,n+j}^{\text{DR}}(X_{n+j}) = \{y : s(X_{n+j}, y) \leq \widehat{q}\}, \text{where} \\ &\widehat{q} = \inf\Big\{t \in \mathbb{R} : \inf_{t' \geq t} \widehat{p}^{(1)}(t') \geq g_{f,\rho}^{-1}(1-\alpha)\Big\}. \qquad (4)\end{aligned}$$

The complete procedure for constructing the doubly robust prediction sets is described in Algorithm 2. Intuitively, when $\widehat{p}^{(1)}(t)$ is sufficiently close to $\mathbb{P}_{(X_{n+j},Y_{n+j}) \sim Q_X \times P_{Y \,|\, X}}(s(X_{n+j}, Y_{n+j}) \leq t)$, $\widehat{q}$ is close to the $g_{f,\rho}^{-1}(1-\alpha)$-th quantile under $Q_X \times P_{Y \,|\, X}$, thereby upper bounding the $(1-\alpha)$-th quantile of S_{n+j} under $Q_{X,Y}$.

In the following, we let $q^*(\xi)$ be the $(g_{f,\rho}^{-1}(1-\alpha)-\xi)$-th quantile of $s(X,Y)$ under $Q_X \times P_{Y\,|\,X}$. The validity of $\widehat{C}^{\mathrm{DR}}_{f,\rho,n+j}(X_{n+j})$ is established in the following theorem.

Theorem 3.7. *For any $k \in \{0,1\}$, assume that*

(1) $\widehat{w}^{(k)}(x) \le w_{\max} \cdot \mathbb{E}_{P_X}[\widehat{w}^{(k)}(X)]$;

(2) $\widehat{m}^{(k)}(x;t) \in [0,1]$ is non-decreasing and right-continuous in t.

Denote the product estimation error by

$$\begin{aligned}\mathrm{EstErr}^{(k)}(t) =& \big\|\mathbb{1}\{s(X,Y)\le t\} - \widehat{m}^{(1-k)}(X;t)\big\|_{L_2(P)}\\ &\times \left\|\frac{\widehat{w}^{(1-k)}(X)}{\mathbb{E}[\widehat{w}^{(1-k)}(X)]} - w(X)\right\|_{L_2(P)},\end{aligned}$$

where $\|\cdot\|_{L_2(P)}$ denotes the L_2-norm under P, and the expectation is taken conditional on $\mathcal{D}_{\mathrm{tr}}^{(1-k)}$ and $\mathcal{D}_{\mathrm{test}}^{(1-k)}$. For a unit $n+j \in \mathcal{I}_{\mathrm{test}}^{(k)}$, there is

$$\begin{aligned}&\mathbb{P}_{(X_{n+j},Y_{n+j})\sim Q}\big(Y_{n+j} \in \widehat{C}^{\mathrm{DR}}_{f,\rho,n+j}(X_{n+j}) \,\big|\, \mathcal{D}_{\mathrm{tr}}^{(1-k)}, \mathcal{D}_{\mathrm{test}}^{(1-k)}\big)\\ &\ge g_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big) - g'_{f,\rho}\big(g^{-1}(1-\alpha)\big)\\ &\quad\times \left\{\sup_{t\in\mathcal{T}(\alpha)} 2\cdot \mathrm{EstErr}^{(k)}(t) + \sqrt{\frac{16w_{\max}^2}{|\mathcal{I}_{\mathrm{tr}}^{(k)}|} + \frac{2}{|\mathcal{I}_{\mathrm{test},j}^{(k)}|}}\right\},\end{aligned}$$

where $g'_{f,\rho}$ is the left derivative of $g_{f,\rho}$, and $\mathcal{T}(\alpha) = [\underline{q}, \bar{q}]$ is a neighborhood around the $g_{f,\rho}^{-1}(1-\alpha)$-th quantile under $Q_X \times P_{Y\,|\,X}$ with

$$\underline{q} = \sup_{0\le t\le q^*(0)} \mathrm{EstErr}(t) + \sqrt{\frac{9w_{\max}^2}{|\mathcal{I}_{\mathrm{tr}}^{(k)}|} + \frac{1}{|\mathcal{I}_{\mathrm{test},j}^{(k)}|}},\ \bar{q} = q^*(0).$$

When $g_{f,\rho}(1) \ge 1-\alpha$, we further have

$$\begin{aligned}&\mathbb{P}_{(X_{n+j},Y_{n+j})\sim Q}\big(Y_{n+j} \in \widehat{C}^{\mathrm{DR}}_{f,\rho,n+j}(X_{n+j}) \,\big|\, \mathcal{D}_{\mathrm{tr}}^{(1-k)}, \mathcal{D}_{\mathrm{test}}^{(1-k)}\big)\\ &\ge 1-\alpha - g'_{f,\rho}\big(g^{-1}(1-\alpha)\big)\\ &\quad\times \left\{\sup_{t\in\mathcal{T}(\alpha)} 2\cdot \mathrm{EstErr}^{(k)}(t) + \sqrt{\frac{16w_{\max}^2}{|\mathcal{I}_{\mathrm{tr}}^{(k)}|} + \frac{2}{|\mathcal{I}_{\mathrm{test},j}^{(k)}|}}\right\}.\end{aligned}$$

The proof of Theorem 3.7 is deferred to Appendix B.3, where we prove a more general result that the prediction set is valid with high probability conditional on the training data; we then show how the general result implies Theorem 3.7.

Theorem 3.7 implies that the miscoverage rate of $\widehat{C}^{\mathrm{DR}}(X_{n+j})$ is the product of the local estimation error plus an $O(n^{-1/2})$ term, where the product term is small if either $\widehat{w}$ is approximately *proportional* to w, or if $\widehat{m}(x;t)$ is close to $\mathbb{P}_{P_{Y\,|\,X}}(s(X,Y)\le t \,|\, X=x)$ in the neighborhood of the $g_{f,\rho}^{-1}(1-\alpha)$-th quantile under $Q_X \times P_{Y\,|\,X}$. Compared with the double robustness result of Yang et al. (2022), our dependence on the estimation error of $\widehat{m}(x;t)$ is local (around the $g^{-1}(1-\alpha)$-th quantile) while that of Yang et al. (2022) is global (for all t). This is achieved through the "monotonization" step, an idea that also appears in Gui et al. (2023).

3.3. Choice of the Robust Parameter ρ

Another important piece of our procedure is the robust parameter ρ. Choosing ρ is a common challenge in the distributionally robust learning literature (see e.g., Rahimian and Mehrotra (2019); Cauchois et al. (2023); Si et al. (2023); Mu et al. (2022) and the references therein). In certain applications, users can specify an appropriate ρ with context-dependent knowledge. For example, in the sensitivity analysis example to be detailed in Section 4, ρ characterizes the magnitude of the odds ratio of treatment probability whether or not conditioning on the confounders. With this interpretation, practitioners may use background knowledge to determine the plausible degree of confounding, or to determine how they would like to trade efficiency for robustness.

When the choice of ρ is not clear a priori, we provide two solutions based on the proposal of Si et al. (2023):

(1) If there is (a small amount of) supervised data from target distribution, i.e., $Q_{X,Y}$, one can estimate an upper bound of ρ, and use the estimator in place of ρ.

(2) When no supervised data in the target distribution is available, we can apply the procedure with a sequence of ρ, obtaining a sequence of prediction sets. Each value of ρ corresponds to a certain level of robustness, and the user can trade off between the level of robustness and efficiency (e.g., the length of the prediction interval).

In the case where ρ is estimated, Theorem 3.8 characterizes the coverage guarantee of WRCP. We only present the result for WRCP here for simplicity; the result extends also to D-WRCP. When none of these are available, it is in principle impossible to identify the $Y\,|\,X$ shift. We nevertheless can heuristically borrow ideas from the omitted variable bias framework in the causal inference literature, where we randomly choose one predictor, and delete it from the dataset, estimating the resulting ρ (between $P_{Y\,|\,X}$ and $P_{Y\,|\,X_{-j}}$). Repeating the above multiple times, we can take the maximum of the ρ's as our robust parameter.

Theorem 3.8. *Under the same assumptions of Theorem 3.4, suppose that $\widehat{\rho}$ is independent of $(\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{test}})$. Denote $\rho^* := \operatorname{ess\,sup}_x\ D_f(Q_{Y\,|\,X=x} \,\|\, P_{Y\,|\,X=x})$. Then for $k \in \{0,1\}$ and any $n+j \in \mathcal{I}_{\mathrm{test}}^{(k)}$, the prediction interval produced by*

Algorithm 1 with the robust parameter taken to be $\widehat{\rho}$ satisfies

$$\mathbb{P}\big(Y_{n+j} \in \widehat{C}_{f,\widehat{\rho},n+j}(X_{n+j})\big) \geq g_{f,\rho^*}\big(g^{-1}_{f,\widehat{\rho}}(1-\alpha)\big) - \frac{1}{2} g'_{f,\rho^*}\big(g^{-1}_{f,\widehat{\rho}}(1-\alpha)\big) \cdot \mathbb{E}_{X\sim P_X}\Big[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\Big],$$

where $g'_{f,\rho}$ is the left derivative of $g_{f,\rho}$. Furthermore, if $g_{f,\widehat{\rho}}(1) \geq 1-\alpha$ and $\widehat{\rho} \geq \rho^*$, then

$$\mathbb{P}\big(Y_{n+j} \in \widehat{C}_{f,\widehat{\rho},n+j}(X_{n+j})\big) \geq 1-\alpha - \frac{1}{2} g'_{f,\rho^*}\big(g^{-1}_{f,\widehat{\rho}}(1-\alpha)\big) \cdot \mathbb{E}_{X\sim P_X}\Big[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\Big],$$

where $g'_{f,\rho}$ is the left derivative of $g_{f,\rho}$.

4. Application: Sensitivity Analysis of Individual Treatment Effects

Our framework can be applied to the sensitivity analysis of individual treatment effects in the presence of confounding factors. To set the stage, we follow the potential outcome framework (Neyman, 1923; Imbens and Rubin, 2015) and suppose that each sample is associated with a set of random variables $(X, U, T, Y(0), Y(1))$, where $X \in \mathcal{X}$ denotes the observed covariates, $U \in \mathcal{U}$ the unobserved confounders, $T \in \{0,1\}$ the binary treatment, and $Y(1), Y(0) \in \mathbb{R}$ the potential outcomes with and without being treated. Here, not all the quantities are observed — the observable variables are (X, T, Y), where the realized outcome $Y = TY(1) + (1-T)Y(0)$ under the *Stable Unit Treatment Value Assumption (SUTVA)*. Assume that the unobserved confounder U satisfies that[2] $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X, U$.

Imagine now there is a cohort of n i.i.d. samples $(X_i, U_i, T_i, Y_i(1), Y_i(0))_{i=1}^n$, where we observe $\mathcal{D} = (X_i, T_i, Y_i)_{i=1}^n$. For a new individual X_{n+1}, we are interested in a prediction interval $\widehat{C}(X_{n+1})$ for individual treatment effect (ITE), $Y(1) - Y(0)$, such that

$$\mathbb{P}\big(Y(1) - Y(0) \in \widehat{C}(X_{n+1})\big) \geq 1-\alpha. \tag{5}$$

Without additional constraints on the unobserved confounders, it is hopeless to obtain an efficient prediction interval achieving (5), since the difference in the treated and control group can be entirely driven by the confounding factor. Previously, Lei and Candès (2021) studies this problem assuming that there are no observed confounders, i.e., $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$; Jin et al. (2023b) adopts the marginal Γ-selection model (Tan, 2006), which allows for unobserved confounders but the influence of U — roughly speaking — is *uniformly* bounded by a constant Γ. The marginal Γ-selection model can be unsatisfactory in some cases, where the influence of U is limited only *on average* but is unbounded with small probability (the corresponding constant Γ is therefore $+\infty$). Such a situation can however be well characterized by the f-sensitivity model (Jin et al., 2022):

Definition 4.1 (The (f,ρ)-selection condition)**.** Suppose $f : \mathbb{R}_+ \mapsto \mathbb{R}$ is a convex function such that $f(1) = 0$, and P is a distribution over $(X, U, T, Y(1), Y(0))$. P satisfies the (f,ρ)-selection condition if for P-almost all x,

$$\int f\Big(\frac{e(X)}{1-e(X)} \frac{1-\bar{e}(X,U)}{\bar{e}(X,U)}\Big)\, \mathrm{d}P_{U \mid X=x, T=1} \leq \rho,$$
$$\int f\Big(\frac{1-e(X)}{e(X)} \frac{\bar{e}(X,U)}{1-\bar{e}(X,U)}\Big)\, \mathrm{d}P_{U \mid X=x, T=0} \leq \rho,$$

where $\bar{e}(x,u) = P(T = 1 \mid X = x, U = u)$ and $e(x) = P(T = 1 \mid X = x)$.

Can we construct a prediction interval achieving (5) under the f-sensitivity model? It turns out that this task is a special case of our proposed framework. To see this, we first reduce the problem to that of inference on the counterfactuals: if we can construct valid prediction intervals for $Y(1)$ and $Y(0)$, respectively, then combining these two intervals and taking a union bound yields a valid interval for the ITE.

Without loss of generality, we focus on $Y(1)$, aiming to construct an interval $\widehat{C}_{f,\rho}(X_{n+1})$ such that $\mathbb{P}(Y(1) \in \widehat{C}_{f,\rho}(X_{n+1})) \geq 1-\alpha$. Since $Y(1)$ can only be observed for the treated units, the training data follows the distribution $P_{Y(1),X \mid T=1}$ while our target distribution is $P_{Y(1),X}$ — there exists a distributional shift. The X shift can be computed as follows

$$w(x) = \frac{\mathrm{d}P_X}{\mathrm{d}P_{X \mid T=1}}(x) = \frac{\mathbb{P}(T=1)}{e(x)} \propto \frac{1}{e(x)},$$

which depends only on the observable propensity score and can be estimated with the data. Next, we consider the distributional shift in $Y(1) \mid X$. By Jin et al. (2022, Lemma 1), under the f-sensitivity model, $D_f(P_{Y(1) \mid X, T=0} \,\|\, P_{Y(1) \mid X, T=1}) \leq \rho$. Then,

$$D_f\big(P_{Y(1) \mid X} \,\|\, P_{Y(1) \mid X, T=1}\big) \leq (1-e(X)) \cdot D_f(P_{Y(1) \mid X, T=0} \,\|\, P_{Y(1) \mid X, T=1}) \leq \rho,$$

where the inequality follows from the convexity of the f-divergence. By now, it should be clear that the distributional shift in our task consists of an estimable X shift and a shift in $Y \mid X$ bounded in f-divergence, and therefore fits into the framework of this paper. For completeness, we present the adaptation of our main proposal to this specific task of sensitivity analysis, as long as results for other types of estimands in Appendix D.

[2] Such an assumption can always be achieved by taking U to be $(Y(1), Y(0))$.

5. Numerical Results

5.1. Simulation Setup and Evaluation Metrics

We empirically evaluate our proposed methods `WRCP` and `D-WRCP` under a variety of simulation settings. In this section, we present several representative settings and leave the other results to Appendix E.

Three benchmarks are considered for comparison: `CP`: standard conformal prediction (Vovk et al., 2005), `WCP`: weighted conformal prediction (Tibshirani et al., 2019), and `RCP`: robust conformal prediction (Cauchois et al., 2023). We consider $X \in \mathbb{R}^{50}$ and $Y \in \mathbb{R}$, and a ground-truth robust parameter $\rho^* = D_{\text{KL}}(Q_{Y\,|\,X} \,\|\, P_{Y\,|\,X}) = 0.01$. The data generating process is detailed in Appendix E.1,

We consider low, medium, and high levels of X shift respectively. For each run of under a simulation setting, a training set $\mathcal{D}_{\text{tr}}$ and a test set $\mathcal{D}_{\text{test}}$ are generated, with $|\mathcal{D}_{\text{tr}}| = |\mathcal{D}_{\text{test}}| = 2000$. We consider $\rho \in \{0.005, 0.01, \ldots, 0.025\}$ and the target coverage 90%. For each ρ, the above experiment is repeated for $N = 100$ runs, and for each method, we compute the averaged coverage rate and prediction interval length averaged over the 100 runs and 50% of the test samples (1000 samples): $\widehat{\text{Coverage}} = \frac{1}{100\times1000}\sum_{i=1}^{100}\sum_{j=1}^{1000}\mathbb{1}\{Y_{ij} \text{ covered}\}$, and $\widehat{\text{Length}} = \frac{1}{100\times1000}\sum_{i=1}^{100}\sum_{j=1}^{1000}\text{Length}_{ij}$. Ideally, a method should have $\widehat{\text{Coverage}} \geq 90\%$ and as small $\widehat{\text{Length}}$ as possible. The implementation details of all the methods are provided in Appendix E.1.

5.2. Simulation Results

Figure 1 presents the simulation results of all methods. As expected, `CP` and `WCP` fail to achieve the desired coverage level 0.9. `RCP` is overly conservative since it also considers the worst-case X shift. Our proposed method `WRCP` and `D-WRCP` achieve approximate validity for a wide range of ρ — in particular, `WRCP` and `D-WRCP` achieve almost exact coverage when $\rho = \rho^*$; as ρ increases, the coverage remains reasonably close to the target level.

The prediction interval length tells a similar story: `CP` and `WCP` have short prediction intervals due to undercoverage; `RCP` often outputs prediction intervals of infinite length (for the purpose of illustration, we replace $+\infty$ with 17 — an upper bound of all the realized lengths — when plotting the results); our methods provide valid and informative prediction intervals.

6. Real Data Applications

We evaluate the performance of all methods on four real datasets: the national study of learning mindsets dataset (Carvalho et al., 2019), the ACS income dataset (Ding et al., 2021), the covid information study datasets (Pennycook et al., 2020; Roozenbeek et al., 2021), and the poverty mapping dataset (Yeh et al., 2020; Koh et al., 2021). The results corresponding to the first two datasets are presented in the main text, while the results for the latter two datasets are deferred to Appendix F. For each task, we implement `WRCP` and `D-WRCP`, as well as the benchmarks `CP`, `WCP`, and `RCP` with the target coverage level 80%. For `WRCP` and `D-WRCP`, we choose $f(t) = t\log t$ (the KL-divergence), and consider a sequence of robust parameter ρ's, reporting the averaged coverage rate and prediction set length/cardinality as a function of ρ. For `RCP`, the robust parameter is chosen as $\rho_{\text{RCP}} = \rho + D_{\text{KL}}(Q_X \,\|\, P_X)$, where $D_{\text{KL}}(Q_X \,\|\, P_X) = \mathbb{E}_{Q_X}[\log(\mathrm{d}Q_X/\mathrm{d}P_X)]$ is estimated with Monte Carlo with the set-aside training data.

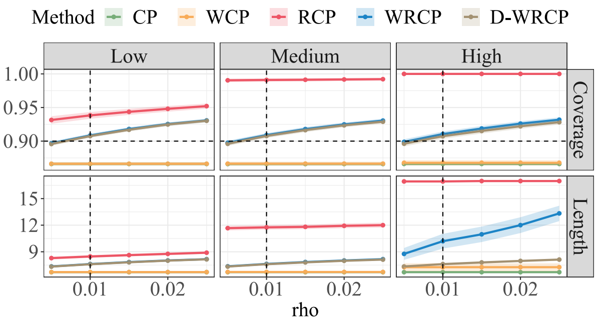

Figure 1. Averaged coverage (top) and prediction interval length (bottom) over $N = 100$ independent runs as a function of the robust parameter ρ, when the amount of X shift is low (left), medium (middle), and high (right). The shaded bars correspond to the 95% confidence intervals. The horizontal dashed line corresponds to the target coverage rate 0.9, and the vertical dashed line is the true robust parameter $\rho^* = 0.01$.

6.1. National Study of Learning Mindsets

The National Study of Learning Mindsets (NSLM) (Yeager et al., 2019; Yeager, 2019) is a randomized study investigating the effect of instilling students with a growth mindset. Based on the results from NSLM, Carvalho et al. (2019) creates an observational study dataset with similar characteristics to the original study. The observational study dataset contains $10{,}391$ students, where $3{,}384$ received the intervention and $7{,}007$ did not. For each student, the dataset records the treatment status T, the outcome Y, and ten covariates: S3, C1, C2, C3, XC, X1, X2, X3, X4, X5.[3] We consider the task of predicting $Y(1)$ in the control group, where we wish

[3] The detailed description of the covariates can be found in Carvalho et al. (2019, Table 1). The original dataset also contains the school id, but we do not include it in our analysis.

to construct a prediction interval $\widehat{C}_{f,\rho}(X)$ such that

$$\mathbb{P}\big(Y(1) \in \widehat{C}_{f,\rho}(X) \,\big|\, T = 0\big) \geq 80\%.$$

Without observing the counterfactuals, the validity of a procedure cannot be evaluated. As an alternative, we create a semi-synthetic dataset based on the NSLM dataset. The data-generating and implementation details are in Appendix F.1.

We consider a sequence of ρ ranging in $\{0.001, 0.005, 0.01, 0.015, \ldots, 0.04\}$, and Figure 2 plots the resulting coverage rate and prediction interval length as a function of ρ. `CP` and `WCP` fail to achieve the desired coverage level, while `RCP` is overly conservative, achieving a much higher coverage rate than the target level. Our methods `WRCP` and `D-WRCP` achieve approximate coverage for a wide range of ρ, and are much more efficient than RCP.

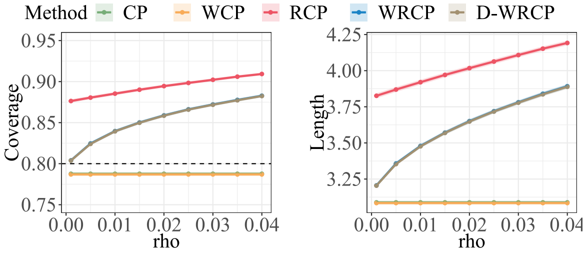

Figure 2. Averaged coverage (left) and prediction interval length (right) over 100 runs as a function of the robust parameter ρ from experiments on the NSLM dataset. The dashed bar corresponds to the 95% confidence interval and the horizontal dashed line corresponds to the target coverage rate 80%.

6.2. ACS Income Dataset

We further evaluate our procedure for a classification task with the ACS income dataset constructed from the US census data (Ding et al., 2021). The target is to predict whether an individual's annual income is above 50,000 dollars. The specific dataset we use is the version preprocessed by Liu et al. (2023), where we choose the data from New York (NY) as the training set, and that from South Dakota (SD) as the target — as discussed in Liu et al. (2023), both X-shift and $Y \mid X$-shift exist between the training and target population. The training and test set contain 103,021 and 4,899 samples, respectively. For each sample, 9 features are recorded, where 3 of them are continuous and 6 categorical. After introducing the dummy variables, the dimension of X comes to 76. The response variable $Y = \mathbb{1}\{\text{income} \geq 50{,}000 \text{ dollars}\}$. The other details are in Appendix F.2.

The robust parameter $\rho \in \{0.005, 0.01, \ldots, 0.04\}$. In each run, we randomly select 2000 samples from the source population, and the same number of samples from the target population. For each ρ, we repeat the above process over 100 random splits and report the averaged coverage and prediction set cardinality.

Figure 3 presents the simulation results of all methods. Ae before, `CP` and `WCP` fail to achieve the desired coverage level 80%; `WCP` improves upon `CP` because it adjusts for the X shift. `RCP` is overly conservative, while our methods again deliver valid and efficient prediction intervals for a wide range of ρ's.

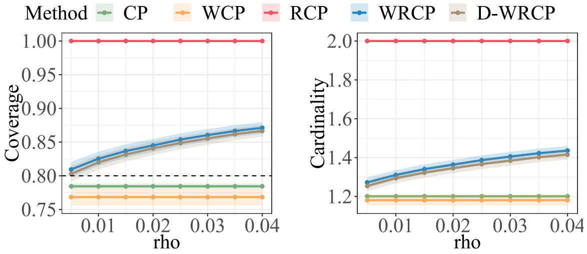

Figure 3. Averaged coverage (left) and prediction set cardinality (right) over 100 runs as a function of the robust parameter ρ from the experiment on ACS income dataset. The other details are the same as in Figure 2.

7. Discussion

In this paper, we provide a fine-grained approach to quantifying the uncertainty of predictive models by distinguishing sources of distributional shift and providing different treatments. We propose two new methods, WRCP and its debiased version D-WRCP, that achieve validity and efficiency under a wide range of distributional shifts, as demonstrated in the simulation and real data experiments. This paper opens up several interesting directions for future work. First, it would be interesting to investigate methods for identifying (an upper bound) of the robust parameter ρ when we have a small amount of supervised data from the target population. Second, can we extend this fine-grained approach to other distributional shift models (e.g., the multi-group model) and improves the efficiency of the corresponding methodologies? Last but not least, there can be other ways to decompose the distributional shifts — it remains to be understood the optimal decomposition in different settings and the corresponding treatments.

Reproducibility

All the numerical results in this paper can be reproduced with the code available at https://github.com/zhimeir/finegrained-conformal-paper.

Acknowledgements

The authors would like to thank Wharton High Performance Computing for the computational resources and the support from the staff members. The authors would also like to thank Ying Jin for the feedback on the manuscript.

Impact Statement

This work is devoted to the uncertainty quantification of predictive models, which ultimately promotes trustworthy machine learning. This applies generally to different aspects of society as such predictive models are increasingly deployed in our society.

References

Angelopoulos, A. N., Bates, S., et al. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845.

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

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Carvalho, C., Feller, A., Murray, J., Woody, S., and Yeager, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2023). Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, (just-accepted):1–22.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490.

Duchi, J., Hashimoto, T., and Namkoong, H. (2023). Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2):649–664.

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406.

Elie-Dit-Cosaque, K. (2020). qosa-indices.

Gendler, A., Weng, T.-W., Daniel, L., and Romano, Y. (2021). Adversarially robust conformal prediction. In *International Conference on Learning Representations*.

Ghosh, S., Shi, Y., Belkhouja, T., Yan, Y., Doppa, J., and Jones, B. (2023). Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, pages 681–690. PMLR.

Guan, L. (2023). Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50.

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

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Hu, X. and Lei, J. (2023). A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association*, pages 1–19.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jin, Y., Guo, K., and Rothenhäusler, D. (2023a). Diagnosing the role of observable distribution shift in scientific replications. *arXiv preprint arXiv:2309.01056*.

Jin, Y., Ren, Z., and Candès, E. J. (2023b). 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*.

Kallus, N. and Zhou, A. (2020). Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in neural information processing systems*, 33:22293–22304.

Kallus, N. and Zhou, A. (2021). Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938.

Lei, L., Sahoo, R., and Wager, S. (2023). Policy learning under biased sample selection. *arXiv preprint arXiv:2304.11735*.

Liu, J., Wang, T., Cui, P., and Namkoong, H. (2023). On the need for a language describing distribution shifts: Illustrations on tabular datasets. *arXiv preprint arXiv:2307.05284*.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. (2020). The effect of natural distribution shift on question answering models. In *International conference on machine learning*, pages 6905–6916. PMLR.

Mu, T., Chandak, Y., Hashimoto, T. B., and Brunskill, E. (2022). Factored DRO: Factored distributionally robust policies for contextual bandits. *Advances in Neural Information Processing Systems*, 35:8318–8331.

Namkoong, H., Yadlowsky, S., et al. (2023). Diagnosing model performance under distribution shift. *arXiv preprint arXiv:2303.02011*.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer.

Park, S., Dobriban, E., Lee, I., and Bastani, O. (2022). PAC prediction sets under covariate shift. In *International Conference on Learning Representations*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., and Rand, D. G. (2020). Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780.

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR.

Qiu, H., Dobriban, E., and Tchetgen, E. T. (2022). Distribution-free prediction sets adaptive to unknown covariate shift. *arXiv preprint arXiv:2203.06126*.

Rahimian, H. and Mehrotra, S. (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Romano, Y., Sesia, M., and Candès, E. J. (2020). Classification with valid and adaptive coverage.

Roozenbeek, J., Freeman, A. L., and van der Linden, S. (2021). How accurate are accuracy-nudge interventions? a preregistered direct replication of pennycook et al.(2020). *Psychological science*, 32(7):1169–1178.

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

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

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

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. (2023). Distributionally robust batch contextual bandits. *Management Science*.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

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

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070.

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, Y., Kuchibhotla, A. K., and Tchetgen, E. T. (2022). Doubly robust calibration of prediction sets under covariate shift. *arXiv preprint arXiv:2203.01761*.

Yeager, D. S. (2019). *The National Study of Learning Mindsets,[United States], 2015-2016*.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., et al. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774):364–369.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*.

Yin, M., Shi, C., Wang, Y., and Blei, D. M. (2022). Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, pages 1–14.

Zhang, Z., Zhan, W., Chen, Y., Du, S. S., and Lee, J. D. (2023). Optimal multi-distribution learning. *arXiv preprint arXiv:2312.05134*.

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.

A. Auxiliary Lemmas

Lemma A.1 (Data processing inequality (Cover, 1999))**.** *Let X, Y, Z denote random variables drawn from a Markov chain in the order (denoted by $X \to Y \to Z$) that the conditional distribution of Z depends only on Y and is conditionally independent of X. Then if $X \to Y \to Z$, we have $I(X;Y) \geq I(X;Z)$, where $I(X;Y)$ is the mutual information between X and Y.*

Lemma A.2 (Adapted from Lemma A.1 of Cauchois et al. (2023))**.** *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a closed convex function such that $f(1) = 0$ and $f(t) < \infty$ for all $t > 0$. The function $g_{f,\rho}(\beta)$ has the following properties.*

(a) *$(\rho, \beta) \to g_{f,\rho}(\beta)$ is a convex function and continuous in $\beta \in [0,1]$ and $\rho \in (0,\infty)$.*

(b) *$g_{f,\rho}(\beta)$ is non-increasing in ρ and non-decreasing in β. Moreover, for all $\rho > 0$, there exists $\beta_0(\rho) := \sup\big\{\beta \in (0,1) \,|\, g_{f,\rho}(\beta) = 0\big\}$, and $g_{f,\rho}(\beta)$ is strictly increasing for $\beta > \beta_0(\rho)$.*

B. Technical Proofs

B.1. Proof of Theorem 3.1

For notional simplicity, let $A := \big\{Y_{n+1} \in \widehat{C}_{f,\rho}(X_{n+1})\big\}$. We aim to show that

$$\mathbb{P}_{(X_{n+1},Y_{n+1})\sim Q_{X,Y}}(A) \geq 1-\alpha.$$

Denote by $\text{Bern}(p)$ the Bernoulli distribution with success probability p. By data processing inequality (Lemma A.1), we have that

$$D_f\big(Q_{Y\,|\,X} \,\|\, P_{Y\,|\,X}\big) \geq D_f\Big(\text{Bern}\big(\mathbb{P}_{Y_{n+1}\sim Q_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})\big) \,\Big\|\, \text{Bern}\big(\mathbb{P}_{Y_{n+1}\sim P_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})\big)\Big).$$

Recall that $Q \in \mathcal{P}(\rho; P)$. We then have

$$\begin{aligned}
\rho &\geq D_f\big(Q_{Y\,|\,X} \,\|\, P_{Y\,|\,X}\big)\\
&\geq D_f\Big(\text{Bern}\big(\mathbb{P}_{Y_{n+1}\sim Q_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})\big) \,\Big\|\, \text{Bern}\big(\mathbb{P}_{Y_{n+1}\sim P_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})\big)\Big)\\
&= \mathbb{P}_{Y_{n+1}\sim P_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D}) \cdot f\left(\frac{\mathbb{P}_{Y_{n+1}\sim Q_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})}{\mathbb{P}_{Y_{n+1}\sim P_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})}\right)\\
&\qquad\qquad + \big(1-\mathbb{P}_{Y\sim P_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})\big) \cdot f\left(\frac{1-\mathbb{P}_{Y\sim Q_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})}{1-\mathbb{P}_{Y\sim P_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})}\right),
\end{aligned}$$

where the last step follows from the definition of the f-divergence. Combining the above and the definition of $g_{f,\rho}$, one can obtain that almost surely

$$g_{f,\rho}\big(\mathbb{P}_{Y\sim P_{Y\,|\,X}}(A)\big) \leq \mathbb{P}_{Y\sim Q_{Y\,|\,X}}(A). \tag{6}$$

Next, we take the expectation over the randomness of the training set $\mathcal{D}$ and X_{n+1}

$$\begin{aligned}
\mathbb{P}_{\mathcal{D},\,(X_{n+1},Y_{n+1})\sim Q_{X,Y}}(A) &= \mathbb{E}_{\mathcal{D},\,X_{n+1}\sim Q_X}\left[\mathbb{P}_{Y_{n+1}\sim Q_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})\right]\\
&\geq \mathbb{E}_{\mathcal{D},X_{n+1}\sim Q_X}\Big[g_{f,\rho}\big(\mathbb{P}_{Y\sim P_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})\big)\Big],
\end{aligned} \tag{7}$$

where the last inequality is because of (6). Since $g_{f,\rho}$ is convex (Lemma A.2), Jensen's inequality implies

$$\mathbb{E}_{\mathcal{D},X_{n+1}\sim Q_X}\Big[g_{f,\rho}\big(\mathbb{P}_{Y\sim P_{Y\,|\,X}}(A\,|\,X_{n+1},\mathcal{D})\big)\Big] \geq g_{f,\rho}\Big(\mathbb{P}_{\mathcal{D},(X_{n+1},Y_{n+1})\sim Q_X\times P_{Y\,|\,X}}(A)\Big). \tag{8}$$

By Tibshirani et al. (2019, Corollary 1) and the construction of $\widehat{C}_{f,\rho}(X_{n+1})$, there is

$$\mathbb{P}_{\mathcal{D},(X_{n+1},Y_{n+1})\sim Q_X\times P_{Y\,|\,X}}(A) \geq g^{-1}_{f,\rho}(1-\alpha).$$

Again leveraging the monotonicity of $g_{f,\rho}(\beta)$ in β, we arrive at

$$\mathbb{P}_{\mathcal{D},(X_{n+1},Y_{n+1})\sim Q_{X,Y}}(A) \geq g_{f,\rho}\big(g^{-1}_{f,\rho}(1-\alpha)\big).$$

When $g_{f,\rho}(1) \geq 1-\alpha$, we prove that $g_{f,\rho}\big(g^{-1}_{f,\rho}(1-\alpha)\big) \geq 1-\alpha$ by contradiction. Suppose otherwise that $g_{f,\rho}\big(g^{-1}_{f,\rho}(1-\alpha)\big) < 1-\alpha$. Then $g^{-1}_{f,\rho}(1-\alpha) < 1$. By the continuity of $g_{f,\rho}(\beta)$ in β, there exists a small $\varepsilon > 0$, such that $g_{f,\rho}(\beta) < 1-\alpha$ for all $\beta \in [g^{-1}_{f,\rho}(1-\alpha), g^{-1}_{f,\rho}(1-\alpha)+\epsilon]$, which contradicts the definition of $g^{-1}_{f,\rho}(1-\alpha)$. The proof is therefore completed.

B.2. Proof of Theorem 3.4

Throughout the proof, we condition on $\mathcal{D}_{\text{tr}}^{(0)} \cup \mathcal{D}_{\text{test}}^{(1-k)}$, and for notational simplicity we do not explicitly write the conditional probability and expectation when the context is clear.

We start by defining a new distribution $\widetilde{Q}_{X,Y} = \widetilde{Q}_X \times P_{Y \mid X}$ such that

$$\frac{d\tilde{Q}_X}{dP_X}(x) = \frac{\widehat{w}^{(k)}(x)}{\mathbb{E}_{X \sim P_X}[\widehat{w}^{(k)}(X)]}.$$

Let $A = \big\{Y_{n+j} \in \widehat{C}_{f,\rho,j}(X_{n+j})\big\}$. If (X_{n+j}, Y_{n+j}) were indeed sampled from $\widetilde{Q}_{X,Y}$, then by Lemma 1 of Tibshirani et al. (2019), we have

$$\mathbb{P}_{\mathcal{D}_{\text{tr}}^{(1)}, (X_{n+j}, Y_{n+j}) \sim \widetilde{Q}_{X,Y}}(A) \geq g_{f,\rho}^{-1}(1-\alpha). \tag{9}$$

Meanwhile, by the definition of the TV distance,

$$\begin{aligned}
&\Big|\mathbb{P}_{(X_{n+j},Y_{n+j}) \sim \widetilde{Q}_X \times P_{Y \mid X}}\big(A \,|\, \mathcal{D}_{\text{tr}}^{(1)}\big) - \mathbb{P}_{(X_{n+j},Y_{n+j}) \sim Q_X \times P_{Y \mid X}}\big(A \,|\, \mathcal{D}_{\text{tr}}^{(1)}\big)\Big| \\
&\leq D_{\text{TV}}\big(\widetilde{Q}_X \times Q_{Y \mid X} \,\|\, Q_X \times Q_{Y \mid X}\big) \\
&= D_{\text{TV}}(\widetilde{Q}_X \,\|\, Q_X) = \frac{1}{2}\mathbb{E}_{X \sim P_X}\Big[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{X \sim P_X}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\Big].
\end{aligned}$$

The above inequality implies that

$$\mathbb{P}_{(X_{n+j},Y_{n+j}) \sim Q_X \times P_{Y \mid X}}\big(A \,\big|\, \mathcal{D}_{\text{tr}}^{(1)}\big) \geq \mathbb{P}_{(X_{n+j},Y_{n+j}) \sim \widetilde{Q}_X \times P_{Y \mid X}}\big(A \,\big|\, \mathcal{D}_{\text{tr}}^{(1)}\big) - \frac{1}{2}\mathbb{E}_{X \sim P_X}\Big[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{X \sim P_X}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\Big].$$

Taking expectation over $\mathcal{D}_{\text{tr}}^{(1)}$, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_{\text{tr}}^{(1)}, (X_{n+j},Y_{n+j}) \sim Q_X \times P_{Y \mid X}}(A) \\
&\geq \mathbb{P}_{\mathcal{D}_{\text{tr}}^{(1)}, (X_{n+j},Y_{n+j}) \sim \widetilde{Q}_{X,Y}}(A) - \frac{1}{2}\mathbb{E}_{X \sim P_X}\Big[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{X \sim P_X}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\Big] \\
&\geq g_{f,\rho}^{-1}(1-\alpha) - \frac{1}{2}\mathbb{E}_{X \sim P_X}\Big[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{X \sim P_X}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\Big],
\end{aligned}$$

where the last step follows from (9). Following the same argument as in the proof of Theorem 3.1 (Eqn. (7) and (8)), we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_{\text{tr}}^{(1)}, (X_{n+1},Y_{n+1}) \sim Q_{X,Y}}(A) \\
&\geq g_{f,\rho}\Big(\mathbb{P}_{\mathcal{D}_{\text{tr}}^{(1)}, (X_{n+j},Y_{n+j}) \sim Q_X \times P_{Y \mid X}}(A)\Big) \\
&\geq g_{f,\rho}\Bigg(g_{f,\rho}^{-1}(1-\alpha) - \frac{1}{2}\mathbb{E}_{X \sim P_X}\Big[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{X \sim P_X}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\Big]\Bigg),
\end{aligned} \tag{10}$$

where the last inequality is because $g_{f,\rho}(\beta)$ is non-decreasing in β. Since $g_{f,\rho}(\beta)$ is convex in $(0,1)$, the left derivative exists and by the separating hyperplane theorem, we further have

$$(10) \geq g_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big) - \frac{1}{2}g'_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big) \cdot \mathbb{E}_{X \sim P_X}\Big[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{X \sim P_X}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\Big].$$

As shown in the proof of Theorem 3.1, when $g_{f,\rho}(1) \geq 1-\alpha$, $g_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big) \geq 1-\alpha$, and we conclude the proof.

B.3. Proof for the Doubly Robust Prediction Intervals

We start by proving that the $D^{\text{DR}}_{f,\rho,n+j}(X_{n+j})$ is training-conditionally valid in Theorem B.1, and then show that Theorem 3.7 is a direct consequence of Theorem B.1.

Theorem B.1. *For any $k \in \{0,1\}$, assume that*

(1) $\widehat{w}^{(k)}(x) \leq w_{\max} \cdot \mathbb{E}_{P_X}[\widehat{w}^{(k)}(X)]$;

(2) $\widehat{m}^{(k)}(x;t) \in [0,1]$ is non-decreasing and right-continuous in t.

Denote the product estimation error by

$$\text{EstErr}^{(k)}(t) = \big\|\mathbb{1}\{s(X,Y;\widehat{\mu}^{(k)}) \leq t\} - \widehat{m}^{(k)}(X;t)\big\|_{L_2(P)}\left\|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{X\sim P_X}[\widehat{w}^{(k)}(X)]} - w(X)\right\|_{L_2(P)},$$

where $\|\cdot\|_{L_2(P)}$ denotes the L_2-norm under P, and the expectation is taken conditional on $\mathcal{D}_{\text{tr}}^{(1-k)}$ and $\mathcal{D}_{\text{test}}^{(1-k)}$. Then for any $\delta > 0$ and any unit $n+j \in \mathcal{I}_{\text{test}}^{(k)}$, with probability at least $1-\delta$,

$$\begin{aligned}&\mathbb{P}_{(X_{n+j},Y_{n+j})\sim Q_{X,Y}}\big(Y_{n+j} \in \widehat{C}^{\text{DR}}_{f,\rho,n+j}(X_{n+j}) \,\big|\, \mathcal{D}_{\text{tr}}^{(k)}, \mathcal{D}_{\text{test},j}^{(k)}, \mathcal{D}_{\text{tr}}^{(1-k)}, \mathcal{D}_{\text{test}}^{(1-k)}\big)\\ &\geq g_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big) - g'_{f,\rho}\big(g^{-1}(1-\alpha)\big) \times \left\{\sup_{t\in\mathcal{T}(\alpha)} \text{EstErr}^{(k)}(t) + \sqrt{\log\Big(\frac{1}{\delta}\Big)\cdot\Big(\frac{9w_{\max}^2}{|\mathcal{I}_{\text{tr}}^{(k)}|} + \frac{1}{|\mathcal{I}_{\text{test},j}^{(k)}|}\Big)}\right\},\end{aligned}$$

where $g'_{f,\rho}$ is the left derivative of $g_{f,\rho}$. If $g_{f,\rho}(1) \geq 1-\alpha$, then with probability at least $1-\delta$,

$$\begin{aligned}&\mathbb{P}_{(X_{n+j},Y_{n+j})\sim Q_{X,Y}}\big(Y_{n+j} \in \widehat{C}^{\text{DR}}_{f,\rho,n+j}(X_{n+j}) \,\big|\, \mathcal{D}_{\text{tr}}^{(k)}, \mathcal{D}_{\text{test},j}^{(k)}, \mathcal{D}_{\text{tr}}^{(1-k)}, \mathcal{D}_{\text{test}}^{(1-k)}\big)\\ &\geq 1-\alpha - g'_{f,\rho}\big(g^{-1}(1-\alpha)\big) \times \left\{\sup_{t\in\mathcal{T}(\alpha)} \text{EstErr}^{(k)}(t) + \sqrt{\log\Big(\frac{1}{\delta}\Big)\cdot\Big(\frac{9w_{\max}^2}{|\mathcal{I}_{\text{tr}}^{(k)}|} + \frac{1}{|\mathcal{I}_{\text{test},j}^{(k)}|}\Big)}\right\}.\end{aligned}$$

Proof. Without loss of generality, assume $k=1$. Throughout, we condition on $\mathcal{D}_{\text{tr}}^{(0)} \cup \mathcal{D}_{\text{test}}^{(0)}$ without explicitly writing the conditioning event when the context is clear. For notational simplicity, we define the normalized weight as

$$\widetilde{w}^{(k)}(x) = \frac{\widehat{w}^{(k)}(x)}{\mathbb{E}_{X\sim P_X}[\widehat{w}^{(k)}(X)]}.$$

In the proof we leave out the dependence on k, writing $\widehat{m}(X;t)$ and $\widetilde{w}(X)$ in place of $\widehat{m}^{(1)}(X;t)$ and $\widetilde{w}^{(1)}(X)$; additionally, we refer to $\mathbb{E}_{(X,Y)\sim P_{X,Y}}$ as $\mathbb{E}_{P_{X,Y}}$, with the same rule applied to the expectation/probability under other distributions, and let $n_{\text{tr}} = |\mathcal{I}_{\text{tr}}^{(1)}|$ and $n_{\text{test}} = |\mathcal{I}_{\text{test},j}^{(1)}|$.

For any $t \in \mathbb{R}$, we define the oracle CDF $F(t) = \mathbb{P}_{Q_X\times P_{Y\,|\,X}}\big(s(X,Y) \leq t\big)$, and for any $\xi \in [0, g_{f,\rho}^{-1}(1-\alpha)]$, the perturbed oracle quantile $q^*(\xi)$ can be equivalently written as

$$q^*(\xi) = \inf\big\{t \in \mathbb{R} : F(t) \geq g_{f,\rho}^{-1}(1-\alpha) - \xi\big\}.$$

Consider the error of margin

$$\Delta = \sup_{\kappa\cdot q^*(\bar{\Delta})\leq t\leq q^*(0)} \text{EstErr}(t) + \sqrt{\log\Big(\frac{1}{\delta}\Big)\cdot\Big(\frac{9w_{\max}^2}{n_{\text{tr}}} + \frac{1}{n_{\text{test}}}\Big)},$$

where $\kappa \in (0,1)$ is a constant that can be arbitrarily close to 1 and

$$\bar{\Delta} = \left\{\sup_{0\leq t\leq q^*(0)} \text{EstErr}(t) + \sqrt{\log\Big(\frac{1}{\delta}\Big)\cdot\Big(\frac{9w_{\max}^2}{n_{\text{tr}}} + \frac{1}{n_{\text{test}}}\Big)}\right\} \wedge g_{f,\rho}^{-1}(1-\alpha).$$

Here, $a \wedge b = \min(a,b)$ and Δ is fully deterministic conditional on $\mathcal{D}_{\text{tr}}^{(0)} \cup \mathcal{D}_{\text{test}}^{(0)}$.

The proof consists of two steps: (1) we show that, with high probability, $\widehat{q} \geq q^*(\Delta)$. and therefore $\widehat{q}$ is no less than the $(g_{f,\rho}^{-1}(1-\alpha) - \Delta)$-th quantile under $Q_X \times P_{Y\,|\,X}$, and (2) $\widehat{q}$ is approximately an upper bound of the $(1-\alpha)$-th quantile under $Q_{X,Y}$.

Step (1). On the event $\{q^*(\Delta) \le \widehat{q}\}$,

$$\mathbb{P}\big(s(X_{n+j}, Y_{n+j}) \le \widehat{q} \,|\, \mathcal{D}_{\text{tr}}^{(1)}, \mathcal{D}_{\text{test},j}^{(1)}\big) = F(\widehat{q}) \overset{\text{(i)}}{\ge} F(q^*(\Delta)) \overset{\text{(ii)}}{\ge} g_{f,\rho}^{-1}(1-\alpha) - \Delta,$$

where step (i) follows from the monotonicity of $F(t)$, and step (ii) is by the definition of $q^*(\Delta)$ and that $F(t)$ is right-continuous. It then suffices to control the probability of $\{q^*(\Delta) > \widehat{q}\}$. Fixing $\delta \in [0,1]$, we aim at showing that

$$\mathbb{P}\big(\widehat{q} < q^*(\Delta)\big) \le \delta.$$

The above is trivial when $\Delta \ge g_{f,\rho}^{-1}(1-\alpha)$. We proceed assuming that $\Delta < g_{f,\rho}^{-1}(1-\alpha)$.

For any $\varepsilon > 0$, we have that $F(q^*(\Delta) - \varepsilon) < g_{f,\rho}^{-1}(1-\alpha) - \Delta$ by the definition of $q^*(\Delta)$. If $\widehat{q} \le q^*(\Delta) - \varepsilon$, then the choice of $\widehat{q}$ implies that

$$\widehat{p}(q^*(\Delta) - \varepsilon) \ge g_{f,\rho}^{-1}(1-\alpha).$$

In other words,

$$\mathbb{P}\big(\widehat{q} \le q^*(\Delta) - \varepsilon\big) \le \mathbb{P}\Big(\widehat{p}(q^*(\Delta) - \varepsilon) \ge g_{f,\rho}^{-1}(1-\alpha)\Big). \tag{11}$$

For better readability, we use $\bar{t}$ to represent $q^*(\Delta) - \varepsilon$, and let $Z_i = \mathbb{1}\{S_i \le \bar{t}\} - \widehat{m}(X_i; \bar{t})$ in the following. The right-hand side of (11) can be further upper bounded as

$$\begin{aligned} &\mathbb{P}\big(\widehat{p}(\bar{t}) \ge g_{f,\rho}^{-1}(1-\alpha)\big) \\ &= \mathbb{P}\bigg(\sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \widetilde{w}(X_i) Z_i + \Big(\sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \widetilde{w}(X_i)\Big)\Big(\frac{1}{n_{\text{test}}} \sum_{i \in \mathcal{I}_{\text{test},j}^{(1)}} \widehat{m}(X_i; \bar{t}) - g_{f,\rho}^{-1}(1-\alpha)\Big) \ge 0\bigg) \\ &\le \mathbb{E}\bigg[\exp\bigg\{\eta\Big(\sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \widetilde{w}(X_i) Z_i + \Big(\sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \widetilde{w}(X_i)\Big)\Big(\frac{1}{n_{\text{test}}} \sum_{i \in \mathcal{I}_{\text{test},j}^{(1)}} \widehat{m}(X_i; \bar{t}) - g_{f,\rho}^{-1}(1-\alpha)\Big)\Big)\bigg\}\bigg], \end{aligned} \tag{12}$$

where $\eta > 0$ is some constant to be determined and the last step follows from Markov's inequality.

Conditional on $\{X_i\}_{i \in \mathcal{I}_{\text{tr}}^{(1)}}$, $Z_i - \mathbb{E}[Z_i \,|\, X_i]$ are $\frac{1}{4}$-subgaussian random variables. Therefore,

$$\mathbb{E}\bigg[\exp\bigg\{\eta\Big(\sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \widetilde{w}(X_i) \cdot \big(Z_i - \mathbb{E}[Z_i \,|\, X_i]\big)\Big)\bigg\} \,\bigg|\, \{X_i\}_{i \in \mathcal{I}_{\text{tr}}^{(1)}}\bigg] \le \exp\Big(\frac{\eta^2 w_{\max}^2 n_{\text{tr}}}{8}\Big).$$

The above implies that

$$\begin{aligned} (12) \le\ & \exp\Big(\frac{\eta^2 w_{\max}^2 n_{\text{tr}}}{8}\Big) \cdot \mathbb{E}\bigg[\exp\bigg\{\eta\Big(\sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \widetilde{w}(X_i)\mathbb{E}[Z_i \,|\, X_i] \\ &\qquad + \Big(\sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \widetilde{w}(X_i)\Big)\Big(\frac{1}{n_{\text{test}}} \sum_{i \in \mathcal{I}_{\text{test},j}} \widehat{m}(X_i; \bar{t}) - g_{f,\rho}^{-1}(1-\alpha)\Big)\Big)\bigg\}\bigg] \\ \le\ & \exp\Big(\frac{\eta^2 w_{\max}^2 n_{\text{tr}}}{8}\Big) \cdot \mathbb{E}\bigg[\exp\bigg\{2\eta\Big(\sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \Big(\widetilde{w}(X_i)\mathbb{E}[Z_i \,|\, X_i] + \frac{1}{n_{\text{test}}} \sum_{\ell \in \mathcal{I}_{\text{test},j}^{(1)}} \widehat{m}(X_\ell; \bar{t}) - g_{f,\rho}^{-1}(1-\alpha)\Big)\Big)\bigg\}\bigg]^{1/2} \\ & \times \mathbb{E}\bigg[\exp\Big(2\eta \sum_{i \in \mathcal{I}_{\text{tr}}^{(1)}} \big(\widetilde{w}(X_i) - 1\big) \cdot \Big(\frac{1}{n_{\text{test}}} \sum_{\ell \in \mathcal{I}_{\text{test},j}^{(1)}} \widehat{m}(X_\ell; \bar{t}) - g_{f,\rho}^{-1}(1-\alpha)\Big)\Big)\bigg]^{1/2}, \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Since $\mathcal{D}_{\text{test}}^{(1)}$ is independent of $\mathcal{D}_{\text{tr}}^{(1)}$, conditional on $\mathcal{D}_{\text{test}}^{(1)}$, there is

$$\mathbb{E}\bigg[\exp\Big(2\eta\sum_{i\in\mathcal{I}_{\text{tr}}^{(1)}}\big(\widetilde{w}(X_i)-1\big)\Big(\frac{1}{n_{\text{test}}}\sum_{\ell\in\mathcal{I}_{\text{test},j}^{(1)}}\widehat{m}(X_\ell;\bar{t})-g_{f,\rho}^{-1}(1-\alpha)\Big)\Big)\,\Big|\,\mathcal{D}_{\text{test}}^{(1)}\bigg]^{1/2}\leq\exp\big(\eta^2w_{\max}^2n_{\text{tr}}\big),$$

where we use the sub-gaussianity of $\widetilde{w}(X_i)$. Recalling that $F(\bar{t})<g_{f,\rho}^{-1}(1-\alpha)-\Delta$, we have that

$$\begin{aligned}&\mathbb{E}\bigg[\exp\bigg\{2\eta\Big(\sum_{i\in\mathcal{I}_{\text{tr}}^{(1)}}\Big(\widetilde{w}(X_i)\mathbb{E}[Z_i\,|\,X_i]+\frac{1}{n_{\text{test}}}\sum_{\ell\in\mathcal{I}_{\text{test},j}^{(1)}}\widehat{m}(X_\ell;\bar{t})-g_{f,\rho}^{-1}(1-\alpha)\Big)\Big)\bigg\}\bigg]^{1/2}\\ \leq{}&\mathbb{E}\bigg[\exp\bigg\{2\eta\Big(\sum_{i\in\mathcal{I}_{\text{tr}}^{(1)}}\Big(\widetilde{w}(X_i)\mathbb{E}[Z_i\,|\,X_i]+\frac{1}{n_{\text{test}}}\sum_{\ell\in\mathcal{I}_{\text{test},j}^{(1)}}\widehat{m}(X_\ell;\bar{t})-F(\bar{t})-\Delta\Big)\Big)\bigg\}\bigg]^{1/2}.\end{aligned}$$

By definition, $F(\bar{t})=\mathbb{P}_{Q_X\times P_{Y\,|\,X}}(s(X,Y)\leq\bar{t})=\mathbb{E}_{P_{X,Y}}[w(X)Z]+\mathbb{E}_{Q_X}[\widehat{m}(X;\bar{t})]$. Therefore,

$$\begin{aligned}&\big|F(\bar{t})-\mathbb{E}_{P_{X,Y}}[\widetilde{w}(X)Z]-\mathbb{E}_{Q_X}[\widehat{m}(X;\bar{t})]\big|\\ \leq{}&\Big|\mathbb{E}_{P_{X,Y}}\Big[\big(w(X)-\widetilde{w}(X)\big)\cdot\big(\mathbb{1}\{s(X,Y)\leq\bar{t}\}-\widehat{m}(X;\bar{t})\big)\Big]\Big|\\ \leq{}&\big\|w(X)-\widetilde{w}(X)\big\|_{L_2(P)}\cdot\big\|\mathbb{1}\{s(X,Y)\leq\bar{t}\}-\widehat{m}(X;\bar{t})\big\|_{L_2(P)}\\ ={}&\text{EstErr}(\bar{t}).\end{aligned}\tag{13}$$

The last inequality follows from the Cauchy-Schwarz inequality.

Next, we focus on the following quantity:

$$\begin{aligned}&\mathbb{E}\bigg[\exp\bigg\{2\eta\Big(\sum_{i\in\mathcal{I}_{\text{tr}}^{(1)}}\Big(\widetilde{w}(X_i)\mathbb{E}[Z_i\,|\,X_i]-\mathbb{E}_{P_{X,Y}}[\widetilde{w}(X)Z]\Big)+\frac{n_{\text{tr}}}{n_{\text{test}}}\sum_{i\in\mathcal{I}_{\text{test},j}^{(1)}}\Big(\widehat{m}(X_i;\bar{t})-\mathbb{E}_{Q_X}[\widehat{m}(X;\bar{t})]\Big)\Big)\bigg\}\bigg]\\ ={}&\mathbb{E}\bigg[\exp\bigg\{2\eta\sum_{i\in\mathcal{I}_{\text{tr}}^{(1)}}\Big(\widetilde{w}(X_i)\mathbb{E}[Z_i\,|\,X_i]-\mathbb{E}_{P_{X,Y}}[\widetilde{w}(X)Z]\Big)\bigg\}\bigg]\\ &\qquad+\mathbb{E}\bigg[\exp\bigg\{\frac{2\eta\cdot n_{\text{tr}}}{n_{\text{test}}}\sum_{i\in\mathcal{I}_{\text{test},j}^{(1)}}\Big(\widehat{m}(X_i;\bar{t})-\mathbb{E}_{Q_X}[\widehat{m}(X;\bar{t})]\Big)\bigg\}\bigg]\\ \leq{}&\exp\Big(2\eta^2w_{\max}^2n_{\text{tr}}+\frac{\eta^2n_{\text{tr}}^2}{2n_{\text{test}}}\Big),\end{aligned}\tag{14}$$

where the second step uses the independence between $\mathcal{D}_{\text{test}}$ and $\mathcal{D}_{\text{tr}}$. Combining (8), (13) and (14) leads to

$$(8)\leq\exp\Big\{-\eta n_{\text{tr}}\big(\Delta-\text{EstErr}(\bar{t})\big)+\eta^2\Big(w_{\max}^2n_{\text{tr}}+\frac{n_{\text{tr}}^2}{4n_{\text{test}}}\Big)\Big\}.$$

Putting everything together, we conclude that

$$\mathbb{P}\Big(\widehat{p}(q^*(\Delta)-\varepsilon)\geq g_{f,\rho}^{-1}(1-\alpha)\Big)\leq\exp\Big\{-\eta n_{\text{tr}}\big(\Delta-\text{EstErr}(\bar{t})\big)+\eta^2\Big(\frac{17w_{\max}^2n_{\text{tr}}}{8}+\frac{n_{\text{tr}}^2}{4n_{\text{test}}}\Big)\Big\}.\tag{15}$$

We choose η to minimize the upper bound above and get

$$\eta=\frac{\Delta-\text{EstErr}(\bar{t})}{\frac{17w_{\max}^2}{4}+\frac{n_{\text{tr}}}{2n_{\text{test}}}},\text{ and correspondingly, (15)}\leq\exp\bigg\{-\frac{\big(\Delta-\text{EstErr}(\bar{t})\big)^2}{\frac{9w_{\max}^2}{n_{\text{tr}}}+\frac{1}{n_{\text{test}}}}\bigg\}.$$

Recall that $\bar{t} = q^*(\Delta) - \varepsilon$. Since $\Delta \le \bar{\Delta}$, $\bar{t} \ge q^*(\bar{\Delta}) - \varepsilon$. For ε sufficiently small, we further have $q^*(\Delta) - \varepsilon > q^*(\bar{\Delta}) \cdot \kappa$. By the definition of Δ, we have

$$\mathrm{EstErr}(\bar{t}) + \sqrt{\log\Big(\frac{1}{\delta}\Big) \cdot \Big(\frac{9w_{\max}^2}{n_{\text{tr}}} + \frac{1}{n_{\text{test}}}\Big)} \le \sup_{\kappa \cdot q^*(\bar{\Delta}) \le t \le q^*(0)} \mathrm{EstErr}(t) + \sqrt{\log\Big(\frac{1}{\delta}\Big) \cdot \Big(\frac{9w_{\max}^2}{n_{\text{tr}}} + \frac{1}{n_{\text{test}}}\Big)} = \Delta.$$

Consequently, we arrive at $\mathbb{P}\big(\widehat{q} \le q^*(\Delta) - \varepsilon\big) \le \delta$. Taking $\varepsilon \to 0$ and by the continuity of the probability measure, we have that $\mathbb{P}\big(\widehat{q} < q^*(\Delta)\big) \le \delta$.

Step (II). Let $A = \{S_{n+j} \le q^*(\Delta)\}$. As in the proof of Theorem 3.1, we have that

$$\mathbb{P}_{Q_{X,Y}}\big(S_{n+j} \le q^*(\Delta)\big) \ge g_{f,\rho}\big(\mathbb{P}_{Q_X \times P_{Y\,|\,X}}\big(S_{n+j} \le q^*(\Delta)\big)\big) \ge g_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha) - \Delta\big).$$

On the event $\{\widehat{q} \ge q^*(\Delta)\}$,

$$\begin{aligned}
\mathbb{P}_{Q_{X,Y}}\big(S_{n+j} \le \widehat{q}\,\big|\,\mathcal{D}_{\text{tr}}^{(1)}, \mathcal{D}_{\text{test},j}^{(1)}\big) &\ge \mathbb{P}_{Q_{X,Y}}\big(S_{n+j} \le q^*(\Delta)\big) \\
&\ge g_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha) - \Delta\big) \\
&\ge g_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big) - g_{f,\rho}'\big(g_{f,\rho}^{-1}(1-\alpha)\big)\Delta,
\end{aligned}$$

where the last step follows from the convexity of $g_{f,\rho}$ and the separating hyperplane theorem. As proved in Theorem 3.1, when $g_{f,\rho}(1) \ge 1-\alpha$, $g_{f,\rho}(g_{f,\rho}^{-1}(1-\alpha)) \ge 1-\alpha$ and we complete the proof. □

Proof of Theorem 3.7 In this proof, we write $\Delta(\delta)$ instead of Δ to emphasize the dependence of Δ on δ. By Theorem B.1, we know that for any $\delta \in (0,1)$, $\mathbb{P}\big(\widehat{q} < q^*\big(\Delta(\delta)\big)\big) \le \delta$.

In the following, we shall consider a sequence of $\delta \in \{2^{-\ell}\}_{\ell=0}^{\infty}$. For each $\ell \in \mathbb{N}$, we let $q_\ell = q^*\big(\Delta(2^{-\ell})\big)$.

$$\begin{aligned}
&\mathbb{P}_{Q_X \times P_{Y\,|\,X}}\big(S_{n+j} \le \widehat{q}\big) - g_{f,\rho}^{-1}(1-\alpha) \qquad (16)\\
&= \sum_{\ell=0}^{\infty} \mathbb{E}\Big[\mathbb{1}\{q_{\ell+1} \le \widehat{q} < q_\ell\} \cdot \Big(\mathbb{P}_{Q_X \times P_{Y\,|\,X}}\big(S_{n+j} \le \widehat{q}\,\big|\,\mathcal{D}_{\text{tr}}^{(k)}, \mathcal{D}_{\text{test},j}^{(k)}\big) - g_{f,\rho}^{-1}(1-\alpha)\Big)\Big] \\
&\qquad\qquad + \mathbb{E}\Big[\mathbb{1}\big\{\widehat{q} \ge q_0\big\} \cdot \Big(\mathbb{P}_{Q_X \times P_{Y\,|\,X}}\big(S_{n+j} \le \widehat{q}\,\big|\,\mathcal{D}_{\text{tr}}^{(k)}, \mathcal{D}_{\text{test},j}^{(k)}\big) - g_{f,\rho}^{-1}(1-\alpha)\Big)\Big] \\
&\ge \sum_{\ell=0}^{\infty} \mathbb{E}\Big[\mathbb{1}\{q_{\ell+1} \le \widehat{q} < q_\ell\} \cdot \big(F(q_{\ell+1}) - g_{f,\rho}^{-1}(1-\alpha)\big)\Big] + \mathbb{E}\Big[\mathbb{1}\big\{\widehat{q} \ge q_0\big\} \cdot \big(F(q_0) - g_{f,\rho}^{-1}(1-\alpha)\big)\Big] \\
&\ge -\sum_{\ell=0}^{\infty} \Delta(2^{-\ell-1}) \cdot \mathbb{P}\big(q_{\ell+1} \le \widehat{q} < q_\ell\big) - \Delta(0) \cdot \mathbb{P}(\widehat{q} \ge q_0),
\end{aligned}$$

where the last step is due to the definition of q_ℓ. Since for any $\ell \in \mathbb{N}$, $\mathbb{P}(\widehat{q} < q^*(\Delta(2^{-\ell}))) \le 2^{-\ell}$, we further have

$$\begin{aligned}
(16) &\ge -\sum_{\ell=0}^{\infty} \Delta(2^{-\ell})\mathbb{P}\big(\widehat{q} < q^*\big(\Delta(2^{-\ell-1})\big)\big) - \Delta(0) \\
&\ge -\sum_{\ell=0}^{\infty} 2^{-\ell-1}\Delta(2^{-\ell-1}) - \Delta(0) \\
&\ge \sup_{\kappa q^* \bar{\Delta} \le t \le q^*(0)} 2 \cdot \mathrm{EstErr}(t) + \sqrt{\frac{16w_{\max}^2}{n_{\text{tr}}} + \frac{2}{n_{\text{test}}}}.
\end{aligned}$$

We now return to the coverage under $Q_{X,Y}$. Again using the step in the proof of Theorem 3.1, we have

$$\begin{aligned}\mathbb{P}_{Q_{X,Y}}(S_{n+j} \leq \widehat{q}) &\geq g_{f,\rho}\big(\mathbb{P}_{Q_X \times P_{Y\,|\,X}}(S_{n+j} \leq \widehat{q})\big) \\ &\geq g_{f,\rho}\Bigg(g_{f,\rho}^{-1} - \sup_{\kappa q^* \bar{\Delta} \leq t \leq q^*(0)} 2 \cdot \text{EstErr}(t) - \sqrt{\frac{16 w_{\max}^2}{n_{\text{tr}}} + \frac{2}{n_{\text{test}}}}\Bigg) \\ &\geq g_{f,\rho}\big(g_{f,\rho}^{-1}(1-\alpha)\big) - g'_{f,\rho}(g_{f,\rho}^{-1}(1-\alpha)) \cdot \Bigg(\sup_{\kappa q^* \bar{\Delta} \leq t \leq q^*(0)} 2 \cdot \text{EstErr}(t) + \sqrt{\frac{16 w_{\max}^2}{n_{\text{tr}}} + \frac{2}{n_{\text{test}}}}\Bigg)\end{aligned}$$

When $g_{f,\rho}(1) \geq 1-\alpha$, $g_{f,\rho}(g_{f,\rho}^{-1}(1-\alpha)) \geq 1-\alpha$. The proof is thus completed.

B.4. Proof of theorem 3.8

Throughout, we condition on $\widehat{\rho}$ and $\mathcal{D}_{\text{tr}}^{(1-k)} \cup \mathcal{D}_{\text{test}}^{(1-k)}$. Fix $k \in \{0,1\}$ and $\{n+j\} \in \mathcal{I}_{\text{test}}^{(k)}$. Define $A = \big\{Y_{n+j} \in \widehat{C}_{f,\widehat{\rho},n+j}(X_{n+j})\big\}$. By the proof of Theorem 1, we have that

$$\mathbb{P}_{(X_{n+j},Y_{n+j})\sim Q_{XY}}(A) \geq g_{f,\rho^*}\Big(\mathbb{P}_{(X_{n+j},Y_{n+j})\sim Q_X \times P_{Y\,|\,X}}(A)\Big).$$

Next, by the intermediate steps in the proof of Theorem 3.4, there is

$$\begin{aligned}P_{(X_{n+j},Y_{n+j})\sim Q_X \times P_{Y\,|\,X}}(A) &\geq P_{(X_{n+j},Y_{n+j})\sim \tilde{Q}_X \times P_{Y\,|\,X}}(A) - \frac{1}{2}\mathbb{E}_{X\sim P_X}\left[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{X\sim P_X}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\right] \\ &\geq g_{f,\widehat{\rho}}^{-1}(1-\alpha) - \frac{1}{2}\mathbb{E}_{X\sim P_X}\left[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{X\sim P_X}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\right].\end{aligned}$$

Combining the above inequalities and since the monotonicity of $g_{f,\rho^*}(\beta)$ in β, we have

$$\begin{aligned}&\mathbb{P}_{(X_{n+j},Y_{n+j})\sim Q_{XY}}\big(Y_{n+j} \in \widehat{C}_{f,\widehat{\rho},n+j}(X_{n+j})\big) \\ &\geq g_{f,\rho^*}\left(g_{f,\widehat{\rho}}^{-1}(1-\alpha) - \frac{1}{2}\mathbb{E}_{X\sim P_X}\left[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{P_X}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\right]\right) \\ &\geq g_{f,\rho^*}\big(g_{f,\widehat{\rho}}^{-1}(1-\alpha)\big) - \frac{1}{2}g'_{f,\rho^*}(g_{f,\widehat{\rho}}^{-1}(1-\alpha)) \cdot \mathbb{E}_{X\sim P_X}\left[\Big|\frac{\widehat{w}^{(k)}(X)}{\mathbb{E}_{P_X}[\widehat{w}^{(k)}(X)]} - w(X)\Big|\right].\end{aligned}$$

When $\widehat{\rho} \geq \rho^*$, $g_{f,\rho^*}(g_{f,\widehat{\rho}}^{-1}(1-\alpha)) \geq g_{f,\widehat{\rho}}(g_{f,\widehat{\rho}}^{-1}(1-\alpha))$. The latter is greater or equal to $1-\alpha$ when $g_{f,\widehat{\rho}}(1) \geq 1-\alpha$, following the proof of Theorem 3.1. The proof is therefore completed.

C. Addtional Algorithmic Details

This section contains the complete algorithmic details of the proposed methods. Algorithm 1 corresponds to the weighted robust conformal prediction (WRCP) method, while Algorithm 2 corresponds to the debiased weighted robust conformal prediction (D-WRCP) method.

D. Additional results of sensitivity analysis under the f sensitivity model

This section collects additional results of adapting our method to the sensitivity analysis of ITE under the f-sensitivity model.

Suppose that the inferential target is $Y(t_1)$ for $t_1 \in \{0,1\}$; and the target population is $T = t_2$, where $t_2 \in \{0,1,\circ\}$, with $\circ$ denoting the whole population. The prediction interval $\widehat{C}_{f,\rho}(X_{n+1})$ should satisfy

$$\mathbb{P}\big(Y_{n+1}(t_1) \in \widehat{C}_{f,\rho}(X_{n+1}) \,\big|\, T = t_2\big) \geq 1-\alpha.$$

Given a set of training data $\mathcal{D}_{\text{tr}} = \{(X_i, T_i, Y_i)\}_{i=1}^n$, we start as before by randomly splitting the data into two folds $\mathcal{D}_{\text{tr}}^{(0)}$ and $\mathcal{D}_{\text{tr}}^{(1)}$. The first fold $\mathcal{D}_{\text{tr}}^{(0)}$ is used for fitting the propensity score function $\widehat{e}(\text{x})$ (if unknown); we also use the unit in $\mathcal{D}_{\text{tr}}^{(0)}$

Algorithm 1 Weighted robust conformal prediction (WRCP)

```
Input: Training set $\mathcal{D}_{\text{tr}} = \{(X_i, Y_i)\}_{i=1}^n$; test data $\mathcal{D}_{\text{test}} = \{X_{n+j}\}_{j=1}^m$; regression algorithm $\mathcal{A}$; classification algorithm $\mathcal{C}$;
       target miscoverage level $\alpha \in (0,1)$; score function $s(x, y; \mu)$; robust parameter $\rho$.
Optional input: likelihood ratio function $w(x)$.

Randomly split $\mathcal{D}_{\text{tr}}$ into two disjoint subsets of equal sizes, $\mathcal{D}_{\text{tr}}^{(0)}$ and $\mathcal{D}_{\text{tr}}^{(1)}$, indexed by $\mathcal{I}_{\text{tr}}^{(0)}$ and $\mathcal{I}_{\text{tr}}^{(1)}$, respectively
Apply $\mathcal{A}$ to $\mathcal{D}_{\text{tr}}^{(0)}$ and obtain the prediction function: $\widehat{\mu} \leftarrow \mathcal{A}(\mathcal{D}_{\text{tr}}^{(0)})$
Compute the nonconformity score $S_i = s(X_i, Y_i)$ for $i \in \mathcal{I}_{\text{tr}}^{(1)}$
if $w(x)$ exists then
    for $j = 1, \ldots, m$ do
        Construct $\widehat{C}_{f,\rho,n+j}(X_{n+j})$ according to (3)
    end
else
    Split $\mathcal{D}_{\text{test}}$ into two disjoint subsets of equal sizes, $\mathcal{D}_{\text{test}}^{(0)}$ and $\mathcal{D}_{\text{test}}^{(1)}$, indexed by $\mathcal{I}_{\text{test}}^{(0)}$ and $\mathcal{I}_{\text{test}}^{(1)}$, respectively
    for $k = 0, 1$ do
        Train a classifier: $\widehat{\mathbb{P}}^{(k)}(A = 1 \mid X = x) \leftarrow \mathcal{C}(\mathcal{D}_{\text{tr}}^{(0)}, \mathcal{D}_{\text{test}}^{(1-k)})$ Construct the estimator $\widehat{w}^{(k)}(x) \leftarrow \frac{\widehat{\mathbb{P}}^{(k)}(A=1 \mid X=x)}{1-\widehat{\mathbb{P}}^{(k)}(A=1 \mid X=x)}$
        for $\ell \in \mathcal{I}_{\text{test}}^{(k)}$ do
            Construct $\widehat{C}_{f,\rho,\ell}(X_\ell)$ according to (3) with $w(x)$ replaced by $\widehat{w}^{(k)}(x)$
        end
    end
end

Output: Prediction sets $\{\widehat{C}_{f,\rho,n+j}(X_{n+j})\}_{j \in [m]}$.
```

such that $T = t_1$ to fit a function $\widehat{\mu}^{(t_1)}$ for predicting $Y(t_1)$. The form of the X shift weight function $w^{(t_1,t_2)}(x)$ is listed in Table 1. Next, for any $j \in [m]$, the prediction interval is constructed as

$$\widehat{C}_{f,\rho,n+j}^{(t_1,t_2)}(x) = \left\{ y \in \mathbb{R} : s(x,y) \leq \text{Quantile}\Big(g_{f,\rho}^{-1}(1-\alpha), \sum_{i \in \mathcal{D}_{\text{tr}}^{(1)}, T_i = t_2} p_i^{(t_1,t_2)}(x) \cdot \delta_{S_i} + p_{n+1}^{(t_1,t_2)}(x) \cdot \delta_\infty \Big) \right\},$$

$$\text{where } p_i^{(t_1,t_2)}(x) = \frac{\widehat{w}^{(t_1,t_2)}(X_i)}{\sum_{j \in \mathcal{D}_{\text{tr}}^{(1)}, T_j = t_2} \widehat{w}^{(t_1,t_2)}(X_j) + \widehat{w}^{(t_1,t_2)}(x)},$$

$$p_{n+1}^{(t_1,t_2)}(x) = \frac{\widehat{w}^{(t_1,t_2)}(x)}{\sum_{j \in \mathcal{D}_{\text{tr}}^{(1)}, T_j = t_2} \widehat{w}^{(t_1,t_2)}(X_j) + \widehat{w}^{(t_1,t_2)}(x)}. \tag{17}$$

Above, $\widehat{w}^{(t_1,t_2)}$ is the estimator for $w^{(t_1,t_2)}$. The complete procedure can be found in Algorithm 3.

t_2 t_1	1	0	○
1	1	$\frac{1-e(x)}{e(x)} \cdot \frac{p_1}{p_0}$	$\frac{p_1}{e(x)}$
0	$\frac{e(x)}{1-e(x)} \cdot \frac{p_0}{p_1}$	1	$\frac{p_0}{1-e(x)}$

Table 1. The form of the X shift function $w^{(t_1,t_2)}(x)$. The function $e(x) = \mathbb{P}(T = 1 \mid X = x)$ is the observed propensity score function, $p_1 = \mathbb{P}(T = 1)$, and $p_0 = \mathbb{P}(T = 0)$.

Algorithm 2 Debiased weighted robust conformal prediction (D-WRCP)

```
Input: Training set $\mathcal{D}_{\text{tr}} = \{(X_i, Y_i)\}_{i=1}^n$; test data $\mathcal{D}_{\text{test}} = \{X_{n+j}\}_{j=1}^m$ regression algorithm $\mathcal{A}$; classification algorithm
       $\mathcal{C}$; conditional CDF fitting algorithm $\mathcal{M}$ target miscoverage level $\alpha \in (0,1)$; score function $s(x,y;\mu)$; robust
       parameter $\rho$.

Randomly split $\mathcal{D}_{\text{tr}}$ into two disjoint subsets of equal sizes, $\mathcal{D}_{\text{tr}}^{(0)}$ and $\mathcal{D}_{\text{tr}}^{(1)}$, indexed by $\mathcal{I}_{\text{tr}}^{(0)}$ and $\mathcal{I}_{\text{tr}}^{(1)}$, respectively
Randomly split $\mathcal{D}_{\text{test}}$ into two disjoint subsets of equal sizes, $\mathcal{D}_{\text{test}}^{(0)}$ and $\mathcal{D}_{\text{test}}^{(1)}$, indexed by $\mathcal{I}_{\text{test}}^{(0)}$ and $\mathcal{I}_{\text{test}}^{(1)}$, respectively
for $k = 0, 1$ do
    Obtain the prediction function: $\widehat{\mu}^{(k)} \leftarrow \mathcal{A}(\mathcal{D}_{\text{tr}}^{(1-k)})$
    Compute the nonconformity score $S_i = s(X_i, Y_i; \widehat{\mu}^{(k)})$ for $i \in \mathcal{I}_{\text{tr}}^{(k)}$
    Train a classifier $\widehat{\mathbb{P}}^{(k)}(A = 1 \mid X = x) \leftarrow \mathcal{C}(\mathcal{D}_{\text{tr}}^{(1-k)}, \mathcal{D}_{\text{test}}^{(1-k)})$
    Construct the estimator for $X$ shift $\widehat{w}^{(k)}(x) \leftarrow \frac{\widehat{\mathbb{P}}^{(k)}(A=1 \mid X=x)}{1-\widehat{\mathbb{P}}^{(k)}(A=1 \mid X=x)}$
    Obtain the estimated conditional CDF of $S$: $\widehat{m}^{(k)} \leftarrow \mathcal{M}(\mathcal{D}_{\text{tr}}^{(k)})$
    for $\ell \in \mathcal{I}_{\text{test}}^{(k)}$ do
        Construct $\widehat{C}^{\text{DR}}_{f,\rho,\ell}(X_\ell)$ according to (4);
    end
end

Output: Prediction sets $\{\widehat{C}^{\text{DR}}_{f,\rho,n+j}(X_{n+j})\}_{j\in[m]}$.
```

E. Additional Simulation Results

This section collects the remaining details of the simulation in Section 5 and results from additional simulation studies. In specific, we evaluate the candidate methods under a pure $Y \mid X$ shift setting, a pure X shift setting, and a nonlinear model setting; the effect of the choice of f is also studied.

E.1. Additional Details of Simulation in Section 5.

For the training data, $X \sim \mathcal{N}(0, I_{50})$ and $Y \mid X \sim X^\top \beta + \mathcal{N}(0,1)$, where $\|\beta\|_0 = 10$ and the nonzero entries take the value 0.47. The target covariate distribution has a shifted mean: $Q_X = \mathcal{N}(\beta_0, I_{50})$, with $\beta_0 = (\eta, -\eta, 0, \cdots, 0)$, where η is a tuning parameter controlling the amount of X shift; in the simulation, we take η to be 0.1, 0.5, and 0.8, corresponding

Algorithm 3 Conformalized counterfactual inference under the f-sensitivity model

```
Input: Training set $\mathcal{D}_{\text{tr}} = \{(X_i, T_i, Y_i)\}_{i=1}^n$; test data $\{X_{n+j}\}_{j=1}^m$ counterfactual type $t_1 \in \{0,1\}$, target population
       $t_2 \in \{0, 1, \circ\}$ outcome fitting algorithm $\mathcal{A}$; propensity score fitting algorithm $\mathcal{E}$ target miscoverage level $\alpha \in (0,1)$;
       score function $s(x,y;\mu)$; sensitivity parameter $\rho$.

Randomly split $\mathcal{D}_{\text{tr}}$ into two subsets of equal sizes $\mathcal{D}_{\text{tr}}^{(0)}$ and $\mathcal{D}_{\text{tr}}^{(1)}$, indexed by $\mathcal{I}_{\text{tr}}^{(0)}$ and $\mathcal{I}_{\text{tr}}^{(1)}$
Apply $\mathcal{A}$ to obtain the outcome regression functions: $\widehat{\mu}^{(t_1)} \leftarrow \mathcal{A}\big(\{(X_i, Y_i) : i \in \mathcal{I}_{\text{tr}}^{(0)}, T_i = t_1\}\big)$
Apply $\mathcal{E}$ to obtain the estimated propensity score function: $\widehat{e} \leftarrow \mathcal{E}(\mathcal{D}_{\text{tr}}^{(0)})$
Compute the nonconformity scores: $S_i = s(X_i, Y_i; \widehat{\mu}^{(t_1)})$ for $i \in \mathcal{I}_{\text{tr}}^{(1)}$ such that $T_i = t_1$
for $\ell \in \mathcal{I}_{\text{test}}$ do
    Construct $\widehat{C}^{(t_1,t_2)}_{f,\rho,\ell}(X_\ell)$ according to (17)
end
Output: Prediction sets $\{\widehat{C}^{(t_1,t_2)}_{f,\rho,n+j}(X_{n+j})\}_{j\in[m]}$.
```

to low, medium, and high X shifts, respectively. The target $Y \mid X$ distribution is specified via

$$\frac{\mathrm{d}Q_{Y \mid X}}{\mathrm{d}P_{Y \mid X}}(x) = \begin{cases} 0.96 & \text{if } \left|Y - X^\top \beta\right| < 1.86; \\ 1.59 & \text{if } \left|Y - X^\top \beta\right| \geq 1.86, \end{cases}$$

For all the candidate methods, we implement the split version, where half of the data is reserved for model fitting and the other half for calibration. The nonconformity score is $s(x, y) = |y - \widehat{\mu}(x)|$, where we fit $\widehat{\mu}(\cdot)$ with cross-validated Lasso (Tibshirani, 1996) using the `scikit-learn` package in python (Pedregosa et al., 2011). For `WCP`, `WRCP` and `D-WRCP`, the covariate likelihood ratio $w(x)$ is estimated via the random forest classifier (Breiman, 2001) in the `scikit-learn` package. For `WRCP`, `D-WRCP`, we use the KL divergence to quantify the distributional shift, i.e., $f(t) = t \log t$, and consider a sequence of robust parameters ρ. For each ρ, the corresponding robust parameter of `RCP` is chosen as $\rho_{\text{RCP}} = \rho + D_{\text{KL}}(Q_X \,\|\, P_X)$ by the chain rule of KL divergence, where $D_{\text{KL}}(Q_X \,\|\, P_X)$ is estimated by plugging in the estimated $\widehat{w}$. For `D-WRCP`, the conditional CDF is estimated by random forest with the python package `qosa-indices` (Elie-Dit-Cosaque, 2020).

E.2. Pure $Y \mid X$ Shift.

We consider a setting where the distributional shift is purely in the conditional relationship between Y and X while the covariate distribution remains invariant, i.e. $P_X = Q_X$. The other settings are the same as in Section 5. The corresponding results are presented in Figure 4. Since the distributional shift is purely in $Y \mid X$, `WCP` suffers the same degree of coverage drop as `CP`. `WRCP` and `D-WRCP` provide valid and tight prediction intervals. The prediction interval of `RCP` is overly conservative due to the estimation error of the X shift.

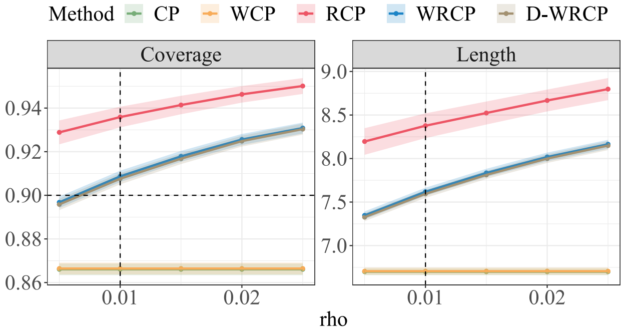

Figure 4. Averaged coverage (left) and prediction interval length (right) under pure $Y \mid X$ shift. The details are otherwise the same as in Figure 1.

E.3. Pure X Shift.

We now consider a pure covariate shift, fixing the conditional relationship between Y and X. The other setup is the same as in Section 5, except that we take η (the X shift parameter) to be 0.1. The results are shown in Figure 5, where `WCP` achieves the desired coverage level; our proposed method is a bit conservative, but less severe than `RCP`.

E.4. Nonlinear Models.

Next, we consider a nonlinear relationship between Y and X. Under P,

$$Y = \frac{1}{(1 + e^{X_1})(1 + e^{-X_2})} + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, 1).$$

The X shift parameter η is set to be 0.1, and all the other settings are the same as in Section 5. The corresponding results are presented in Figure 6, where the message is similar to the linear case: `WCP` is not able to address the distributional shift, `RCP` is overly conservative, while our proposed methods provide valid and tight prediction intervals.

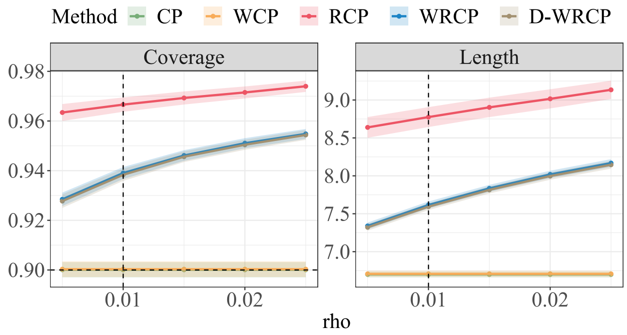

Figure 5. Averaged coverage (left) and prediction interval length (right) under pure X shift. The details are otherwise the same as in Figure 1.

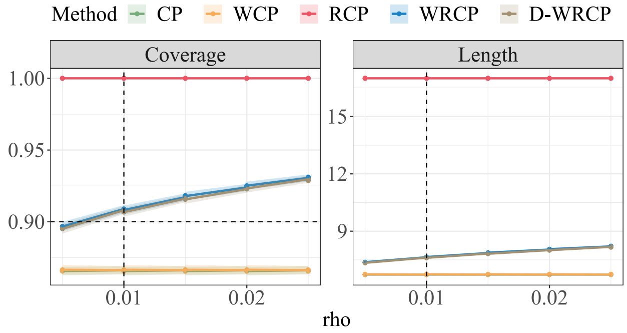

Figure 6. Averaged coverage (left) and prediction interval length (right) under the nonlinear model experiment. The details are otherwise the same as in Figure 1.

E.5. Effect of f.

Finally, we investigate the effect of the choice of f on the performance of our proposed methods. We consider the same setup as in Section 5 with $\rho^* = 0.01$ (under KL divergence) and $\eta = 0.5$. We vary the choice of f in the implementation of `WRCP` to be (1) $f(x) = x \log x$ (KL divergence), (2) $f(x) = \frac{1}{2}|x-1|$ (TV distance), and (3) $f(x) = (x-1)^2$ (χ^2 divergence). Figure 7 demonstrates the averaged coverage and prediction interval length under different choices of f. The results show that the choice of f has a rather mild effects on the performance of `WRCP`.

F. Addtional Real Data Results

F.1. Implementation Details for the NSLM Dataset

Following the strategy of Carvalho et al. (2019), we generate synthetic potential outcomes from the following model:

$$Y(t) = \mu(x) + \tau(x_1, x_2, c_1) \cdot t + \epsilon, \text{ for } t = 0, 1.$$

Above, x denotes all the covariates for a student, and x_1, x_2, c_1 corresponds to X1, X2 and C1, respectively. The baseline function μ is obtained by fitting a generalized additive model (Hastie, 2017) on the control arm of the original data, and ϵ is sampled with replacement from the sum of the residual from the fitted model on the original data and a noise term $\mathcal{N}(0, 0.025)$. The form of the treatment effect is:

$$\tau(x_1, x_2, c_1) = 0.228 + 0.05 \cdot \mathbb{1}\{x_1 < 0.07\} - 0.05 \cdot \mathbb{1}\{x_2 < -0.69\} - 0.08 \cdot \mathbb{1}\{c_1 \in \{1, 13, 14\}\}.$$

Confounding is introduced by removing X1 and X2.

For each run, we randomly select half of the treated units and half of the control units for model fitting; the other half of the treated units are reserved for calibration, while the other half of the control units for evaluation. The nonconformity

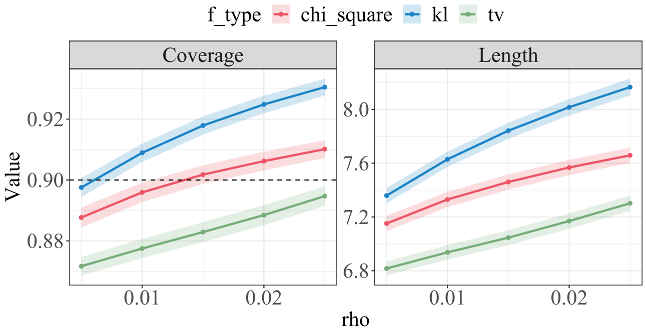

Figure 7. Averaged coverage (left) and prediction interval length (right) with different choices of f. The details are otherwise the same as in Figure 1.

score function $s(x,y) = |y - \widehat{\mu}(x)|$. Both the regression function $\widehat{\mu}(x)$ and the propensity score function $e(x)$ are fitted with random forest. With each ρ, we repeat the above process for 100 random splits.

F.2. Implementation Details for the ACS Income Dataset

Since Y is binary, we adopt the generalized inverse quantile conformity score introduced by Romano et al. (2020), and the prediction set is a subset of $\{0,1\}$. The weight function $\widehat{w}$ is estimated via XGBoost (Chen and Guestrin, 2016) with the hyperparameters provided by Liu et al. (2023), and the outcome model $\widehat{\mu}$ is fitted with random forest.

F.3. COVID Information Studies

The covid information studies investigate how a “nudge” for thinking about the accuracy of information can affect the people’s ability to discern fake news when sharing COVID-related headlines. The original study of 856 participants (Pennycook et al., 2020) is first conducted, followed by a replication study (Roozenbeek et al., 2021) of 1,583 participants. The original study found a significant interaction term between the intervention and the validity of the headline, while the replication study also found a significant interaction, but with a much smaller magnitude (more details about the comparison between the two studies can be found in Jin et al. (2023a)).

As discussed in Jin et al. (2023a), the discrepancy between the two results can be attributed to the distributional shift in the both the covariates and $Y \mid X$. Here, instead of estimating the treatment effect, we consider the task of predicting a participant’s rating for willingness to share a headline. Each sample in the dataset corresponds to a participant, where the outcome is the rating for their willingness to share a headline; the predictors include the treatment status (i.e., whether a nugde is sent), the validity of the news, and 10 other covariates.[4] After removing the samples with missing values, the training and test set consist of 811 and 1,583 samples, respectively. Each run splits the training and test sets into two halves for model fitting and calibration; The weight function $\widehat{w}$ and the outcome model $\widehat{\mu}$ are both estimated via random forest. The robust parameter $\rho \in \{0.005, 0.01, \ldots, 0.04\}$, and for each ρ We repeat the above process for 100 random splits.

Figure 8 demonstrates the results of all methods. For the purpose of visualization, we replace the infinite prediction interval length with 2 (an upper bound of all the finite realized lengths) when plotting the averaged prediction interval length. In this example, we again see that `CP` and `WCP` fail to achieve the desired coverage level, with `WCP` being slightly better than `CP` due to the adjustment for X shift. The proposed methods `WRCP` and `D-WRCP` achieve approximate coverage for a wide range of ρ’s, and are much more efficient than `RCP`.

F.4. Poverty Estimation with Satellite Images

In this example, we evaluate the candidate methods on the poverty mapping dataset (Yeh et al., 2020), where the task is to estimate the poverty rate in different regions of the world with the help of satellite images. The data is obtained and preprocessed with the WILDS python package (Koh et al., 2021).

[4] In the original datasets, each participant was asked to rate 30 headlines. In our analysis, the outcome and the covariates are all averaged over the 30 headlines.

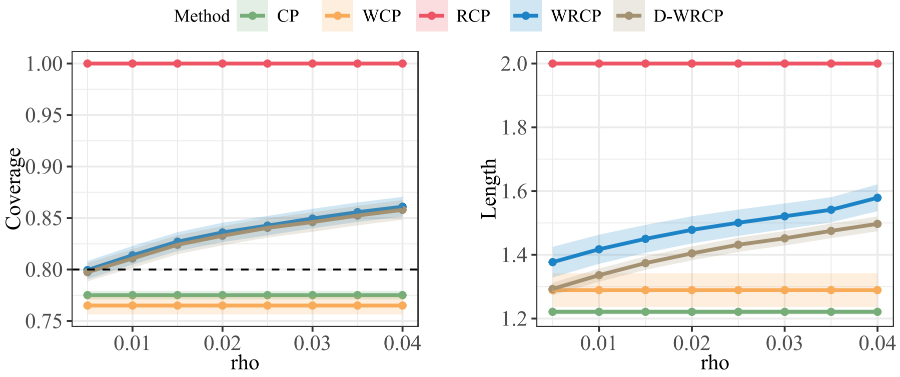

Figure 8. Averaged coverage (left) and prediction interval length (right) over 100 runs as a function of the robust parameter ρ from the experiment on covid study datasets. The other details are the same as in Figure 2.

The covariate X corresponds to a 224×224-pixel image, and the outcome Y is a real-valued asset wealth index (computed from Demographic and Health Surveys data). In our implementation, the training set consists of 500 samples, and the test set consists of 200 samples. We fit a convolutional neural network (CNN) for $\widehat{\mu}$; for $\widehat{w}$, the variational autoencoder (VAE) is used to learn a representation of the images, upon which a random forest classifier is trained to estimate the X shift.

We consider $\rho \in \{0.1, 0.12, \ldots, 0.18\}$, repeating the above process for 100 random splits under each ρ. Figure 9 shows the results of `CP`, `WCP`, `RCP`, and `WRCP`. We again see that `CP` and `WCP` show significant under-coverage, while `RCP` is overly conservative. `WRCP` delivers valid and tight prediction intervals for a wide range of ρ's.

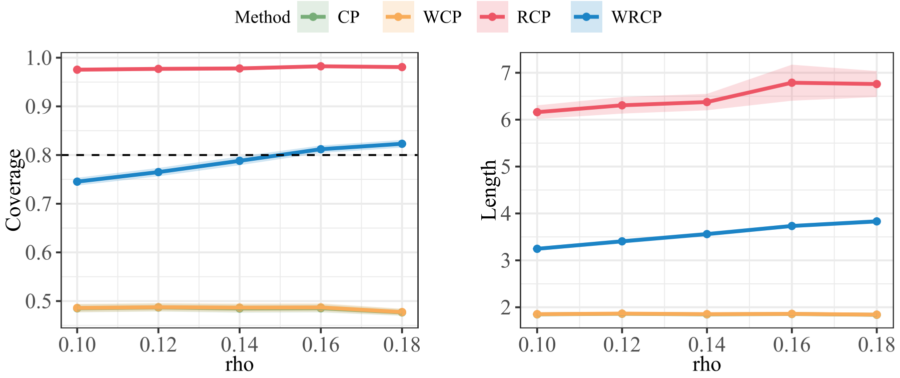

Figure 9. Averaged coverage (left) and prediction interval length (right) over 100 runs as a function of the robust parameter ρ from the experiment on poverty estimation. The other details are the same as in Figure 2.