Doubly Robust Calibration of Prediction Sets under Covariate Shift

Yachong Yang*[1], Arun Kumar Kuchibhotla†[2], and Eric Tchetgen Tchetgen‡[1]

[1]Department of Statistics, University of Pennsylvania
[2]Department of Statistics & Data Science, Carnegie Mellon University

December 14, 2022

Abstract

Conformal prediction has received tremendous attention in recent years and has offered new solutions to problems in missing data and causal inference; yet these advances have not leveraged modern semiparametric efficiency theory for more robust and efficient uncertainty quantification. In this paper, we consider the problem of obtaining distribution-free prediction regions accounting for a shift in the distribution of the covariates between the training and test data. Under an *explainable covariate shift* assumption analogous to the standard missing at random assumption, we propose three variants of a general framework to construct well-calibrated prediction regions for the unobserved outcome in the test sample. Our approach is based on the efficient influence function for the quantile of the unobserved outcome in the test population combined with an arbitrary machine learning prediction algorithm, without compromising asymptotic coverage. We establish that the resulting prediction sets eventually attain nominal coverage in large samples. This guarantee is a consequence of the product bias form of our proposal which implies correct coverage if either the propensity score or the conditional distribution of the response is estimated sufficiently well. Our results also provide a framework for construction of doubly robust prediction sets of individual treatment effects, under the unconfoundedness condition. We further discuss aggregation of prediction sets from different machine learning algorithms for optimal prediction and illustrate the performance of our methods in both synthetic and real data. Finally, inspired by sensitivity analysis in missing data, we briefly discuss how our proposal could be extended to account for departures from the explainable covariate shift setting.

1 Introduction

Prediction is a major focus of modern machine learning literature. Most machine learning methods are designed for point prediction, but accurately quantifying the uncertainty associated with a given point prediction algorithm remains an important challenge in many applications. Given independent and identically distributed (i.i.d.) pairs (X_i, Y_i), $i = 1, \ldots, N$, from a distribution $P = P_X \otimes P_{Y|X}$ supported on $\mathcal{X} \times \mathbb{R}$ (e.g., $\mathcal{X} = \mathbb{R}^d, d \geq 1$), and given a desired nominal coverage rate $1 - \alpha \in (0, 1)$, the goal of prediction with well-calibrated uncertainty quantification is to build a prediction set $\widehat{C}_{N,\alpha}$, such that

$$\mathbb{P}\big(Y_f \in \widehat{C}_{N,\alpha}(X_f)\big) \geq 1 - \alpha, \tag{1}$$

where the probability is taken over the marginal distribution of all the training data along with (X_f, Y_f). Note that (1) does not imply conditional coverage $\mathbb{P}\big(Y_f \in \widehat{C}_{N,\alpha}(X_f) | X_f = x_f\big) \geq 1 - \alpha$, which is known

*Email and address: yachong@wharton.upenn.edu, Academic Research Building, 265 S 37th St, Philadalphia, PA, US.
†Email and address: arunku@cmu.edu, Baker Hall, 4909 Frew St, Pittsburgh, PA, US.
‡Email and address: ett@wharton.upenn.edu, Academic Research Building, 265 S 37th St, Philadalphia, PA, US.

to be impossible without assumptions over the underlying distribution as shown in Barber et al. (2019). This goal, however, can be *approximately* achieved where *approximately* is meant either asymptotically or by conditioning on X_f belonging to a set A rather than X_f being equal to, say x_f. Conformal prediction introduced by Vovk et al. (2005) provides a simple and finite-sample valid solution to (1) without any assumption on the distribution P, requiring only that the training data and (X_f, Y_f) to be exchangeable (jointly). It provides a valid prediction set by wrapping around any point prediction algorithm, irrespective of what the point prediction algorithm is.

Recently several works have considered the extension of conformal prediction methodology to the case of non-exchangeable data; see Section 2 for a review. Our work lies in this space. One important distinction from the exchangeable case is that finite sample coverage guarantees are generally impossible for non-exchangeable data without very restrictive assumptions. Formally, we consider the problem of prediction with uncertainty quantification under covariate shift where the completely observed training data is drawn i.i.d. from $P_X \otimes P_{Y|X}$ but the test point that needs to be covered by a prediction set is drawn from $Q_X \otimes P_{Y|X}$. This problem, which we term *explainable covariate shift* problem, was first posed by Tibshirani et al. (2019) who provides a solution called "Weighted Conformal Prediction" (WCP) which assumes that the covariate shift is known via dQ_X/dP_X. This method has been extended by Lei and Candès (2020) to allow for unknown covariate shift.

The prediction problem under covariate shift can be equivalently stated as predicting the label for a given feature vector observed from a different covariate distribution. In terms of real-world applications, prediction under covariate shift is important in semi-supervised and transfer learning settings. In health care and related problems, it is often the case that the amount of labeled data is limited in comparison to unlabeled data. This is, particularly, true with electronic health record (EHR) data where labeling the response can be costly and/or laborious (Chakrabortty and Cai, 2018). Instead of assuming the data are missing completely at random (i.e., observations are chosen to be labeled completely at random), it may be preferable to allow for the possibility that they are labeled based on observed covariate features. Prediction in this case is same as prediction under covariate shift as proved in Section 3.2 and can be useful either for imputation or for understanding the spread in the response distribution. Recently, Lei and Candès (2020) and Jin et al. (2021) showed how to use prediction under covariate shift to construct prediction intervals for individual treatment effect under traditional causal inference assumptions. This can potentially be more useful in understanding the impact of a treatment at the individual level than the standard average treatment effect; please see appendix S.2 for more details.

Therefore, prediction under covariate shift is a building block of several important prediction problems and an improvement upon the weighted conformal method of Tibshirani et al. (2019) would lead to advancements in many other directions as well, which is exactly what we aim to do in this work. We reconsider the problem of prediction under covariate shift from a missing data point of view and provide a novel solution that is more robust and computationally efficient using modern semiparametric efficiency theory.

2 Conformal prediction: literature review

There are different forms of guarantees that one might consider for the validity of a prediction set. A test point (X, Y) is covered by a set C if and only if $\mathbb{1}\{Y \in C(X)\} = 1$. Often in practice, the set C is constructed based on a training data and the test point (X, Y) is independent of C. If the prediction set denoted as $\widehat{C}$ is computed from a training data and the test point (X, Y) is drawn from a distribution P, then the (prediction) miscoverage loss with respect to P is given by

$$L_P(\widehat{C}) := \mathbb{E}_{(X,Y)\sim P}[\mathbb{1}\{Y \notin \widehat{C}(X)\}|\widehat{C}] = \int \mathbb{1}\{y \notin \widehat{C}(x)\} dP(x, y).$$

Observe that $L_P(\widehat{C})$ is a random variable if $\widehat{C}$ is a random set, which often it is. Further, note that there is no requirement that the training data used to construct $\widehat{C}$ comes from the same distribution P as the test data (X, Y) in defining the loss $L_P(\widehat{C})$.

By a $(1-\alpha)$-prediction set, one might expect/ask for $L_P(\widehat{C}) \leq \alpha$. But this is too much to expect/ask in general without very restrictive assumptions on P or $\widehat{C}$. More actionable goals are to require $L_P(\widehat{C})$ to be less than α either in expectation or with some specific probability with respect to the randomness of $\widehat{C}$. We refer to the goal of $\mathbb{E}[L_P(\widehat{C})] \leq \alpha$ as **joint coverage**. This is the problem solved by conformal prediction (as introduced by Vovk et al. (2005)) and many of its variants in the literature. The goal of $\mathbb{P}(L_P(\widehat{C}) \leq \alpha) \geq 1-\delta$ is referred to as (α, δ) **probably approximately correct (PAC)** in the literature, which belongs to the class of classical tolerance regions — sets that cover a pre-specified fraction of the population distribution, see e.g. Guttman (1967) and Krishnamoorthy and Mathew (2009). Notably, the PAC prediction set depends on an additional parameter δ which is not required by the conformal approach. We note that, in general, neither will imply the other. Vovk (2012) shows that the split conformal prediction sets with $(1-\alpha)$ joint coverage also satisfy $(\alpha + f_N(\delta), \delta)$-PAC guarantee, for an explicitly computable $f_N(\delta)$.

2.1 Under Exchangeability

We will now provide a brief introduction to the literature on conformal prediction with some emphasis on split conformal prediction that will be important to understanding the current work. Vovk et al. (2005) first introduced a version of conformal prediction called the transductive conformal method (later described in Lei et al. (2013) as the full conformal method) that requires fitting the learning algorithm to samples $(X_i, Y_i), 1 \leq i \leq N$ and $X_{N+1} = x$ for all $x \in \mathcal{X}$. While this method makes full use of the data for prediction, it is computationally intensive in practice. Papadopoulos et al. (2002) proposed an alternative method called the inductive conformal method (or the split conformal method in Lei and Wasserman (2014)) which splits the data into two different parts, and the learning algorithm is trained only on the first part and the prediction set is constructed using conformal "scores" on the second part of the data. For concreteness, we describe the split conformal algorithm here in the regression setting with $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$. Let $\mathcal{A} : \mathcal{X} \to \mathbb{R}$ be any prediction algorithm trained on the first split of the data, i.e., for any $x \in \mathcal{X}$, $\mathcal{A}(x)$ is the point prediction for Y. If m is the number of observations in the second split, $R_i = |Y_i - \mathcal{A}(X_i)|, i = 1, \ldots, m$ are residuals computed on the second split of the data, and $\widehat{Q}_\alpha$ is the $\lceil (m+1)(1-\alpha) \rceil$-th largest element of R_i's, then for any (X_{N+1}, Y_{N+1}) that is exchangeable with $(X_i, Y_i), i = 1, \ldots, N$, it holds that $\mathbb{P}(|Y_{N+1} - \mathcal{A}(X_{N+1})| \leq \widehat{Q}_\alpha) \geq 1 - \alpha$. In particular, for $\widehat{C}_\alpha(x) = \{y : |y - \mathcal{A}(x)| \leq \widehat{Q}_\alpha\}$, we have $\mathbb{P}(Y_{N+1} \in \widehat{C}_\alpha(X_{N+1})) \geq 1 - \alpha$. Both full and split conformal algorithms require only the assumption of exchangeable data for coverage validity. To overcome potential statistical inefficiency due to data splitting, several papers including Barber et al. (2021), Romano et al. (2020) and Kim et al. (2020) introduced aggregation techniques such as jackknife+, CV+, bootstrap after jackknife+ to make better use of the data. All aforementioned prediction set constructions are valid under exchangeability of the training data as well as the test point (X, Y) to be covered, as discussed in Kuchibhotla (2020). Also, see Solari and Djordjilović (2022).

The above works consider joint coverage as the criterion for valid prediction set and focus on exchangeability for coverage validity. There also is a separate line of research including, but not limited to, Györfi and Walk (2020) and Yang and Kuchibhotla (2021) on prediction sets based on i.i.d. data assumption and concentration inequalities. The advantage is that one attains coverage guarantees conditional on the training data used to construct the set, which may be more informative and automatically implies unconditional (PAC) coverage. To elaborate on this line of thought and as an initial glimpse into the current work, consider the split conformal prediction procedure described above. In that description, under i.i.d. data assumption, the residuals computed on the second split of the data $R_i = |Y_i - \mathcal{A}(X_i)|, i = 1, \ldots, m$ are i.i.d. along with $R = |Y - \mathcal{A}(X)|$, irrespective of what $\mathcal{A}(\cdot)$ is. We aim to find $\widehat{r}_\alpha$ such that $\mathbb{P}(|Y - \mathcal{A}(X)| \leq \widehat{r}_\alpha | \mathcal{A}) \geq 1 - \alpha$. The concentration inequalities approach to prediction sets can now proceed as follows: the well-known DKW inequality (see e.g. Massart (1990)) implies that

$$\mathbb{E}\left[\sup_t \left| \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{|Y_i - \mathcal{A}(X_i)| \leq t\} - \mathbb{P}\big(|Y - \mathcal{A}(X)| \leq t | \mathcal{A}\big) \right| | \mathcal{A} \right] \leq \sqrt{\frac{2}{m}}.$$

Hence, an $\widehat{r}_\alpha$ at which the empirical CDF of the residuals is at least $(1-\alpha) + \sqrt{2/m}$ yields a valid prediction set of joint coverage of at least $1 - \alpha$. This also implies that if we just use the sample $(1-\alpha)-$quantile of

the residuals, then this would give a prediction set that has an approximate coverage of $1-\alpha$ with a slack at most $\sqrt{2/m}$. Moreover, the application of DKW-type inequalities here also readily yields a PAC guarantee. The well-developed theory of efficient estimation in semiparametric theory shows that the sample quantile is an optimal estimator of the population quantile (Pfanzagl, 1976). This optimality theory is leveraged throughout the manuscript in deriving well-calibrated prediction sets under covariate shift.

2.2 Under Non-exchangeability

Several authors have considered the problem of prediction sets for non-exchangeable data. The literature can be divided into three parts: (1) works that consider independent data but with potential non-identical distributions, (2) works that consider prediction set construction for dependent data including time series, and (3) works that are agnostic to the randomness structure in the data. Our current work belongs to the first category. In the following, we only review the works directly related to our work and leave the review of the latter two categories to Section S.1 of the apppendix.

Tibshirani et al. (2019) introduced the problem of prediction set construction under covariate shift. Building on Tibshirani et al. (2019), Lei and Candès (2020) worked in a covariate shift setting, establishing asymptotic joint coverage under certain conditions, accounting for estimation of the covariate density ratio. Cauchois et al. (2020) provided a method that produces approximate valid prediction set for any test distribution in an f-divergence ball around the training population and also discussed how to estimate the expected data shift and build robustness to it. Lei and Candès (2020) and Kivaranovic et al. (2020) constructed prediction sets for counterfactuals and individual treatment effects (ITE). The former work uses the classical SUTVA assumption and rewrites the problem in terms of covariate shift assumption. The latter work assumes covariates are independent of treatment assignment and therefore implicitly rules out any covariate shift. These works represent the first connection of conformal prediction to causal inference.

In this paper, we will focus on the specific form of non-exchangeability called *covariate shift* considered in Tibshirani et al. (2019) and Lei and Candès (2020) where the training data $\mathcal{D}^{\text{tr}}$ is composed of two parts $\mathcal{D}_P^{\text{tr}}$ and $\mathcal{D}_Q^{\text{tr}}$, and random variables (X_i, Y_i)'s in $\mathcal{D}_P^{\text{tr}}$ are i.i.d. from $P_X \otimes P_{Y|X}$, while random variables X_i's in $\mathcal{D}_Q^{\text{tr}}$ are i.i.d. from Q_X with missing response. The goal is to obtain a prediction set such that the probability a new data pair (X_f, Y_f) falls into this prediction set is larger than some nominal level $1-\alpha$ where (X_f, Y_f) is from $Q_X \otimes P_{Y|X}$. As noted, all prior works focused primarily on achieving this goal asymptotically with the exception of Tibshirani et al. (2019) who assumed the covariate shift is known (i.e., known dQ_X/dP_X), in which case they achieved finite sample coverage guarantee. We prove in this paper that it is impossible to construct a *non-trivial* finite sample valid prediction set without complete a priori knowledge of either the covariate shift or the conditional distribution of Y_f given X_f; see Theorem 1. We thus resort to the goal of constructing a prediction set with an asymptotic coverage of at least $1-\alpha$.

Our approach to construction of prediction sets under covariate shift requires estimation of the quantile of a univariate function of (X_f, Y_f) akin to the conformal score. This task involves nuisance parameters that must be estimated sufficiently well to ensure the asymptotic coverage guarantee of our prediction sets. Fortunately, as we establish using modern semiparametric theory, this task can be accomplished in a robust fashion by using the efficient influence function as an estimating equation for the quantile. As we show, the efficient influence function is endowed with a double robustness property (see e.g. Scharfstein et al. (1999), Robins (2000) and Bang and Robins (2005)) which ensures that the coverage bias of our prediction sets can be made negligible even if the nuisance parameters are estimated at nonparametric rates using highly adaptive machine learning algorithms.

Before completing the review of the relevant literature, we mention a concurrent work Qiu et al. (2022) that uses similar connection to semiparametric statistics as ours does but targets a different type of asymptotic PAC guarantee than we are able to provide in Theorem 4, a detailed discussion comparing the two goals and corresponding methods is given at the end of Section 3.1.

Organization The remainder of the paper is organized as follows. In Section 3, we formally introduce the explainable covariate shift problem, providing the missing data formulations of the problem that gives

notation used throughout the paper and also makes its connection to causal inference. In Section 4, we formally establish that it is impossible to construct a well-calibrated prediction set that is informative without a priori knowledge about the covariate shift such as knowledge of the ratio of covariate densities, in the sense that any valid prediction set would with high probability have infinite Lebesgue measure. In Section 5, we provide our first doubly robust algorithm for the explainable covariate shift problem, using a sample splitting strategy which attains nominal asymptotic prediction coverage. In Section 6, we provide a second doubly robust algorithm that makes more efficient use of the observed data by avoiding sample splitting, with guaranteed validity under certain regularity conditions we establish. Section 7 reports simulation studies (both synthetic and real data) validating the theoretical results of the proposed doubly robust methods, and comparing them to the weighted conformal prediction method of Tibshirani et al. (2019). In Section 8, we consider the problem of aggregating a collection of prediction sets by providing an explicit algorithm adapted from Yang and Kuchibhotla (2021) in an effort to optimize prediction accuracy; we illustrate the efficiency of the proposed algorithm through simulations. In Section 9, we relax the explainable covariate shift assumption by allowing for the presence of latent covariate shift encoded in a sensitivity parameter and discuss the efficient influence function. This is the first step in extending the framework of the current paper to account for a departure from explainable covariate shift as well as constructing prediction sets for ITE under unmeasured confounding. Further elaboration of Section 9 will appear elsewhere. Finally in Section 10, we conclude the paper with a brief discussion.

Proofs of all results and supporting lemmas are provided in the supplementary where for convenience, the sections and equations are prefixed with "S." and "E.", respectively.

3 Our Problem and Notation

In this section, we provide a formal description of the *explainable* covariate shift problem, and introduce both missing data and counterfactual formulations of the problem.

3.1 The Covariate Shift Problem

The common assumption that the training and test data follow a common probability distribution can fail in practice, for example training data may be collected under stringent laboratory conditions that cannot be met when deployed in clinical practice; likewise, image training data may be obtained in one region whereas the test data may be collected in another. Therefore, it is important to consider situations where training and test distributions are different, also known as *covariate shift*, which we formalize below. Assuming we have training data $\mathcal{D}^{\text{tr}}$ composed of two parts $\mathcal{D}_P^{\text{tr}}$ and $\mathcal{D}_Q^{\text{tr}}$, where

$$\mathcal{D}_P^{\text{tr}} \;:=\; \{Z_i = (X_i, Y_i) : 1 \le i \le n\} \quad \text{and} \quad \mathcal{D}_Q^{\text{tr}} \;:=\; \{Z_i = X_i : n+1 \le i \le N\}, \tag{2}$$

and random variables in $\mathcal{D}_P^{\text{tr}}$ are i.i.d. from $P_X \otimes P_{Y|X}$, while random variables in $\mathcal{D}_Q^{\text{tr}}$ are i.i.d. from Q_X. Note that in $\mathcal{D}_Q^{\text{tr}}$ only data on covariates are available and the outcome/response is missing, and thus the subject of prediction. In such setting, a covariate shift problem is said to be present as the covariates in $\mathcal{D}_Q^{\text{tr}}$ are sampled from Q_X which may be different from P_X (the distribution of covariates in $\mathcal{D}_P^{\text{tr}}$). This setting readily extends to the more general case where one also observes samples from $Q_X \otimes P_{Y|X}$, and/or samples from $Q_X \otimes Q_{Y|X}$; settings that are closely related to *transfer learning* (see e.g. Kpotufe and Martinet (2018) and Reeve et al. (2021)) and in the special case when P_X and Q_X are identical, this reduces exactly to the setting of *semi-supervised learning* (see e.g. Zhu and Goldberg (2009) and Zhang et al. (2019)). Though several variants of the covariate shift problem have been of interest in statistics and ML literatures, the specific setup considered has recently generated renewed interest in ML literature. While the focus has usually been on approaches to account for covariate shift while conducting model selection such as regression, see e.g. Sugiyama et al. (2007), Quiñonero-Candela et al. (2008), Bickel et al. (2009), Reddi et al. (2015),

Chen et al. (2016), the goal here is to obtain a prediction set such that

$$\mathbb{P}\big(Y_f \in \widehat{C}_{N,\alpha}(X_f)\big) \geq 1-\alpha, \quad \text{whenever} \quad (X_f, Y_f) \sim Q_X \otimes P_{Y|X}. \tag{3}$$

This problem was first posed in Tibshirani et al. (2019) where they developed a weighted version of conformal prediction that produces valid prediction sets when the likelihood ratio between training and test distributions is known. The idea of their construction is similar to importance sampling Monte Carlo. If, as in most practical settings, the likelihood ratio between the two distributions is unknown and therefore must be estimated, they empirically demonstrate in a low-dimensional setting that approximate coverage might still be possible via simulation studies. However, Tibshirani et al. (2019) do not formally consider the extent to which bias in estimating the likelihood ratio propagates to impact coverage. In addition, as noted by the authors of that paper, for every new test point (x_f, y_f), their prediction set that involves a weighted quantile has to be recalculated, and this would be computationally intensive. Specifically, their approach requires that the test point x_f is specified in advance.

In this work we will construct prediction regions $\widehat{C}_N$ that are determined by one-dimensional functions of (X_f, Y_f), i.e., we take an arbitrary function $(x, y) \mapsto R(x, y) \in \mathbb{R}$ and estimate the quantile of $R(X_f, Y_f)$. For now, we will think of $R(\cdot, \cdot)$ as a fixed non-stochastic function. In practice, $R(\cdot, \cdot)$ is a conformal score computed from an independent sample; examples include $R(x, y) = |y - \mathcal{A}(x)|$ (regression residual, Lei et al., 2018) or $R(x, y) = \max\{\widehat{q}_{\alpha/2}(x) - y, y - \widehat{q}_{1-\alpha/2}(x)\}$ (conformalized quantile residual with estimated conditional quantiles $\widehat{q}_{\alpha/2}(\cdot)$ and $\widehat{q}_{1-\alpha/2}(\cdot)$, Romano et al., 2019). If r_α is the smallest $(1-\alpha)$-quantile of $R(X_f, Y_f)$ in the target population $Q_X \otimes P_{Y|X}$, then

$$\mathbb{P}\big(R(X_f, Y_f) \leq r_\alpha\big) \geq 1-\alpha,$$

and hence, for $C_\alpha = \{(x, y) : R(x, y) \leq r_\alpha\}$, we have $\mathbb{P}\big((X_f, Y_f) \in C_\alpha\big) \geq 1-\alpha$. This result holds, irrespective of the choice of function $R(\cdot, \cdot)$. Note that the assumption that Y_i conditional on X_i for $1 \leq i \leq n$ has the same distribution as Y_f conditional on X_f essentially means that the covariate shift problem can be completely accounted for by conditioning on observed covariates X, hence the reference to this setting as "explainable covariate shift" problem. It implies that the conditional distribution of $R(X_i, Y_i)$ given X_i coincides with the conditional distribution of $R(X_f, Y_f)$ given X_f, i.e.,

$$\mathbb{P}\big(R(X_i, Y_i) \in B \big| X_i = x\big) = \mathbb{P}\big(R(X_f, Y_f) \in B \big| X_f = x\big), 1 \leq i \leq n \tag{4}$$

for all Borel sets $B \subseteq \mathbb{R}$. We borrow results from semiparametric theory and estimate r_α based on its efficient influence function, which is intimately related to efficient influence function of the average treatment effect among the treated (ATT) functional; we then combine this influence function with arbitrary training map $R(\cdot, \cdot)$ and establish that the resulting prediction set has asymptotic nominal coverage with coverage bias of a product form which implies correct coverage if either the likelihood ratio between the two distributions or the conditional distribution of R given X can be estimated sufficiently well, also known as double robustness. Note that in our construction of the prediction set $\widehat{C}_{N,\alpha}$, the mapping R does not depend on the test point x_f at which prediction is needed. Tibshirani et al. (2019) approached (3) using the equation

$$\mathbb{P}_{(X_f,Y_f)\sim Q_X\otimes P_{Y|X}}(R(X_f, Y_f) \leq \theta) = \mathbb{E}_{(X,Y)\sim P_X\otimes P_{Y|X}}\left[\mathbb{1}\{R(X, Y) \leq \theta\}\frac{dQ_X}{dP_X}(X)\right]. \tag{5}$$

This implies that consistent estimation of dQ_X/dP_X allows for consistent estimation of r_α. But, given (4), consistent estimation of the conditional distribution of $R(X, Y)$ given X also allows for consistent estimation of r_α. This gives a hint at double robustness in estimating r_α which we will formalize later in the paper.

In Lei and Candès (2020), the authors proposed a method that targets the covariate shift problem under the framework of counterfactual prediction in a causal inference setting, which we consider in Section S.2. Interestingly, the coverage of their prediction set has bias of the order of the minimum of two errors, that of the prediction ML algorithm and that of the estimated covariate likelihood ratio, a property that appears to hold under the so-called conformal quantile regression (CQR) which restricts the choice of conformal score

to a quantile regression function for the outcome in view; it is unclear whether similar robustness extends beyond CQR. In contrast, while our approach equally applies to the counterfactual prediction framework considered by Lei and Candès (2020), as we establish, its coverage bias is guaranteed to be of the order of the product of two errors, that of an estimated quantile function for R with that of the covariate likelihood ratio, an immediate consequence of double robustness. Therefore, the bias of our coverage error rate can be substantially smaller relative to that of Lei and Candès (2020). In addition, the product bias property of the proposed method is guaranteed to hold for any ML technique used to empirically construct R, therefore making our approach potentially more general than theirs. Park et al. (2021) uses probably approximately correct (PAC) prediction sets (tolerance regions that cover a pre-specified fraction of the population distribution) for deep learning models including in the presence of covariate shift that also requires prior knowledge on the shift of the distributions. Notably, the PAC prediction set depends on an additional parameter δ which is not required by our approach.

Concurrently to our paper, Qiu et al. (2022) consider a related prediction setting, also drawing from modern semiparametric theory, but mainly focusing on a particular form of asymptotic PAC guarantee with covariate shift. We view these contributions as complementary. Specifically, in our covariate shift setting, asymptotic joint coverage is interpreted as: $\mathbb{P}(Y_f \in \widehat{C}_{N,\alpha}(X_f)) \geq 1-\alpha-o(1)$ as $N \to \infty$. In contrast, the (α, δ)-PAC guarantee can be met approximately with negligible errors in either α, δ or both. The target of Qiu et al. (2022) is to develop $(\alpha, \delta+o(1))$-PAC prediction set. The methodology developed in the current paper provides $(\alpha + o_p(1), \delta)$-PAC guarantee along with the joint coverage guarantee. Whichever version of asymptotic PAC guarantee is more useful depends on the application. Notably, in order to minimize the impact of bias due to nuisance parameter estimation, Qiu et al. (2022) leverage the product bias property of the efficient influence function for the coverage probability of their prediction set, while we leverage the product bias structure of the efficient influence function for the $(1-\alpha)$-quantile of the statistic generating the prediction set; these two influence functions are equal up to a multiplicative constant for a given coverage guarantee, although the difference in their use leads to a different guarantee. Importantly, their PAC approach involves the construction of a valid confidence interval for the coverage probability which in turn requires a regular asymptotic linear estimator of coverage, in which case, their product bias must be of order smaller than root-n. Our proposed approach does not have this requirement, and therefore our guarantee is attainable even if the product bias is of order larger than root-n provided that at least one of our nuisance functions is consistent.

3.2 Reformulation as a Missing Data Problem

In this section, we formulate the covariate shift problem in a missing data framework, and introduce notation used throughout the remainder of the paper. This reformulation allows us to make use of the modern theory of semiparametric statistics. Recall the training data (2). For each (X_i, Y_i) contained in $\mathcal{D}_P^{\text{tr}}$, define $R_i = R(X_i, Y_i)$ and set $T_i = 0$. For each (X_i, Y_i) contained in $\mathcal{D}_Q^{\text{tr}}$, R_i is unobserved because the corresponding Y_i is unobserved. Hence the observed data $\mathcal{D} = \mathcal{D}_P^{\text{tr}} \cup \mathcal{D}_Q^{\text{tr}}$ can be succinctly written as $Z_i = \big(X_i, T_i, (1-T_i)R_i\big), 1 \leq i \leq N$ such that

$$\mathbb{P}(X_i \in A | T_i = 0) =: P_X(A) \quad \text{and} \quad \mathbb{P}(X_i \in A | T_i = 1) =: Q_X(A), \tag{6}$$

while

$$\mathbb{P}(R_i \in B | T_i = 0, X_i = x) = \mathbb{P}(R_i \in B | T_i = 1, X_i = x) =: P_{Y|X=x}(B) \quad \text{a.e.} \quad x. \tag{7}$$

Equations in (6) signify that the covariates in $\mathcal{D}_P^{\text{tr}}$ are distributed as P_X and that the covariates in $\mathcal{D}_Q^{\text{tr}}$ are distributed as Q_X. Condition (7), on the other hand, signifies that the conditional distribution of the response Y given X is the same for $\mathcal{D}_P^{\text{tr}}$, $\mathcal{D}_Q^{\text{tr}}$, and the future data (X_f, Y_f). Condition (7) restates condition (4) in terms of T_i.

Condition (7) implies that R_i is independent of T_i conditional on X_i for all $1 \leq i \leq N$; this is denoted by $R_i \perp T_i | X_i$ and is equivalent to the missing at random (MAR) assumption in missing data literature. This assumption, which is not testable without an additional condition, essentially states that there is no

unmeasured factor that is related with both R and T. For identification, we further assume that for any Borel set B,

$$P_X(B) = 0 \text{ implies } Q_X(B) = 0. \tag{8}$$

This is the same as assuming the measure Q_X is absolutely continuous w.r.t P_X. In other words, the support of $X|T = 1$ is contained in the support of $X|T = 0$. Assumption (8) is needed for (5), which is a crucial component of the double robustness of our methodology.

Summarizing the above discussion, the covariate shift assumption which states that only the covariate distributions between $\mathcal{D}_P^{\text{tr}}$ and $\mathcal{D}_Q^{\text{tr}}$ can be different but not the conditional distributions of the response given the covariates is equivalent to the MAR assumption.

For any $\theta \in \mathbb{R}$, define sets

$$\widehat{C}(\theta; x) := \big\{y : R(x, y) \leq \theta\big\}.$$

In this notation, under condition (8), we aim to find a data-dependent random variable $\widehat{r}_\alpha$ such that for any $(X_f, Y_f) \sim Q_X \otimes P_{Y|X}$,

$$\mathbb{P}\big(Y_f \in \widehat{C}(\widehat{r}_\alpha; X_f)\big) = \mathbb{P}_{(X,Y)\sim Q_X \times P_{Y|X}}(R(X, Y) \leq \widehat{r}_\alpha) = \mathbb{P}(R(X, Y) \leq \widehat{r}_\alpha | T = 1) \geq 1 - \alpha, \tag{9}$$

while r_α, the "target" of $\widehat{r}_\alpha$ is defined to be the smallest real number such that

$$\mathbb{P}\big(R(X, Y) \leq r_\alpha | T = 1\big) \geq 1 - \alpha.$$

Because we do not observe random variables R when $T = 1$, this goal is not achievable in finite samples without restrictive assumptions such as a known $\mathbb{P}(T = 1|X)/\mathbb{P}(T = 0|X)$ (Tibshirani et al., 2019). We provide a random variable $\widehat{r}_\alpha$ so that (9) is achieved with a slack that converges to zero as N tends to ∞.

4 Impossibility of finite sample coverage

Recall our aim from (9). Resorting to semiparametric theory, in most cases, implies that the resulting coverage guarantee is only asymptotic. In our problem of covariate shift, one can prove that it is impossible to construct a finite sample valid *non-trivial* prediction set without the knowledge of either the covariate shift or the conditional distribution of Y given X. Here, by a non-trivial prediction set, we mean a set with a finite Lebesgue measure. Lemma S1 of Section S6.1 of Qiu et al. (2022) prove an analogous but weaker result for PAC guarantee as they establish a result similar to (11) of the following Theorem, however (12) appears to be an entirely novel contribution.

Theorem 1. *Suppose the observed data consists of n i.i.d. tuples $(X_i, T_i, (1 - T_i)Y_i)$. Further assume that $\mathcal{X} \subseteq \mathbb{R}^d$ is the support of X_i and $\mathcal{Y} \subseteq \mathbb{R}$ is the support of Y_i. Let $\bar{\mathcal{P}}^0$ be the set of all distributions $\bar{P}^0$ on the random vector $\bar{O} = (X, T, Y)$ such that T is independent of Y given X ($T \perp Y|X$), and the joint distribution of (X, Y) is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X} \times \mathcal{Y}$.*

Suppose that a (possibly randomized) prediction set $\widehat{C}_\alpha$ has finite-sample joint coverage guarantee in the target population, that is,

$$\sup_{\bar{P}^0 \in \bar{\mathcal{P}}^0} \mathbb{P}_{\bar{P}^0}(Y \notin \widehat{C}_\alpha(X) \mid T = 1) \leq \alpha, \quad \textit{for some} \quad \alpha \in (0, 1). \tag{10}$$

Then, for any $\bar{P}^0 \in \bar{\mathcal{P}}^0$ and a.e. $y \in \mathcal{Y}$ with respect to the Lebesgue measure,

$$\mathbb{P}_{\bar{P}^0}\big(y \notin \widehat{C}_\alpha(X)\big) \leq \alpha. \tag{11}$$

Furthermore, $\widehat{C}_\alpha(X)$ would at least cover one of the end points $\mathcal{Y}$ with probability at least $1 - \alpha$, and hence if $\mathcal{Y} = \mathbb{R}$, then

$$\mathbb{E}_{\bar{P}^0}[Leb(\widehat{C}_\alpha(X))] = \infty. \tag{12}$$

We defer the proof to the Section S.8 of the appendix. The proof is based on the lack of a non-trivial test for the problem of conditional independence hypothesis testing proved in Shah and Peters (2020). The connection to conditional independence testing can be seen from the fact that any prediction set that is valid under the conditional independence of T and Y given X can be rewritten as a valid test for the hypothesis that T is conditionally independent of Y given X; see the proof in Section S.8 for more details.

Given the lack of finite-sample valid non-trivial prediction, we resort to finding an efficient asymptotically valid prediction set based on semi-parametric theory in the following sections.

5 Methodology with split data

In this section, we discuss our novel prediction set construction under the covariate shift setting using semiparametric theory.

Recall our notation $Z = (X, T, (1-T)R)$. Suppose that one is interested in the qth-quantile of a random variable $R|T=1$, denoted $\theta_0 := \inf\{r : F(r) \geq q\}$, where F is the CDF of $R|T=1$. An estimator $\widehat{\theta}$ is said to be asymptotically linear if it satisfies

$$\sqrt{N}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(Z_i) + o_p(1), \quad \mathbb{E}[\psi(Z)] = 0, \mathbb{E}\big[\psi(Z)^\top \psi(Z)\big] < \infty,$$

as $N \to \infty$. Clearly, the asymptotic variance of $\widehat{\theta}$ is then $\mathbb{E}\big[\psi(Z)\psi(Z)^\top\big]/N$. The function $\psi(z)$ is referred to as the influence function, following terminology of Hampel (1974). Furthermore, as the model is non-parametric in the sense that the observed data distribution is not restricted and all such distributions are regular, (see Chapter 2 of Bickel et al. (1993) on regularity), any estimator satisfying the above expansion is said to attain the semiparametric efficiency bound, and $\psi(\cdot)$ is said to be the efficient influence function of θ_0 in the nonparametric model. For more on influence functions and semiparametric theory, see e.g. Newey (1990), Chapter 25 of Van der Vaart (2000) and van der Vaart (2002).

We now state the efficient influence function for the $(1-\alpha)$-th quantile of $R|T=1$ under the MAR assumption. For every $x \in \mathcal{X}$ and $r \in \mathbb{R}$, define

$$\begin{aligned} \pi^\star(x) &:= \mathbb{P}(T=1|X=x)/\mathbb{P}(T=0|X=x), \\ m^\star(r,x) &:= \mathbb{E}[\mathbb{1}\{R \leq r\}|X=x]. \end{aligned}$$

The function $\pi^\star(\cdot)$ represents the true density ratio of the covariates among labeled and unlabeled data. The function $m^\star(\cdot,\cdot)$ represents the true conditional mean function that by assumption (4) is common for both labeled and unlabeled data.

We now state the efficient influence function for the quantile of interest under regularity conditions for the data distribution that will motivate our proposed method, where it should be noted that strictly speaking, the regularity conditions are actually not needed for the theoretical guarantees of our proposed methods.

Lemma 1. *Suppose* $\mathbb{E}[\mathbb{1}\{T=0\}\pi^{\star 2}(X)] = \mathbb{E}\mathbb{P}^2(T=1|X)/\mathbb{P}(T=0|X)$ *is finite and that the density of the conditional distribution of* $R|T=1$ *at* r_α *is bounded away from zero. Then the efficient influence function of the* $(1-\alpha)$*-quantile of* $R|T=1$ *in the nonparametric model for* Z *which allows the distribution of* Z *to remain unrestricted under the condition that it is regular, is given up to a proportionality constant by*

$$\begin{aligned} \psi(z) = \mathrm{IF}(r_\alpha, x, r, t; \pi^\star, m^\star) = \mathbb{1}\{t=0\}\pi^\star(x)\Big[\mathbb{1}\{r \leq r_\alpha\} - m^\star(r_\alpha, x)\Big] \\ + \mathbb{1}\{t=1\}\Big[m^\star(r_\alpha, x) - (1-\alpha)\Big]. \end{aligned} \tag{13}$$

Proof. Hahn (1998) gave a derivation of the influence function for the average treatment effect among the treated (ATT). We adapt their proof to that of the conditional quantile and give the complete derivation of (13) along with a basic introduction to semiparametric theory needed to derive the result in Section S.3 of the supplementary. □

For any two functions $\pi(\cdot)$ and $m(\cdot,\cdot)$, let

$$\mathrm{IF}(\theta, x, r, t; \pi, m) := \mathbb{1}\{t=0\}\pi(x)\Big[\mathbb{1}\{r \le \theta\} - m(\theta, x)\Big] + \mathbb{1}\{t=1\}\Big[m(\theta, x) - (1-\alpha)\Big],$$

ignoring the scaling factor. Note that $\mathrm{IF}(\theta, x, r, t; \pi, m)$ is only a function of $(x, t, (1-t)r)$ because the term that depends on r has a multiplicative factor of $\mathbb{1}\{t=0\}$. Let $P[f]$ denote integration conditional on the training sample. For example, for any θ, π, m that are potentially data-dependent,

$$P[\mathrm{IF}(\theta, x, r, t; \pi, m)] = \int \mathrm{IF}(\theta, x, r, t; \pi, m) dP_{R|X=x}(r|x) dP_{T|X=x}(t|x) dP_X(x).$$

This is a random variable if θ or π or m are random.

First, we draw a key connection between the desired coverage and the aforementioned influence function.

Lemma 2. *Let $\pi : \mathcal{X} \to \mathbb{R}_+$ and $m : \mathbb{R} \times \mathcal{X} \to [0,1]$ be any two functions. Then for every (potentially) data-dependent $\theta \in \mathbb{R}$, the representation*

$$\begin{aligned} \mathbb{P}_{(X,Y)\sim Q_X \otimes P_{Y|X}}\big(Y \in \widehat{C}(\theta; X) \mid \theta\big) &= \mathbb{P}(Y \in \widehat{C}(\theta; X) | \theta, T=1) \\ &= 1 - \alpha + \frac{P[\mathrm{IF}(\theta, X, R, T; \pi, m)]}{\mathbb{P}(T=1)}, \end{aligned} \tag{14}$$

holds true, whenever either of the following holds true:

1. *$\pi(x) = \pi^\star(x)$ for all x; or*
2. *$m(\gamma, x) = m^\star(\gamma, x)$ for all γ and x.*

Proof. See Section S.4 for a proof. □

Note that even if data-dependent, an estimate of θ remains independent of a future observation (X, Y). Lemma 2 has a key implication that $\mathrm{IF}(\cdots)$ is a doubly robust influence function. Because

$$\mathbb{P}\big(Y \in \widehat{C}(\theta; X) \mid \theta, T=1\big) = \mathbb{P}\big(R(X,Y) \le \theta \mid \theta, T=1\big),$$

taking $\theta = r_\alpha$, the quantile of $R(X,Y)$ conditional on $T=1$,[1] Lemma 2 implies that

$$P[\mathrm{IF}(r_\alpha, X, R, T; \pi, m)] = 0, \quad \text{if either} \quad \pi \equiv \pi^\star \text{ or } m \equiv m^\star. \tag{15}$$

Because r_α is a constant, $P[\mathrm{IF}(r_\alpha, X, R, T; \pi, m)] = \mathbb{E}[\mathrm{IF}(r_\alpha, X, R, T; \pi, m)]$.

Remark. *Note that our results do not actually require uniqueness of a solution to $\mathbb{E}\big[\mathrm{IF}(\theta, x, r, t; \pi^\star, m^\star)\big] = 0$. This could arise for example in settings when R is discrete. In principle, our result would continue to hold for any element θ of a solution set.*

Property (15) is a double robustness property in that the expectation is zero, as long as one of π and m is the true function. This property implies that when, as would generally be the case in practice, $\pi^\star$ and $m^\star$ are estimated, the resulting bias is of the following product form, where we introduce the notation $\|f(\cdot)\|_2$ as the L_2-norm of $f(\cdot)$, where $\|f(\cdot)\|_2 := [\int f^2(x) M(\mathrm{d}x)]^{1/2}$, and $M(\cdot)$ is the probability measure of X that for any Borel set B, $M(B)$ is given by

$$\begin{aligned} M(B) &= \mathbb{P}(X \in B) \\ &= \mathbb{P}(X \in B | T=1)\mathbb{P}(T=1) + \mathbb{P}(X \in B | T=0)\mathbb{P}(T=0) \\ &= Q_X(B)\mathbb{P}(T=1) + P_X(B)\mathbb{P}(T=0). \end{aligned}$$

[1] We assume here that $\mathbb{P}_{(X,Y)\sim Q_X \otimes P_{Y|X}}(R(X,Y) \le r_\alpha) = 1-\alpha$, which is mild as one can add small Gaussian noise to $R(X,Y)$.

Theorem 2. *For any functions $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot,\cdot)$, it holds that*

$$\sup_{\gamma\in\mathbb{R}}\Big|P\big[\mathrm{IF}(\gamma,X,R,T;\widehat{\pi},\widehat{m})-\mathrm{IF}(\gamma,X,R,T;\pi^\star,m^\star)\big]\Big| \\ \leq \|\widehat{\pi}-\pi^\star\|_2 \sup_{\gamma}\|\widehat{m}(\gamma,\cdot)-m^\star(\gamma,\cdot)\|_2.$$

Proof. See Section S.5 for a proof. □

Theorem 2 implies that $P[\mathrm{IF}(\gamma,X,R,T;\widehat{\pi},\widehat{m})]$ converges to $P[\mathrm{IF}(\gamma,X,R,T;\pi^\star,m^\star)]$ as long as one of $\pi^\star$ and $m^\star$ is estimated consistently.

Lemma 2 is the main building block for our methodology. In order to ensure approximately correct coverage of $1-\alpha$, we need to find θ such that $P[\mathrm{IF}(\theta,X,R,T;\pi,m)]$ is approximately zero, with either $\pi\equiv\pi^\star$ or $m\equiv m^\star$. In practice where we do not have access to either $\pi^\star$ or $m^\star$ and even if we know either of them, one cannot compute $P[\mathrm{IF}(\theta,X,R,T;\pi,m)]$ without access to the true distribution of $(X,(1-T)R,T)$. Our methodology, hence, is as follows. We construct estimators $\widehat{\pi}(\cdot)$ and $\widehat{m}(\cdot,\cdot)$ such that $\|\widehat{\pi}-\pi\|_2=o_p(1)$ and $\|\widehat{m}-m\|_2=o_p(1)$ for some $\pi(\cdot)$ and $m(\cdot,\cdot)$, and either $\pi\equiv\pi^\star$ or $m\equiv m^\star$. Then we find the smallest $\widehat{\theta}$ such that

$$\mathbb{P}_N\big[\mathrm{IF}(\widehat{\theta},X,R,T;\widehat{\pi},\widehat{m})\big]:=\frac{1}{N}\sum_{1\leq i\leq N}\mathrm{IF}(\widehat{\theta},X_i,R_i,T_i;\widehat{\pi},\widehat{m})\geq 0. \tag{16}$$

We can prove under certain regularity conditions on $\widehat{\pi}$ and $\widehat{m}$ that

$$\mathbb{P}_N\big[\mathrm{IF}(\widehat{\theta},X,R,T;\widehat{\pi},\widehat{m})\big]-P\big[\mathrm{IF}(\widehat{\theta},X,R,T;\pi,m)\big]=o_p(1), \tag{17}$$

even for a data-dependent $\widehat{\theta}$. Then, Lemma 2 implies that $\widehat{\theta}$ satisfying (16) also satisfies

$$\mathbb{P}_{(X,Y)\sim Q_X\otimes P_{Y|X}}\big(Y\in\widehat{C}(\widehat{\theta};X)\mid\widehat{\theta}\big)\geq(1-\alpha)+o_p(1).$$

This yields the desired coverage guarantee (9). In finding $\widehat{\theta}$ and proving (17), one can avoid restrictive regularity conditions (such as smoothness or Donsker class) on $\widehat{\pi},\widehat{m}$ by splitting the data into two parts, using the first part to determine $\widehat{\pi},\widehat{m}$ and using the second part to compute $\mathbb{P}_{\mathcal{I}_2}[\mathrm{IF}(\cdots)]$. The detailed split sample procedure is succinctly described in Algorithm 1.

In step 2 of Algorithm 1, the training method $\mathcal{A}$ can be a regression estimator of Y on X leading to $\widehat{\mu}$ and the nested sets, for example, could be $\mathcal{F}_t=\{(x,y):|y-\widehat{\mu}(x)|\leq t, t\geq 0\}$. This corresponds to using $R=R(x,y)=|y-\widehat{\mu}(x)|$. Alternatively, one can also consider a training method that leads to conditional quantile estimator $\widehat{q}_{\alpha/2}(\cdot)$ and $\widehat{q}_{1-\alpha/2}(\cdot)$. The nested sets, for example, could be $\mathcal{F}_t=\{(x,y):y\in[\widehat{q}_{\alpha/2}(x)-t,\widehat{q}_{1-\alpha/2}(x)+t]\}$. This corresponds to the map $R=R(x,y)=\max\{\widehat{q}_{\alpha/2}(x)-y,y-\widehat{q}_{1-\alpha/2}(x)\}$ which is the conformal score of the conformalized quantile regression (CQR) method of Romano et al. (2019). For discrete/categorial response Y, conditional probability of each class $\mathbb{P}(Y=j|X=x)$ could be used to construct the nested sets and the map R, see for example, Section 4 of Kuchibhotla and Berk (2021).

Now we list some assumptions that will be used in the following theorems.

(A1) $(X_i,T_i,(1-T_i)R_i), i\in\mathcal{I}_2$ are independent and identically distributed random vectors satisfying condition (7).

(A2) The functions $(\theta,x)\mapsto\widehat{m}(\theta,x)$ and $x\mapsto\widehat{\pi}(x)$ are bounded, i.e., there exist m_0 and π_0 such that for all $\theta\in\mathbb{R}$ and $x\in\mathbb{R}^d$, $|\widehat{m}(\theta,x)|\leq m_0$ and $|\widehat{\pi}(x)|\leq\pi_0$.

(A3) The estimator $\widehat{m}(\theta,x)$ is a nondecreasing function of θ.

Assumptions **(A1)** and **(A2)** are both standard conditions where we note that $m^\star(\cdot,\cdot)$ is a conditional CDF contained in the unit interval [0,1]. For assumption **(A3)**, because $m^\star(\theta,x)$ is a conditional CDF which must be monotonically nondecreasing in θ, any given estimator $\widetilde{m}$ can be improved upon by transforming it

Algorithm 1: Split doubly robust prediction

Input: Training data $\mathcal{D}^{\text{tr}} = \mathcal{D}_P^{\text{tr}} \cup \mathcal{D}_Q^{\text{tr}}$; Coverage probability $1-\alpha$, a training method $\mathcal{A}$ and estimators $\widehat{\pi}, \widehat{m}$, the point for prediction x.

Output: A valid prediction set $\widehat{C}_\alpha(x)$.

1 Split training data $\mathcal{D}^{\text{tr}}$ randomly into $\mathcal{D}_1$ and $\mathcal{D}_2$, where $\mathcal{D}_1 = \{Z_i \in \mathcal{D}^{\text{tr}}, i \in \mathcal{I}_1\}$ and $\mathcal{D}_2 = \{Z_i \in \mathcal{D}^{\text{tr}}, i \in \mathcal{I}_2\}$.

2 Fit the training method $\mathcal{A}$ on $\mathcal{D}_1$ and using fitted method $\mathcal{A}$, construct an increasing (nested) sequence of sets $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$. Here $\mathcal{T}$ is a subset of $\mathbb{R}$. The nested sets $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ can depend arbitrarily on $\mathcal{D}_1$.

3 For each $i \in \mathcal{I}_2$ that satisfies $T_i = 0$, define the conformal score

$$r_i \;=\; r(Z_i) := \inf\{t \in \mathcal{T} :\, Z_i \in \mathcal{F}_t\}.$$

4 Fit estimators $\widehat{\pi}, \widehat{m}$ on $\mathcal{D}_1$ and find the smallest $\widehat{\theta} = \widehat{r}_\alpha$ such that $\mathbb{P}_{\mathcal{I}_2}[\text{IF}(\widehat{\theta}, X, R, T; \widehat{\pi}, \widehat{m})] \geq 0$, where

$$\begin{aligned}\mathbb{P}_{\mathcal{I}_2}[\text{IF}(\widehat{\theta}, X, R, T; \widehat{\pi}, \widehat{m})] \;=\; & \frac{1}{|\mathcal{I}_2|}\sum_{i\in\mathcal{I}_2} \mathbb{1}\{t_i = 0\}\widehat{\pi}(x_i)\Big[\mathbb{1}\{r_i \leq \widehat{\theta}\} - \widehat{m}(\widehat{\theta}, x_i)\Big] \\ & + \frac{1}{|\mathcal{I}_2|}\sum_{i\in\mathcal{I}_2} \mathbb{1}\{t_i = 1\}\Big[\widehat{m}(\widehat{\theta}, x_i) - (1-\alpha)\Big].\end{aligned}$$

5 **return** the prediction set $\widehat{C}_\alpha(x) := \big\{y : R(x,y) \leq \widehat{r}_\alpha\big\}$.

into a monotone estimator $\widehat{m}^*$ such that $\|\widehat{m}^* - m\| \leq \|\widetilde{m} - m\|$, see e.g. the first two properties of Proposition 2 of Chernozhukov et al. (2009). We state their result in Proposition 1 in S.9 of the supplementary for completeness. Given this proposition, it is natural that we restrict the estimator $\widehat{m}(\cdot,\cdot)$ to the class of functions that are non-decreasing in their first argument. Hence we impose assumption **(A3)**.

Under assumptions **(A1)**–**(A3)**, we now provide a coverage guarantee for the prediction set $\widehat{C}_\alpha$ returned by Algorithm 1. Observe that following Lemma 2 and the discussion surrounding (16) and (17), we obtain

$$\begin{aligned}\mathbb{P}\Big(Y \in \widehat{C}_\alpha(X) \mid \mathcal{D}^{\text{tr}}, T = 1\Big) - (1-\alpha) \;=\; & \frac{\mathbb{P}_{\mathcal{I}_2}[\text{IF}(\widehat{r}_\alpha, X, R, T; \widehat{\pi}, \widehat{m})]}{\mathbb{P}(T=1)} \\ & + \frac{P[\text{IF}(\widehat{r}_\alpha, X, R, T; \widehat{\pi}, \widehat{m})] - \mathbb{P}_{\mathcal{I}_2}[\text{IF}(\widehat{r}_\alpha, X, R, T; \widehat{\pi}, \widehat{m})]}{\mathbb{P}(T=1)} \\ & + \frac{P[\text{IF}(\widehat{r}_\alpha, X, R, T; \pi^\star, m^\star)] - P[\text{IF}(\widehat{r}_\alpha, X, R, T; \widehat{\pi}, \widehat{m})]}{\mathbb{P}(T=1)} \\ \geq\; & 0 + \mathbf{I} + \mathbf{II}.\end{aligned} \tag{18}$$

Here we use the fact that $\widehat{r}_\alpha$ satisfies $\mathbb{P}_{\mathcal{I}_2}[\text{IF}(\widehat{r}_\alpha, X, R, T; \widehat{\pi}, \widehat{m})] \geq 0$. The term **II** in (18) can be bounded in absolute value using Theorem 2. We now provide a bound on **I** in Theorem 3 below under assumptions **(A1)**–**(A3)**. Theorem 3 actually proves the tail and expectation bound for

$$\sup_{\theta\in\mathbb{R}} |\mathbb{P}_{\mathcal{I}_2}[\text{IF}(\theta, X, R, T; \widehat{\pi}, \widehat{m})] - P[\text{IF}(\theta, X, R, T; \widehat{\pi}, \widehat{m})]|.$$

Theorem 3. *Under assumption* **(A1)**, *for any estimators* $\widehat{\pi}, \widehat{m}$ *satisfying assumptions* **(A2)** *and* **(A3)**, *there exists a universal constant* $\mathfrak{C}$ *such that for any* $\delta > 0$,

$$\mathbb{P}\left(|\mathbf{I}| \leq \frac{\mathfrak{C}}{\mathbb{P}(T=1)}\sqrt{\frac{(m_0 + \pi_0 + 1 - \alpha)^2 \log\big(1/\delta\big) + (m_0 + \pi_0)^2}{|\mathcal{I}_2|}} \,\middle|\, \mathcal{D}_1\right) \;\geq\; 1 - \delta.$$

Moreover, there exists a universal constant $\mathfrak{C}'$ *such that*

$$\mathbb{E}\left[|\mathbf{I}|\middle|\mathcal{D}_1\right] \;\leq\; \frac{\mathfrak{C}'}{\mathbb{P}(T=1)}\sqrt{\frac{(m_0+\pi_0+1-\alpha)^2+(m_0+\pi_0)^2}{|\mathcal{I}_2|}} \leq \frac{\mathfrak{C}'}{\mathbb{P}(T=1)}\frac{m_0+\pi_0+1}{\sqrt{|\mathcal{I}_2|}}.$$

Proof. See Section S.6 of the appendix. □

This theorem is proved using techniques from empirical process theory by bounding $\sup_\theta\left|\mathbb{P}_N[\mathrm{IF}(\widehat{r}_\alpha,\dots)]-\mathbb{P}[\mathrm{IF}(\widehat{r}_\alpha,\dots)]\right|$. It gives a convergence rate of $\mathbb{P}_N[\mathrm{IF}(\widehat{r}_\alpha,\dots)]$ to $\mathbb{P}[\mathrm{IF}(\widehat{r}_\alpha,\dots)]$ that scales as $O(N^{-1/2})$, if $|\mathcal{I}_1| \asymp |\mathcal{I}_2| \asymp N$.

Using the definition of $\widehat{r}_\alpha$ and then combining Theorems 2 and 3 together with (18) yields the following main result.

Theorem 4. *Under assumption* **(A1)**, *for any estimators* $\widehat{\pi}, \widehat{m}$ *satisfying assumptions* **(A2)** *and* **(A3)**, *there exists a universal constant* $\mathfrak{C}$ *such that for any* $\delta > 0$ *with probability at least* $1-\delta$,

$$\begin{aligned}\mathbb{P}_{(X,Y)\sim Q_X\times P_{Y|X}}\left(Y \in \widehat{C}(\widehat{r}_\alpha; X) \mid \mathcal{D}^{\mathrm{tr}}\right) \geq 1-\alpha \\ - \frac{\|\widehat{\pi}-\pi^\star\|_2}{\mathbb{P}(T=1)}\sup_\theta\|\widehat{m}(\theta,\cdot)-m^\star(\theta,\cdot)\|_2 \\ - \mathfrak{C}\frac{(m_0+\pi_0+1)}{\mathbb{P}(T=1)}\sqrt{\frac{\log(1/\delta)+1}{|\mathcal{I}_2|}}.\end{aligned} \tag{19}$$

Moreover,

$$\begin{aligned}\mathbb{P}_{(X,Y)\sim Q_X\otimes P_{Y|X}}\left(Y \in \widehat{C}(\widehat{r}_\alpha; X)\right) \;\geq\; (1-\alpha) - \mathbb{E}\left[\frac{\|\widehat{\pi}-\pi^\star\|_2}{\mathbb{P}(T=1)}\sup_\theta\|\widehat{m}(\theta,\cdot)-m^\star(\theta,\cdot)\|_2\right] \\ - \frac{\mathfrak{C}}{\mathbb{P}(T=1)}\frac{m_0+\pi_0+1}{\sqrt{|\mathcal{I}_2|}}.\end{aligned} \tag{20}$$

Proof. This is a direct result of Lemma 2, (18), and Theorem 3. □

Equation (19) of Theorem 4 provides a coverage guarantee conditional on the training data and (20) provides a bound on the unconditional coverage. Note that the slack for the coverage is the sum of two terms: the product bias from the estimation of $\pi^\star$ and $m^\star$ and a term of order $O(N^{-1/2})$ if we assume the two splits $\mathcal{I}_1$ and $\mathcal{I}_2$ are of similar size. Proposition 2a of Vovk (2012) establishes a conditional prediction coverage guarantee which also involves a slack analogous to the last term of (19). The miscoverage error slacks in (19) and (20) have clear meaning. The first slack (product of errors) comes from the double robustness property of IF and the second slack (of order $|\mathcal{I}_2|^{-1/2}$) comes from approximating $P[\mathrm{IF}(\cdots)]$ with $\mathbb{P}_{\mathcal{I}_2}[\mathrm{IF}(\cdots)]$.

Remark. *Theorem 4 only provides lower bounds on the (conditional or unconditional on* $\mathcal{D}^{\mathrm{tr}}$*) coverage probability. Without a continuity assumption on the distribution of* $R(X,Y)$ *when* $(X,Y) \sim Q_X \otimes P_{Y|X}$, *it is not possible to provide an upper bound. If there is an* r_α *such that* $P[\mathrm{IF}(r_\alpha, X, R, T; \pi^\star, m^\star)] = 0$, *then the conclusions of Theorem 4 can be made two-sided. The condition of existence of* r_α *which makes* $P[\mathrm{IF}(\cdots)] = 0$ *is same as saying that there exists an* r_α *such that* $\mathbb{P}(R(X,Y) \leq r_\alpha | T=1) = 1-\alpha$, *i.e., there are no jumps in the distribution of* $R(X,Y)|T=1$ *at* $(1-\alpha)$*-th quantile, or equivalently, the distribution function takes the value of* $1-\alpha$. *Under this condition, inequality* (20), *for instance, can be strengthened to*

$$\begin{aligned}\left|\mathbb{P}_{(X,Y)\sim Q_X\otimes P_{Y|X}}\left(Y \in \widehat{C}(\widehat{r}_\alpha; X)\right) - (1-\alpha)\right| \leq \mathbb{E}\left[\frac{\|\widehat{\pi}-\pi^\star\|_2}{\mathbb{P}(T=1)}\sup_\theta\|\widehat{m}(\theta,\cdot)-m^\star(\theta,\cdot)\|_2\right] \\ + \frac{\mathfrak{C}}{\mathbb{P}(T=1)}\frac{m_0+\pi_0+1}{\sqrt{|\mathcal{I}_2|}}.\end{aligned}$$

Similar strengthening also holds for (19). *See the proof of Theorem 4 for details.*

6 Methodology without sample splitting

In this section, we provide an alternative methodology that builds upon Algorithm 1 but is potentially more efficient by avoiding sample splitting. The procedure is summarized in Algorithm 2.

Algorithm 2: Full doubly robust prediction

Input: Training data $\mathcal{D}^{\text{tr}} = \mathcal{D}_P^{\text{tr}} \cup \mathcal{D}_Q^{\text{tr}}$; Coverage probability $1-\alpha$, a training method $\mathcal{A}$ and estimators $\widehat{\pi}, \widehat{m}$, the point for prediction x.

Output: A valid prediction set $\widehat{C}_\alpha(x)$.

1 Fit the training method $\mathcal{A}$ on $\mathcal{D}^{\text{tr}}$ and using fitted method $\mathcal{A}$, construct an increasing (nested) sequence of sets $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$. Here $\mathcal{T}$ is a subset of $\mathbb{R}$. The nested sets $\{\mathcal{F}_t\}_{t\in\mathcal{T}}$ can depend arbitrarily on $\mathcal{D}^{\text{tr}}$.

2 For each $i \in [N]$ that satisfies $T_i = 0$, define the conformal score

$$r_i \;=\; r(Z_i) := \inf\{t \in \mathcal{T} :\, Z_i \in \mathcal{F}_t\}.$$

3 Fit estimators $\widehat{\pi}, \widehat{m}$ on $\mathcal{D}^{\text{tr}}$ and find the smallest $\widehat{\theta} = \widehat{r}_\alpha$ such that $\mathbb{P}_N[\text{IF}(\theta, X, R, T; \widehat{\pi}, \widehat{m})] \geq 0$, where

$$\begin{aligned}\mathbb{P}_N\big[\text{IF}(\theta, X, R, T; \widehat{\pi}, \widehat{m})\big] \;&=\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{t_i = 0\}\widehat{\pi}(x_i)\big[\mathbb{1}\{r_i \leq \theta\} - \widehat{m}(\theta, x_i)\big] \\ &+\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{t_i = 1\}[\widehat{m}(\theta, x_i) - (1-\alpha)].\end{aligned}$$

4 **return** the prediction set $\widehat{C}_\alpha(x) := \{y : R(x,y) \leq \widehat{r}_\alpha\}$.

Similar as in (18), let $\mathbf{I} := \{P[\text{IF}(\widehat{r}_\alpha, X, R, T; \widehat{\pi}, \widehat{m})] - \mathbb{P}_N[\text{IF}(\widehat{r}_\alpha, X, R, T; \widehat{\pi}, \widehat{m})]\}/\mathbb{P}(T=1)$. We now provide a bound on $\mathbf{I}$ in Theorem 5. Note that in addition to assumptions **(A1)**–**(A3)**, because we are doing the training and evaluating on the same dataset, we need some additional assumptions on the classes of estimators in order to apply results from empirical processes to ensure $\mathbf{I}$ converges to zero.

(A4) Assume that $\widehat{\pi}(\cdot)$ and its limit $\pi(\cdot)$ are in function classes $\mathcal{F}_\pi$ and $\widehat{m}(\cdot,\cdot)$ and its limit $m(\cdot,\cdot)$ are in $\mathcal{F}_m$, such that for some $\alpha_\pi, \alpha_m \geq 0$, the covering numbers satisfy $\forall \varepsilon > 0$,

$$\log N(\varepsilon, \mathcal{F}_\pi, L_2(Q)) \leq C\varepsilon^{-\alpha_\pi}, \text{ and } \log N(\varepsilon, \mathcal{F}_m, L_2(Q)) \leq C\varepsilon^{-\alpha_m},$$

where C is some constant and Q is any discrete probability measure, and the covering number $N\big(\varepsilon, \mathcal{F}, L_r(Q)\big)$ is defined in the same way as in Kearns et al. (1994).

Theorem 5. *Under assumption* **(A1)**, *for any estimators* $\widehat{\pi}, \widehat{m}$ *satisfying assumptions* **(A2)**, **(A3)** *and* **(A4)**, *there exists a universal constant* $\mathfrak{C}$ *such that for any* $\delta > 0$,

$$\mathbb{P}\Bigg(|\mathbf{I}| \leq \mathfrak{C}\Bigg\{N^{-1/(\alpha_m \vee 2)}\big(1 + \mathbb{1}\{\alpha_m = 2\}\log N\big) + N^{-1/(\alpha_\pi \vee 2)}\big(1 + \mathbb{1}\{\alpha_\pi = 2\}\log N\big) + \sqrt{\frac{(m_0 + \pi_0 + 1 - \alpha)^2 \log\big(\frac{1}{\delta}\big)}{N}}\Bigg\}/\mathbb{P}(T=1)\Bigg) \geq 1 - \delta,$$

where $\vee$ *is the maximum operator. Moreover, there exists a universal constant* $\mathfrak{C}'$ *such that*

$$\mathbb{E}\left[|\mathbf{I}|\right] \;\leq\; \frac{\mathfrak{C}'}{\mathbb{P}(T=1)}\left(N^{-1/(\alpha_m \vee 2)}\big(1 + \mathbb{1}\{\alpha_m = 2\}\log N\big) + N^{-1/(\alpha_\pi \vee 2)}\big(1 + \mathbb{1}\{\alpha_\pi = 2\}\log N\big) + \frac{m_0 + \pi_0 + 1 - \alpha}{\sqrt{N}}\right).$$

The proof of this theorem is in Section S.7 of the appendix.

Using the definition of $\widehat{r}_\alpha$ and combining Theorems 2 and 3 together with (18) yields the following result for the final coverage of the prediction region from Algorithm 2.

Theorem 6. *Under assumption* **(A1)**, *for any estimators* $\widehat{\pi}, \widehat{m}$ *satisfying assumptions* **(A2)**, **(A3)** *and* **(A4)**, *there exists a universal constant* $\mathfrak{C}$ *such that for any* $\delta > 0$ *with probability at least* $1 - \delta$,

$$\mathbb{P}_{(X,Y)\sim Q_X \times P_{Y|X}}\Big(Y \in \widehat{C}(\widehat{r}_\alpha; X) \mid \mathcal{D}^{\text{tr}}\Big) \geq 1 - \alpha - \frac{1}{\mathbb{P}(T=1)}\Bigg(\|\widehat{\pi} - \pi^\star\|_2 \sup_\theta \|\widehat{m}(\theta,\cdot) - m^\star(\theta,\cdot)\|_2$$
$$+ N^{-1/(\alpha_m \vee 2)}\big(1 + \mathbb{1}\{\alpha_m = 2\}\log N\big) + N^{-1/(\alpha_\pi \vee 2)}\big(1 + \mathbb{1}\{\alpha_\pi = 2\}\log N\big) + \mathfrak{C}(m_0 + \pi_0 + 1)\sqrt{\frac{\log\big(1/\delta\big) + 1}{N}}\Bigg).$$

Moreover,

$$\mathbb{P}_{(X,Y)\sim Q_X \otimes P_{Y|X}}\Big(Y \in \widehat{C}(\widehat{r}_\alpha; X)\Big) - (1-\alpha)$$
$$\geq -\frac{\mathfrak{C}}{\mathbb{P}(T=1)}\Bigg(N^{-1/(\alpha_m \vee 2)}\big(1 + \mathbb{1}\{\alpha_m = 2\}\log N\big) + N^{-1/(\alpha_\pi \vee 2)}\big(1 + \mathbb{1}\{\alpha_\pi = 2\}\log N\big) + \frac{m_0 + \pi_0 + 1}{\sqrt{N}}\Bigg)$$
$$- \mathbb{E}\left[\frac{\|\widehat{\pi} - \pi^\star\|_2}{\mathbb{P}(T=1)} \sup_\theta \|\widehat{m}(\theta,\cdot) - m^\star(\theta,\cdot)\|_2\right].$$

Proof. This is a direct result of Lemma 2, (18), and Theorem 5. □

Comparing the results of Theorem 6 with Theorem 4 in terms of the slack in the miscoverage probability, we notice that the full data based prediction set can have larger miscoverage error than the split data version. However, in terms of how close $\widehat{r}_\alpha$ is to r_α, we expect the full data version to perform better compared to the split data version. This is expected because with full data version the quantile estimator is based on N observations rather than $|\mathcal{I}_2|$ observations which is smaller than N. The variance of the quantile estimator can be better up to a constant of $N/|\mathcal{I}_2|$. We do not pursue these variance comparisons for $\widehat{r}_\alpha$ as that is not our goal.

7 Simulation Studies

In practice, unless $R(\cdot,\cdot)$ is a map that is independent of data, we further split $\mathcal{D}_1$ of Algorithm 1 and use the first split to train $R(\cdot,\cdot)$, while the second split is used to estimate the two nuisance parameters $\pi^\star(\cdot)$ and $m^\star(\cdot,\cdot)$. We include the two algorithms in Algorithm 1 and Algorithm 2 under the name "DRP w. three splits" and "DRP w. full data" respectively. We also include simulation results where R, $\widehat{m}$ and $\widehat{\pi}$ are trained on the same split and the remainder of the data is used for validation under the name "DRP w. two splits" in Section S.10.1 of the appendix. For both synthetic data and the real data set we use two kinds of score functions to estimate $R(\cdot,\cdot)$,

1. absolute residual score $|y - \widehat{\mu}(x)|$ where $\mu(x) := \mathbb{E}(Y|X=x)$ is estimated by ridge regression,
2. conditional quantile regression [2];

The nuisance parameters $\pi^\star$ and $m^\star(\cdot,\cdot)$ are estimated through SuperLearner[3] that includes both Random-Forest and generalized linear model (GLM). To avoid numerical issues, we clip the propensity score at 0.99

[2] the cutting edge algorithm employed in Lei and Candès (2020) paper

[3] SuperLearner uses cross-validation to estimate the performance of multiple machine learning models and then creates an optimal weighted average of those models using the test data. This approach has been proven to be asymptotically as accurate as the best possible prediction algorithm that is tested. For details please refer to Polley and van der Laan (2010).

to prevent $\widehat{\pi}$ from becoming unbounded. The proposed methods are compared against the weighted conformal prediction (WCP) method proposed in Tibshirani et al. (2019). Note that in the weighted conformal prediction method, the prediction interval is given by

$$\widehat{C}_n(x) = \mu_0(x) \pm \text{ Quantile } \left(1-\alpha; \sum_{i=1}^{n} p_i^w(x)\delta_{|Y_i-\mu_0(X_i)|} + p_{n+1}^w(x)\delta_\infty\right),$$

where $p^w(x)$ is a function which depends on the likelihood ratio between the two covariate distributions, or $\pi^\star(x)$. Therefore, when the distribution shift is too "large", i.e. $p_{n+1}^w(x)$ is larger than α, the width becomes ∞. And indeed we observe infinite widths for this method over 50% of the time across all the predicted points and the Monte Carlo replications on synthetic data and over 90% of the time on real data. For illustrative purposes, we truncate the width for WCP at 10 and 50 on synthetic data and real data respectively, and demonstrate the mean width from 500 Monte Carlo simulations.

It is shown in Lei and Candès (2020) that the CQR score would guarantee asymptotic conditional coverage $\mathbb{P}\big(Y_f \in \widehat{C}_{N,\alpha}(X_f)|X_f = x_f\big)$, we also conduct an experiment using this score for our method under the setting of Section 7.2, where we specify 200 test points of X_f (generated from standard normal distributions $N(0, I_4)$) and for each test point, 100 Y_f's following the distribution (21) are generated to test if they fall into the prediction sets trained by our method and WCP, where we report the average (reflected points) for each test point in Figure 3,which shows the coverage and width for each test point that is represented by the L_2 norm (the X axis). We also fit a smoothing spline (with default parameters of the R function **smooth.spline**) with these points.

7.1 Real data

We demonstrate the use of conformal prediction in the covariate shift setting in an empirical example. We re-analyze the data set used in Tibshirani et al. (2019) which is the airfoil data set from the UCI Machine Learning Repository which contains $N = 1503$ observations of a response Y (scaled sound pressure level of NASA airfoils), and a vector of covariates X with $d = 5$ dimension (log frequency, angle of attack, chord length, free-stream velocity, and suction side log displacement thickness). Label missingness was then generated as in Tibshirani et al. (2019) using a propensity score model analogous to the one specified in Section 7.2 below.

7.2 Synthetic data

Here we use the setting from Kang and Schafer (2007) where for each unit $i = 1, \ldots, n = 2000$, suppose that $(x_{i1}, x_{i2}, x_{i3}, x_{i4})^\top$ is independently distributed as $N(0, I_4)$ where I_4 is the 4×4 identity matrix. The y_i's are generated as

$$y_i = 210 + 27.4x_{i1} + 13.7x_{i2} + 13.7x_{i3} + 13.7x_{i4} + \varepsilon_i, \tag{21}$$

where $\varepsilon_i \sim N(0, 1)$, and the true propensity scores are

$$\mathbb{P}(T = 1|x_i) = \text{expit}\left(-x_{i1} + 0.5x_{i2} - 0.25x_{i3} - 0.1x_{i4}\right), \text{ where } \text{expit}(x_i^\top \alpha) = \frac{\exp(x_i^\top \alpha)}{1 + \exp(x_i^\top \alpha)}.$$

7.3 Simulation results

The mean coverage and width from 500 Monte Carlo simulations using the absolute residual score are shown in Table 1, where the middle column corresponds to synthetic data results and the rightmost column to real data results. Figure 1 display histograms of coverage and width from 500 runs on real data using either the absolute residual score and the CQR score while 2 shows the use of absolute residual score. For synthetic data, we keep the results using the two scores separate, with the absolute residual score in Figure 2 and the CQR score to Appendix S.10.2 for easier comparison between our DRP methods and the WCP method. Below we make a few observations on simulation results:

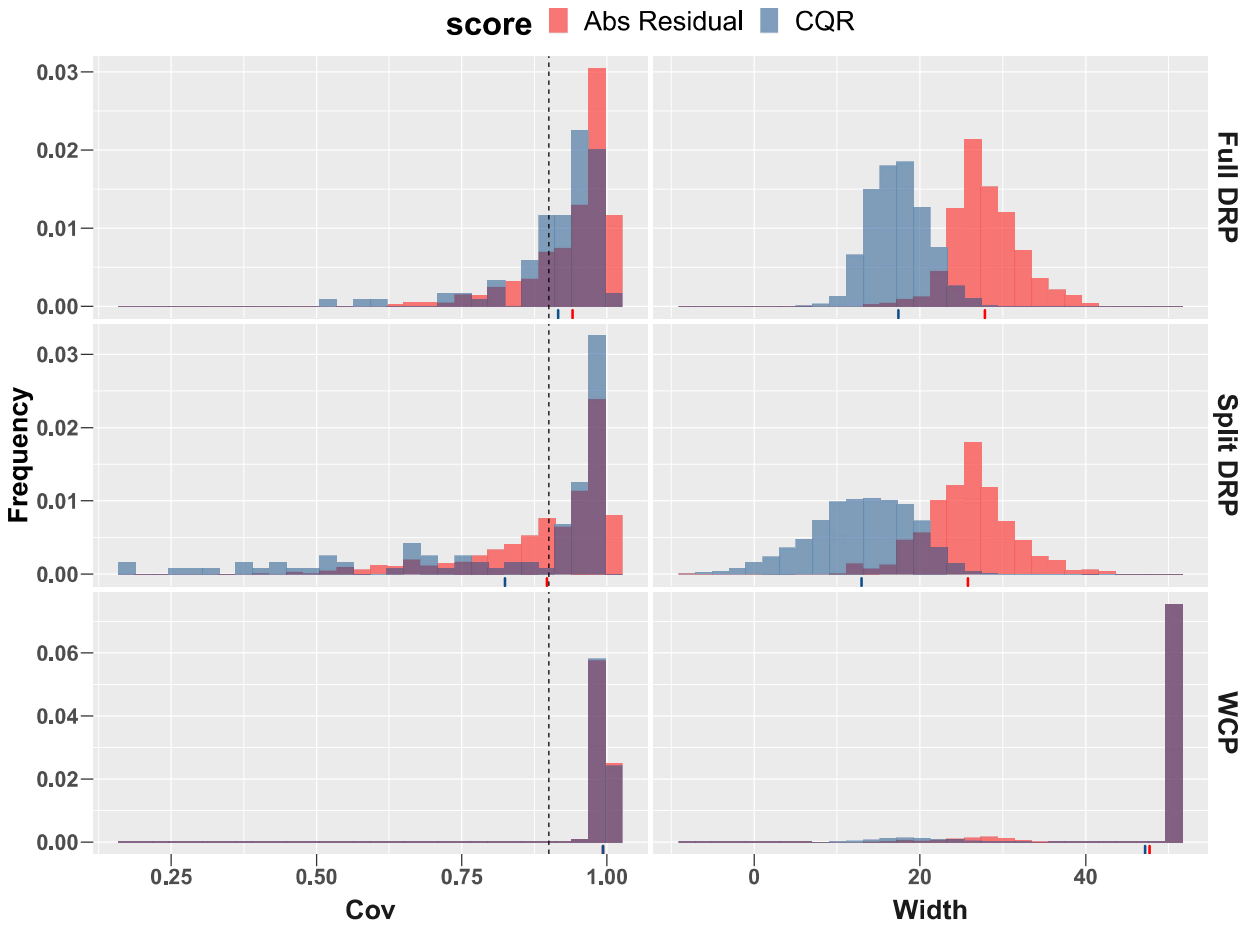

Figure 1: Histograms of coverage and width of Double Robust Prediction (DRP) and Weighted Conformal Prediction (WCP) on real data through either absolute residual score or the CQR score, where the width is truncated at 50 when WCP produces infinity width.

- WCP produces wider width and therefore, tends to over-cover by a considerable amount (by more than 7% over the nominal coverage of 90%).
- Doubly robust prediction with full data and multiple splits have similar performance with valid coverage. In practice, because using three splits guaranteed nominal coverage in sufficiently large sample with fewest assumptions, we recommend this approach as it does not appear to suffer much efficiency loss.
- For the conditional coverage simulation results described in the last paragraph at the start of Section 7, we see that DRP has similar coverage and much smaller width compared to WCP. Note that they both attain desired coverage when the norm of the test data is less than 2, which is what the L_2 norm of a standard 4-dimensional normal r.v. concentrates to. As the norm gets past 2, there are less data points and hence we get fewer observations.

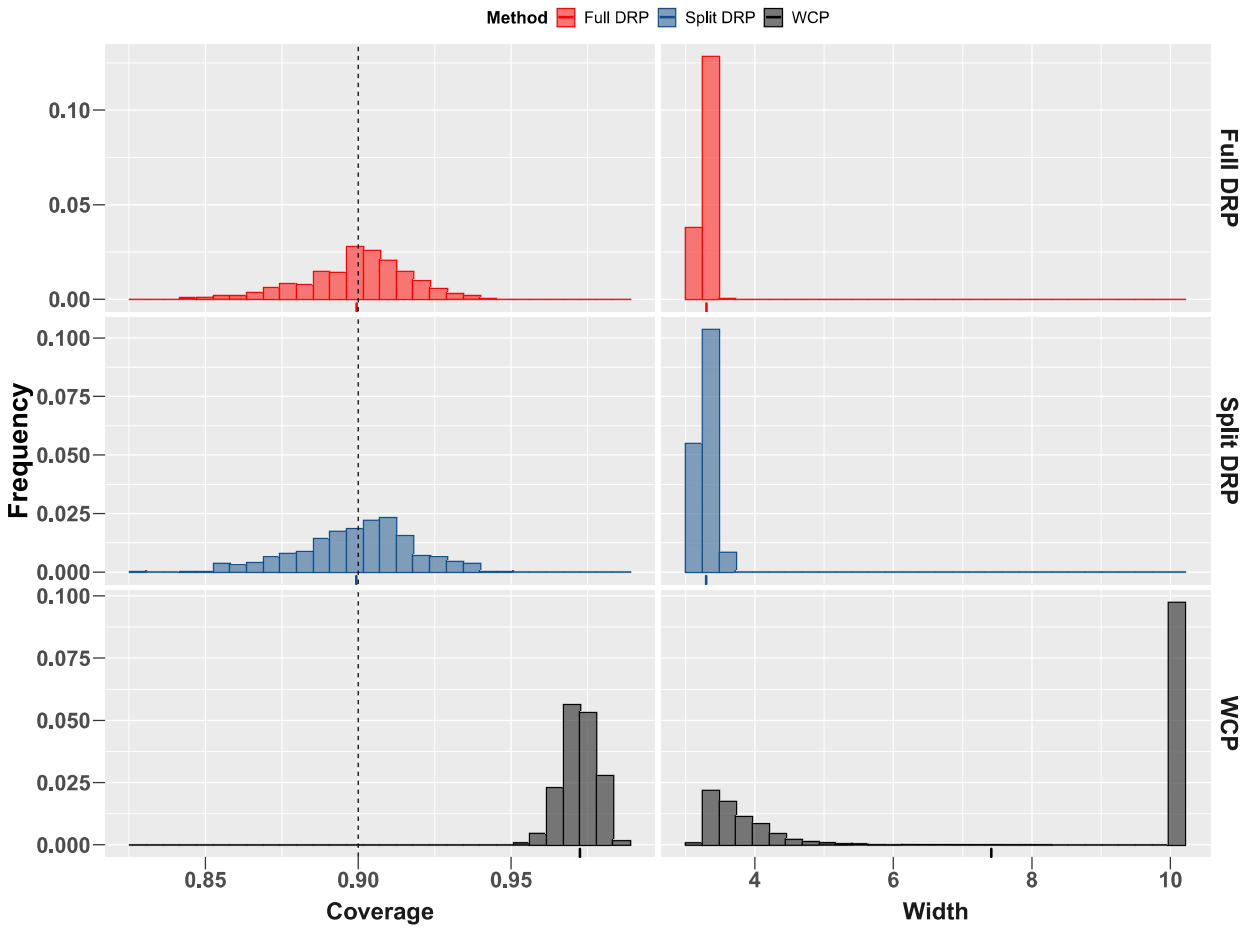

Figure 2: Histograms of coverage and width of Doubly Robust Prediction (DRP) and Weighted Conformal Prediction (WCP) on synthetic data using the absolute residual score. The width is truncated at 10 for WCP.

Mean coverage and width	Synthetic data		Real data	
from 500 monte carlo runs	Coverage	Width	Coverage	Width
DRP w. full data	0.90	3.29	0.94	27.85
DRP w. three splits	0.90	3.30	0.90	25.79
DRP w. two splits	0.90	3.29	0.88	25.19
WCP	0.97	7.41	0.99	47.71

Table 1: Coverage and width of DRP and WCP on synthetic and real data. Clearly, DRP improves on WCP in terms of width while maintaining coverage close to the nominal level of 0.9.

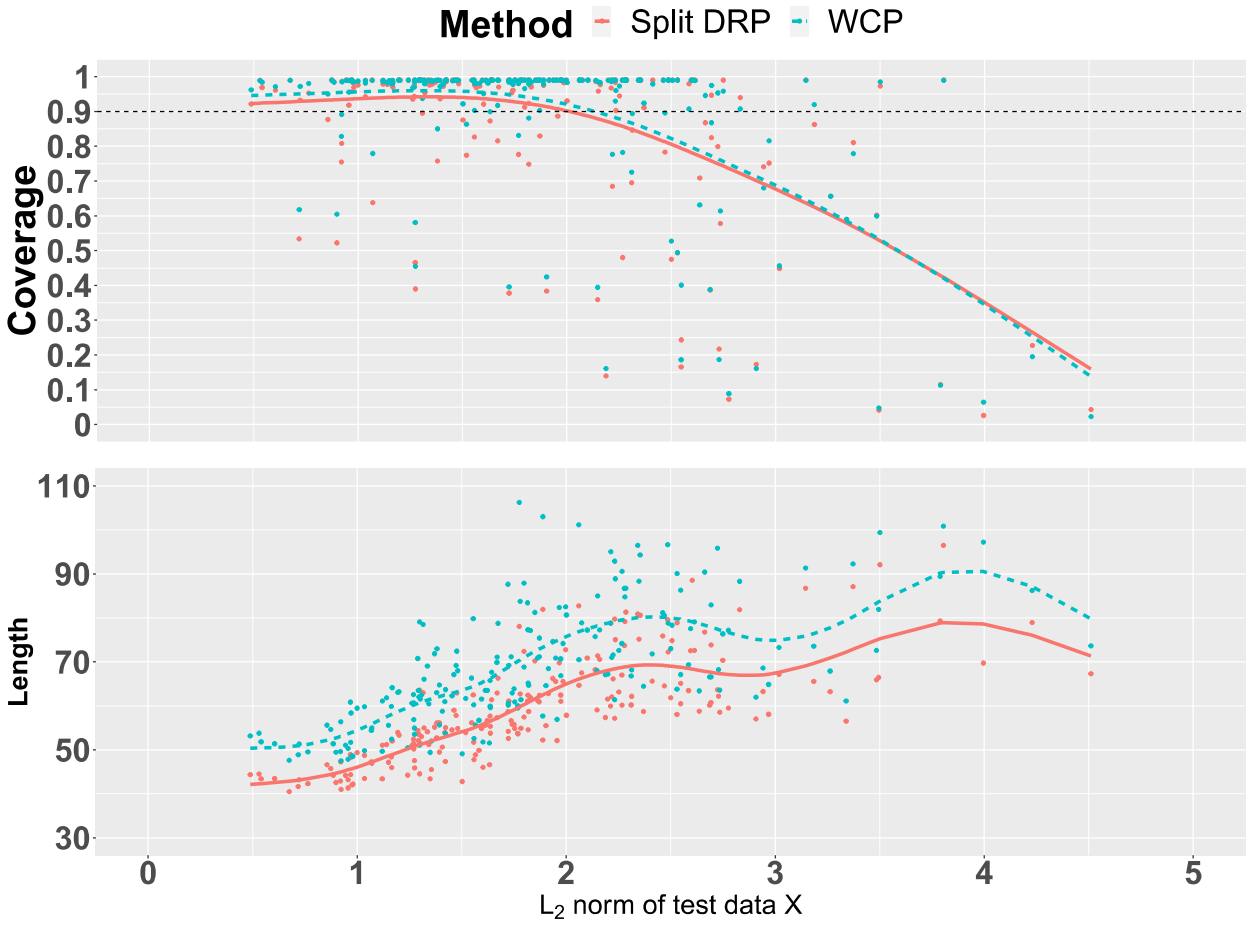

Figure 3: Estimated conditional coverage and length at 200 test points with Split DRP and WCP using the CQR score. The setting is described in the last paragraph at the start of Section 7. The points are the average of the coverage and width from 100 Monte Carlo simulations and the lines are drawn by fitting a smoothing splines for those points.

8 Aggregation of prediction sets

When, as typically the case in practice, multiple training methods are available, we propose to combine the proposed approach with that of Yang and Kuchibhotla (2021) in order to construct a better prediction set with smaller width. The detailed procedure is given in Algorithm 3 and can be shown by arguments given in Yang and Kuchibhotla (2021) to retain coverage validity while attaining the smallest width in large samples with high probability.

Table 2 reports results from a simulation study comparing the proposed doubly robust prediction algorithm to select an optimal tuning parameter for ridge regression-based conformal score, with cross-validation aimed at minimizing the ridge regression MSE. The simulation results confirm the proposed algorithm's ability to preserve marginal prediction coverage while optimizing prediction interval width, both in synthetic and real data sets.

Algorithm 3: Efficient doubly robust prediction

Input: Training data $\mathcal{D}^{\text{tr}} = \mathcal{D}_P^{\text{tr}} \cup \mathcal{D}_Q^{\text{tr}}$ split into $\mathcal{D}_1$ and $\mathcal{D}_2$, where $\mathcal{D}_1 = \{Z_i \in \mathcal{D}^{\text{tr}}, i \in \mathcal{I}_1\}$ and $\mathcal{D}_2 = \{Z_i \in \mathcal{D}^{\text{tr}}, i \in \mathcal{I}_2\}$; Estimators $\widehat{\pi}, \widehat{m}$, and training methods $\mathcal{A}_k, k \in [K]$; The prediction point x.

Output: A valid prediction set $\widehat{C}_\alpha^{\text{EFCP}}(x)$ with smallest width.

Fit training methods $\mathcal{A}_1, \ldots, \mathcal{A}_K$ on $\mathcal{D}_1$ and for each fitted method $\mathcal{A}_k$, construct an increasing (nested) sequence of sets $\{\mathcal{F}_t^{(k)}\}_{t \in \mathcal{T}}$. Here $\mathcal{T}$ is a subset of $\mathbb{R}$.

For each $i \in \mathcal{I}_2$ that satisfies $T_i = 0$, define the conformal score

$$r_k(Z_i) := \inf\{t \in \mathcal{T} : \, Z_i \in \mathcal{F}_t^{(k)}\}.$$

Fit estimators $\widehat{\pi}, \widehat{m}$ on $\mathcal{D}_1$ and solve for $\widehat{\theta} = \widehat{r}_{\alpha,k}$ as the solution to $\mathbb{P}_{\mathcal{I}_2}\big[\text{IF}(\widehat{\theta}, \mathcal{D}_2, r_k; \widehat{\pi}, \widehat{m})\big] = 0$, where

$$\begin{aligned}\mathbb{P}_{\mathcal{I}_2}[\text{IF}(\widehat{\theta}, X, R, T; \widehat{\pi}, \widehat{m})] \;=\; & \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \mathbb{1}\{t_i = 0\} \widehat{\pi}(x_i) \big[\mathbb{1}\{r_i \le \widehat{\theta}\} - \widehat{m}(\widehat{\theta}, x_i)\big] \\ & + \; \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \mathbb{1}\{t_i = 1\} [\widehat{m}(\widehat{\theta}, x_i) - (1-\alpha)].\end{aligned}$$

Compute the corresponding conformal prediction set as $\widehat{C}_k(x) \;:=\; \{y : \, r_k(x,y) \le \widehat{r}_{\alpha,k}\}$.

Set

$$\widehat{k} := \underset{1 \le k \le K}{\arg\min} \; \text{Width}(\widehat{C}_k(x)).$$

Here $\text{Width}(\cdot)$ can be any measure of width or volume of a prediction set. The quantity $\widehat{k}$ need not be unique and any minimizer can be chosen.

return the prediction set $\widehat{C}_{\widehat{k}}$ as $\widehat{C}_\alpha^{\text{EFCP}}$.

Mean coverage and width	Synthetic data		Real data	
from 500 monte carlo runs	Coverage	Width	Coverage	Width
Efficient DRP	0.90	3.30	0.89	27.30
DRP w. CV	0.89	3.32	0.84	18.61

Table 2: Coverage and width of efficient doubly robust prediction and doubly robust prediction with cross-validation on synthetic and real data. Efficient DRP improves on CV in terms of width while maintaining coverage close to the nominal level of 0.9.

9 Sensitivity analysis for latent covariate shift

Thus far, we have assumed that the covariate shift problem is primarily due to observed covariates, which we have denoted explainable covariate shift (or equivalently that outcomes for the target population are missing at random), an assumption that cannot be confirmed empirically without invoking a different non-testable assumption. In this section, we relax this assumption $P_{Y|X} = Q_{Y|X}$ and propose a sensitivity analysis for obtaining doubly robust calibrated prediction sets accounting for a latent covariate shift problem encoded

in a sensitivity parameter. We define a sensitivity function as

$$\gamma^\star(x,y) \;=\; \log\left(\frac{P_{Y=y|X}}{Q_{Y=y|X}}\Big/\frac{P_{Y=y_0|X}}{Q_{Y=y_0|X}}\right),$$

where y_0 is any baseline value for Y. Here $\gamma^\star(x,y)$ is the sensitivity analysis function encoding a hypothetical departure from the assumption that X suffices to account for the covariate shift problem with $\gamma^\star(x,y_0)=0$, and $\gamma^\star(x,y)=0$ for all y recovers the standard assumption of explainable covariate shift. For simplicity we take $y_0=0$. Our sensitivity analysis is inspired by a semiparametric approach for accounting for data missing not at random (MNAR), due to Chapter 5 of Robins et al. (2000) in the missing data literature. Formally, the sensitivity function $\gamma^\star$ can be represented in the missing data notation as

$$\gamma^\star(x,y) \;=\; \log\frac{\mathbb{P}(T=0|X=x,Y=y)\mathbb{P}(T=1|X,Y=0)}{\mathbb{P}(T=0|X,Y=0)\mathbb{P}(T=1|X=x,Y=y)}.$$

For any three functions $\eta(\cdot)$, $m(\cdot,\cdot)$, and $\gamma(\cdot,\cdot)$, let

$$\begin{aligned}\mathrm{IF}(\theta,x,y,r,t;\eta,m,\gamma) := {}& \mathbb{1}\{t=0\}\exp\{-\eta(x)-\gamma(x,y)\}\Big[\mathbb{1}\{r\le\theta\}-m(\theta,x)\Big]\\ &+\mathbb{1}\{t=1\}\Big[m(\theta,x)-(1-\alpha)\Big],\end{aligned}$$

where the two nuisance functions are

$$\begin{aligned}\eta^\star(x) &:= \log\frac{\mathbb{P}(T=0|X=x,Y=0)}{\mathbb{P}(T=1|X=x,Y=0)}\\ m^\star(\theta,x) &:= \mathbb{P}(R\le\theta|X=x,T=1).\end{aligned}$$

Per the sensitivity analysis framework, we assume $\gamma^\star(x,y)=\gamma(x,y)$ is known. See also Example 2 of Tsiatis (2014) for some explanations on why this is a nonparametric identified model. In this framework we have the following theorem.

Theorem 7. *Under the assumption that* $\mathbb{E}\frac{\mathbb{P}^2(T=1|X,Y)}{\mathbb{P}(T=0|X,Y)}$ *is finite and that the density of the conditional distribution of* $R|T=1$ *at* r_α *is bounded away from zero, the efficient influence function of* r_α*, the* $(1-\alpha)-$*quantile of* $R|T=1$ *in the nonparametric model which allows the distribution of* Z *to remain unrestricted, and the conditional log odds ratio function relating* T *to* R *given* X *is known to equal* $\gamma^\star$ *is given up to a proportionality constant by*

$$\begin{aligned}\psi(z)=\mathrm{IF}(r_\alpha,x,y,r,t;\eta^\star,m^\star,\gamma^\star) = {}& \mathbb{1}\{t=0\}\exp\{-\eta^\star(x)-\gamma^\star(x,y)\}\Big[\mathbb{1}\{r\le r_\alpha\}-m^\star(r_\alpha,x)\Big]\\ &+\mathbb{1}\{t=1\}\Big[m^\star(r_\alpha,x)-(1-\alpha)\Big]\end{aligned}$$

Furthermore, the moment function $\mathrm{IF}(r_\alpha,x,y,r,t;\eta,m,\gamma^\star)$ *satisfies the double robustness property that*

$$\mathbb{E}\big[\mathrm{IF}(r_\alpha,x,y,r,t;\eta,m,\gamma^\star)\big]=0,$$

if either $\eta=\eta^\star$ *or* $m=m^\star$.

Proof. See Section S.11 for a proof. □

Theorem 7 gives the efficient influence function of r_α, which provides a moment equation for r_α. And this can be leveraged as an estimating equation similar to Lemma 1 and Theorem 4 to yield a valid prediction set with the product bias from estimating the nuisance functions. In this case, estimation of nuisance functions is not straightforward and not further pursued in this paper. We also include a review of existing literature on this topic. The recent paper Jin et al. (2021) generalized the covariate shift setting to distributional shift

where the joint distribution of covariates and response can be different between the test and training data. The motivation is sensitivity analysis for individual treatment effects. They proposed a robust conformal prediction algorithm that builds upon the weighted conformal inference method from Tibshirani et al. (2019). The paper derives prediction sets that achieve marginal coverage if the propensity score is known exactly, allowing for latent covariate shift of a magnitude bounded by a known sensitivity parameter. They also proposed a second algorithm where the coverage is attained with high probability $1 - \delta$ conditional on the training data, and showed their results provide tight prediction sets in some cases that cannot be improved upon under their assumptions. The coverage bias of both methods is first order. Both their methods require the test point to be specified in advance so that they can compute a different quantile for every new x and thus may be computationally intensive. In terms of sensitivity analysis for unmeasured confounding, we note that their paper considers a different sensitivity framework than ours, as they posit the existence of an unobserved confounder U which together with X completely accounts for confounding. They encode the magnitude of unmeasured confounding in terms of upper and lower bounds for the likelihood ratio of density of U in the treated and untreated samples over the support of U. Thus, their sensitivity parameter appears to capture both the extent of residual confounding, but also reflects aspects of the density of U that may not be of scientific interest, and for which an investigator may not have any prior information. An implication of this choice of parameterization is that for a given sensitivity bound, the bound may reflect small amount of confounding over a wide support of U, or a large amount of confounding, over a narrow support of U, rendering such sensitivity analysis difficult to interpret. In contrast, we favor the approach of Robins et al. (2000) with a more direct sensitivity analysis, which in the counterfactual setting, encodes the departure from unconfoundedness in terms of a likelihood ratio for the counterfactual outcome in view of the treated and control arms conditional on observed covariates. Yin et al. (2021) also studied the sensitivity analysis of ITE using a similar approach as the first method proposed in Jin et al. (2021) while their method of analysis offers a different perspective.

10 Discussion

This paper has proposed three separate algorithms to construct prediction regions which are adaptive to unknown covariate shift between a population from which labeled data are available and an unlabeled population for which outcome prediction is in view. Our three proposed methods have been described as "Split doubly robust prediction", "Full doubly robust prediction" and "Efficient doubly robust prediction". The paper provided a rigorous analysis of the coverage properties of these algorithms, notably establishing that all have coverage bias of a product form and providing formal conditions under which all are asymptotically well-calibrated, in the sense of attaining the nominal coverage rate in large samples. "Split doubly robust prediction" has coverage guarantees in large samples under minimal conditions, but requires one to use a non-negligible subset of the data for training; in contrast, "Full doubly robust prediction" uses the entire data set both for training and prediction, and attains the nominal coverage in large samples under relatively stronger conditions. "Efficient doubly robust prediction" combines "Split doubly robust prediction" and the EFCP algorithm from Yang and Kuchibhotla (2021), which is empirically shown to potentially outperform a standard cross-validation approach. We conjecture that the proposed efficient doubly robust prediction algorithm is nearly as efficient as an oracle with a priori knowledge of the optimal prediction interval, although formally proving this result is left to future work. An important advantage of our framework is that its large sample efficiency and validity guarantees hold for any collection of machine learning techniques and their respective tuning parameters, under the relatively mild requirement that at least one of two estimated nuisance functions is consistent, without necessarily requiring fast convergence rates for the latter.

Another important contribution of the paper is to draw upon a key equivalence between the explainable covariate shift problem, the MAR assumption in the missing data literature, and the notion of unconfoundedness in the causal inference literature, to develop a sensitivity analysis approach to evaluate the extent to which prediction regions might be impacted by hypothetical departures from this assumption. Notably, the proposed methods readily extend to accommodate such sensitivity analysis via a slight modification of our procedure to incorporate a sensitivity parameter, without compromising the product bias or double robust-

ness property of the approach. Fully developing prediction inference for this sensitivity analysis framework however requires care in estimation of nuisance functions which, due to space limitation, we plan to consider in future work. Overall, this paper reveals and leverages deep connections between modern literatures of semiparametric theory, missing data and causal inference, and emerging methods for well-calibrated prediction inference. To the best of our knowledge, such connections have previously not been drawn upon as deliberately as shown to be possible in this work which we hope will generate both interest and further developments towards even more robust and efficient well-calibrated prediction. One possible line of future investigation might be to build on recent theory of higher order influence functions due to Robins and colleagues, see e.g. Robins et al. (2008) and Robins et al. (2017), which in principle could be used to reduce the second order product bias obtained in this paper to a higher order product bias, therefore potentially improving finite sample coverage over a wider range of regimes.

References

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *arXiv:1903.04684*, 2019.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

Emmanuel J Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*, 2021.

Matias D Cattaneo, Yingjie Feng, and Rocio Titiunik. Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880, 2021.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C Duchi. Robust validation: Confident predictions even when distributions shift. *arXiv preprint arXiv:2008.04267*, 2020.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

Xiangli Chen, Mathew Monfort, Anqi Liu, and Brian D Ziebart. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pages 1270–1279. PMLR, 2016.

Victor Chernozhukov, Ivan Fernandez-Val, and Alfred Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009.

Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On Learning Theory*, pages 732–749. PMLR, 2018.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48), 2021a.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864, 2021b.

Y Fan and S Park. Sharp bounds on the distribution of the treatment effect and their statistical inference,% forthcoming in econometric theory. 2007.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *arXiv preprint arXiv:2106.00170*, 2021.

Irwin Guttman. Statistical tolerance regions. *Classical and Bayesian*, 1967.

László Györfi and Harro Walk. Nearest neighbor based conformal prediction. *Pub. Inst. Stat. Univ. Paris*, Special issue in honour of Denis Bosq's 80th birthday(63):173–190, 2020.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *arXiv preprint arXiv:2111.12161*, 2021.

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539, 2007.

Michael J Kearns, Umesh Virkumar Vazirani, and Umesh Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

Byol Kim, Chen Xu, and Rina Foygel Barber. Predictive inference is free with the jackknife+-after-bootstrap. *Advances in Neural Information Processing Systems*, 33, 2020.

Danijel Kivaranovic, Robin Ristl, Martin Posch, and Hannes Leeb. Conformal prediction intervals for the individual treatment effect. *arXiv preprint arXiv:2006.01474*, 2020.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pages 1882–1886. PMLR, 2018.

Kalimuthu Krishnamoorthy and Thomas Mathew. *Statistical tolerance regions: theory, applications, and computation*, volume 744. John Wiley & Sons, 2009.

Arun K Kuchibhotla and Richard A Berk. Nested conformal prediction sets for classification with applications to probation data. *arXiv preprint arXiv:2104.09358*, 2021.

Arun Kumar Kuchibhotla. Exchangeability, conformal prediction, and rank tests. *arXiv preprint arXiv:2005.06095*, 2020.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*, 2020.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

Whitney K Newey. Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135, 1990.

Roberto I Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. Split conformal prediction for dependent data. *arXiv preprint arXiv:2203.15885*, 2022.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. Pac prediction sets under covariate shift. *arXiv preprint arXiv:2106.09848*, 2021.

J Pfanzagl. Investigating the quantile of an unknown distribution. In *Contribution to Applied Statistics*, pages 111–126. Springer, 1976.

Dimitris N Politis. *Model-Free Prediction and Regression: A Transformation-Based Approach to Inference*. Springer, 2015.

Eric C. Polley and Mark J. van der Laan. Super learner in prediction. 2010.

Hongxiang Qiu, Edgar Dobriban, and Eric Tchetgen Tchetgen. Distribution-free prediction sets adaptive to unknown covariate shift. *arXiv preprint arXiv:2203.06126*, 2022.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Sashank Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Adaptive transfer learning. *The Annals of Statistics*, 49(6):3618–3649, 2021.

James Robins, Lingling Li, Eric Tchetgen, and Aad van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

James M Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN, 2000.

James M Robins, Andrea Rotnitzky, and Daniel O Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, pages 3543–3553, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel J Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, pages 3538–3548, 2020.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Formal mode of statistical inference for causal effects. *Journal of statistical planning and inference*, 25(3):279–292, 1990.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Aldo Solari and Vera Djordjilović. Multi split conformal prediction. *Statistics & Probability Letters*, 184: 109395, 2022.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss, 2012.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Jiaye Teng, Zeren Tan, and Yang Yuan. T-sci: A two-stage conformal inference algorithm with guaranteed coverage for cox-mlp. In *International Conference on Machine Learning*, pages 10203–10213. PMLR, 2021.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Anastasios Tsiatis. Sensitivity analysis: A semi-parametric perspective. In *Handbook of Missing Data Methodology*, pages 403–428. Chapman and Hall/CRC, 2014.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Aad W van der Vaart and Jon A Wellner. *Weak convergence and empirical processes*. Springer, 1996.

AW van der Vaart. Semiparametric statistics, ecole d'ete de saint-flour 1999. in "lectures on probability theory and statistics", 2002.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490, 2012.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Yachong Yang and Arun Kumar Kuchibhotla. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.

Mingzhang Yin, Claudia Shi, Yixin Wang, and David M. Blei. Conformal sensitivity analysis for individual treatment effects, 2021.

Margaux Zaffran, Aymeric Dieuleveut, Olivier Féron, Yannig Goude, and Julie Josse. Adaptive conformal predictions for time series. *arXiv preprint arXiv:2202.07282*, 2022.

Anru Zhang, Lawrence D Brown, and T Tony Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019.

Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

Supplement to "Doubly robust calibration of prediction sets under covariate shift"

Abstract

This supplement contains the proofs to all the main results in the paper and some supporting lemmas.

S.1 A review of conformal inference with non-exchangeable data

Politis (2015) considered a transformation based approach that transforms non-i.i.d. data to i.i.d. data and applies i.i.d. data prediction sets on the transformations. For example, in an AR(1) model, transform the observed data to obtain estimated innovations which are assumed i.i.d. and predict the future innovation, then observed data forecast is obtained using the AR(1) model estimates and the prediction of future innovation. Chernozhukov et al. (2018) dealt with time series data where they developed a randomization method by including block structure in the permutation scheme and showed asymptotic validity under some modeling assumptions on the conformity score when exchangeability fails, see also Chernozhukov et al. (2021a) and Chernozhukov et al. (2021b) from the same authors under more general settings. Cattaneo et al. (2021) recently described a different approach to obtain prediction regions with time series data in a synthetic control framework. Recently, Oliveira et al. (2022) proved that the split conformal prediction methodology retains asymptotic coverage guarantee for several dependent data settings.

In a separate strand of work, Gibbs and Candès (2021) developed an adaptive approach based on ideas from conformal inference that builds predictions sets in an online setting where the data generating distribution is allowed to vary over time and established coverage validity over the long term; see also follow-up work in Zaffran et al. (2022) for an adaptive, tuning-free method. These works partly combine ideas from online learning and sequential prediction literature.

Candès et al. (2021) developed a method for survival analysis subject to administrative censoring that has approximate marginal coverage if the censoring mechanism or the conditional survival function is estimated well, and Teng et al. (2021) focused on a similar censoring scenario under a Cox proportional hazards model under the strong ignorability condition. A recent paper Barber et al. (2022) designed a new technique for non-exchangeable data that does not treat data points symmetrically and is robust against distribution shift. In almost all of these cases, the coverage guarantee is attained asymptotically as the number of training samples diverges to infinity.

S.2 Connection to Causal Inference

In this section, we briefly discuss how the goal of prediction in the covariate shift setting connects with that of prediction of counterfactuals and individual treatment effects in a potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990). Lei and Candès (2020) were the first to formally draw this connection. Given N subjects, let $A_i \in \{0,1\}$ denote a binary treatment indicator, $(Y_i(1), Y_i(0))$ be the pair of potential outcomes for unit i, and X_i be the corresponding vector of measured covariates needed to control of confounding. We assume that

$$\big(Y_i(1), Y_i(0), A_i, X_i\big) \overset{\text{i.i.d.}}{\sim} (Y(1), Y(0), A, X),$$

where $(Y(1), Y(0), A, X)$ is a random vector. Under the stable unit treatment value assumption (SUTVA) commonly assumed in the literature (see e.g. Rubin (1990)), the observed dataset consists of triples $(Y_i^{\text{obs}}, A_i, X_i)$ where

$$Y_i^{\text{obs}} = \begin{cases} Y_i(1), & A_i = 1, \\ Y_i(0), & A_i = 0. \end{cases}$$

For regularity, we assume the distributions of $X|A=1$ and $X|A=0$ are absolutely continuous with respect to each other. In this counterfactual setting, we wish to predict the individual treatment effect (ITE)

$\tau_i := Y_i(1) - Y_i(0)$, which cannot be observed because for every unit only one potential outcome is observed while the other is missing. In order to obtain such prediction, we make the standard unconfoundedness assumption (also known as strong ignorability condition) that $(Y(1), Y(0)) \perp A \,|\, X$, the counterfactual analog of MAR. Our approach thus yields prediction intervals for ITE with valid coverage for subjects in the study, for whom one potential outcome is observed; below, we briefly also discuss how one might obtain prediction intervals for subjects not in the study, for whom both potential outcomes are missing, but a covariate X is available.

For any treated unit i in the study, i.e. with $A_i = 1$, we construct a prediction interval $\widehat{C}_i^{\text{ITE}}$ for τ_i such that $\widehat{C}_i^{\text{ITE}} = Y_i^{\text{obs}} - \widehat{C}_0(X_i)$, where $\widehat{C}_0(x)$ satisfies

$$\mathbb{P}\big(Y(0) \in \widehat{C}_0(X) \mid A = 1\big) \geq 1 - \alpha; \tag{E.1}$$

Similarly, for any untreated unit, we construct $\widehat{C}_i^{\text{ITE}} = \widehat{C}_1(X_i) - Y_i^{\text{obs}}$, where $\widehat{C}_1(x)$ satisfies

$$\mathbb{P}\big(Y(1) \in \widehat{C}_1(X) \mid A = 0\big) \geq 1 - \alpha. \tag{E.2}$$

Thus, such construction has a guaranteed coverage for τ_i because

$$\begin{aligned}\mathbb{P}\big(Y_i(1) - Y_i(0) \in \widehat{C}_i^{\text{ITE}}\big) &= \mathbb{P}(A_i = 1)\mathbb{P}(Y_i(0) \in \widehat{C}_0(X_i) | A_i = 1) + \mathbb{P}(A_i = 0)\mathbb{P}(Y_i(1) \in \widehat{C}_1(X_i) | A_i = 0) \\ &\geq (1-\alpha)\big(\mathbb{P}(A_i = 1) + \mathbb{P}(A_i = 0)\big) = 1 - \alpha, \quad \text{if (E.1) and (E.2) both hold.}\end{aligned}$$

A more general goal than (E.1) would be to build a prediction set $\widehat{C}_0$ such that

$$\mathbb{P}_{(X,Y(0)) \sim Q_X \times P_{Y(0)|X}} \Big(Y(0) \in \widehat{C}_0(X)\Big) \geq 1 - \alpha, \tag{E.3}$$

where Q_X is some distribution for X. Note that (E.1) can be seen as a special case of (E.3) with $Q_X = P_{X|A=1}$. Based on the untreated samples, we learn the distribution of $Y(0) \mid X, A = 0$. And by the unconfoundedness assumption, this has the same distribution as $Y(0) \mid X, A = 1$, and also $Y(0) \mid X$.

A more challenging goal potentially of interest in many settings, might be to obtain prediction sets for future subjects not in the study, for whom neither $Y_i(1)$ nor $Y_i(0)$ is observed, but X_i is available. Without any knowledge of the relationship between the distributions of $Y(1)$ and $Y(0)$, a simple approach would be to obtain a pair of prediction intervals at level $1-\alpha/2$, namely $(\widehat{Y}_1^L(x), \widehat{Y}_1^R(x))$ for $Y(1)$ and $(\widehat{Y}_0^L(x), \widehat{Y}_0^R(x))$ for $Y(0)$. Then taking the difference of the two sets such that $\widehat{C}^{\text{ITE}} = \big(\widehat{Y}_1^L(x) - \widehat{Y}_0^R(x), \widehat{Y}_1^R(x) - \widehat{Y}_0^L(x)\big)$ would in principle yield a valid $(1-\alpha)$–coverage for $Y(1) - Y(0)$. See, e.g. Section 4.1 of Lei and Candès (2020). We also note that Fan and Park (2007) has provided sharp bounds on the distribution of the treatment effect which might aid in building a more precise prediction set.

S.3 Semiparametric theory and influence functions

In this section, we provide the definition of an influence function in the literature of semi-parametrics theory and give the derivation of (13).

Definition 1. *Given a semiparametric model $\mathcal{F}$, a law F^* in $\mathcal{F}$, and a class $\mathcal{A}$ of reg. parametric submodels of $\mathcal{F}$, a real valued functional*

$$\theta : \mathcal{F} \to \mathbb{R}$$

is said to be a **pathwise differentiable** *or regular parameter at F^* wrt $\mathcal{A}$ in model $\mathcal{F}$ iff there exists $\psi_{F^*}(x)$ in $\mathcal{L}_2(\dot{F}^*)$ such that for each submodel in $\mathcal{A}$, say indexed by t and with $F^* = F_{t^*}$, and score, say $S_t(t^*) = s_t(X; t^*)$ at t^*, it holds that*

$$\frac{\partial}{\partial t}\theta(F_t)\bigg|_{t=t^*} = \mathbb{E}_{F^*}\left[\psi_{F^*}(X) S_t(t^*)\right]$$

$\psi_{F^*}(.)$ is called a **gradient** *of θ at F^* (wrt $\mathcal{A}$). If, in addition, $\psi_{F^*}(X)$ has mean zero under F^*, $\psi_{F^*}(X)$ is most commonly referred to as an* **influence function** *of the functional θ at F^*.*

In our case of finding the $(1-\alpha)$-quantile for $R|T=1$, let t denote the index for the parametric submodels, $\theta(t)$ be the desired $(1-\alpha)$-quantile for $R|T=1$ with $\theta(t^*) := r_\alpha$. Let $u(R;\theta) := \mathbb{1}(R \le \theta | T=1) - (1-\alpha)$ with $\theta(t^*)$ satisfies that $\mathbb{E}_t[u(R;\theta(t))] = 0$. Then,

$$\begin{aligned}
0 &= \frac{\partial}{\partial t}\mathbb{E}_t\big[u(R;\theta(t))\big]\big|_{t=t^*} \\
&= \frac{\partial}{\partial t}\mathbb{E}_t\big[u(R;\theta(t^*))\big]\big|_{t=t^*} + \frac{\partial}{\partial \theta}\mathbb{E}_{t^*}[u(R;\theta)]|_{\theta=\theta(t^*)}\frac{\partial\theta(t)}{\partial t}\bigg|_{t=t^*}.
\end{aligned}$$

From this we conclude that

$$\begin{aligned}
\frac{\partial\theta(t)}{\partial t}\bigg|_{t=t^*} &= -\bigg\{\frac{\partial}{\partial\theta}\mathbb{E}_{t^*}\big[u(R;\theta)\big]\big|_{\theta=\theta(t^*)}\bigg\}^{-1}\frac{\partial}{\partial t}\mathbb{E}_t\big[u(R;\theta(t^*))\big]\big|_{t=t^*} \\
&\propto \frac{\partial}{\partial t}\mathbb{E}_t\bigg\{\mathbb{E}_t\Big(\mathbb{1}\{R\le\theta\} \mid T=1, X\Big)\Big|T=1\bigg\}\bigg|_{t^*} \\
&= \frac{\partial}{\partial t}\mathbb{E}_t\bigg\{\mathbb{E}_t\Big(\mathbb{1}\{R\le\theta\} \mid T=0, X\Big)\Big|T=1\bigg\}\bigg|_{t^*} \\
&= \frac{\partial}{\partial t}\mathbb{E}_t\bigg\{\mathbb{E}_{t^*}\Big(\mathbb{1}\{R\le\theta\} \mid T=0, X\Big)\;\Big|\;T=1\bigg\}\bigg|_{t^*} \\
&\qquad + \frac{\partial}{\partial t}\mathbb{E}_{t^*}\bigg\{\mathbb{E}_t\Big(\mathbb{1}\{R\le\theta\} \mid T=0, X\Big)\;\Big|\;T=1\bigg\}\bigg|_{t^*} \\
&=: I + II.
\end{aligned}$$

Let $S(Z) := S((1-T)Y, T, X) = \mathbb{1}\{T=0\}S(Y|T=0,X) + S(T|X) + S(X)$ be the score vector for all the observed data. We analyze the two terms separately. For the first term, because $R \perp T|X$, we have that

$$\begin{aligned}
I &= \mathbb{E}_{t^*}\Big[E_{t^*}\big(\mathbb{1}\{R\le\theta\} \mid T=0, X\big)S_{X|T=1}\;\Big|\;T=1\Big] \\
&= \mathbb{E}_{t^*}\bigg[\frac{\mathbb{1}\{T=1\}}{\mathbb{P}(T=1)}E_{t^*}\big(\mathbb{1}\{R\le\theta\} \mid T=0, X\big)S_{X|T=1}\bigg] \\
&= \mathbb{E}_{t^*}\bigg[\frac{\mathbb{1}\{T=1\}}{\mathbb{P}(T=1)}\mathbb{P}(R\le\theta \mid X)S_{X|T}\bigg] \\
&= \mathbb{E}_{t^*}\bigg[\frac{\mathbb{1}\{T=1\}}{\mathbb{P}(T=1)}\big\{\mathbb{P}(R\le\theta \mid X) - \mathbb{E}(\mathbb{P}(R\le\theta \mid X)|T)\big\}S_{X|T}\bigg] \\
&= \mathbb{E}_{t^*}\bigg[\frac{\mathbb{1}\{T=1\}}{\mathbb{P}(T=1)}\big\{\mathbb{P}(R\le\theta \mid X) - \mathbb{E}(\mathbb{P}(R\le\theta \mid X)|T)\big\}S_{X,T}\bigg] \\
&= \mathbb{E}_{t^*}\bigg[\frac{\mathbb{1}\{T=1\}}{\mathbb{P}(T=1)}\big\{\mathbb{P}(R\le\theta \mid X) - \mathbb{E}(\mathbb{P}(R\le\theta \mid X)|T)\big\}S_Z\bigg] \\
&= \mathbb{E}_{t^*}\bigg[\mathbb{1}\{T=1\}\Big\{\mathbb{E}_{t^*}\big(\mathbb{1}\{R\le\theta\} \mid X\big) - (1-\alpha)\Big\}S(Z)\bigg]/\mathbb{P}(T=1),
\end{aligned}$$

For the second term, for brevity let $A := \mathbb{1}\{R \le \theta\}$,

$$
\begin{aligned}
II &= \frac{\partial}{\partial t}\mathbb{E}_{t^*}\left\{\frac{\mathbb{1}\{T=1\}}{\mathbb{P}(T=1)}\mathbb{E}_t\Big(\mathbb{1}\{R\le\theta\} \mid T=0, X\Big)\right\}\Big|_{t^*} \\
&= \mathbb{E}_{t^*}\left\{\frac{\mathbb{1}\{T=1\}}{\mathbb{P}(T=1)}\mathbb{E}_{t^*}\Big(AS_{A|T=0,X} \mid T=0, X\Big)\right\} \\
&= \mathbb{E}_{t^*}\left\{\frac{\mathbb{1}\{T=1\}}{\mathbb{P}(T=1)}\mathbb{E}_{t^*}\Big(AS_{A|T=0,X}(A)\Big|T=0, X\Big)\right\} \\
&= \mathbb{E}_{t^*}\left[\mathbb{E}_{t^*}\left\{AS_{A|T=0,X}(A, T=0, X)\Big|T=0, X\right\} \;\Big|\; T=1\right] \\
&= \mathbb{E}_{t^*}\left[\mathbb{E}_{t^*}\left\{\frac{A\mathbb{1}\{T=0\}}{\mathbb{P}(T=0|X)}S_{A|T=0,X}(A, T=0, X)\Big|X\right\} \;\Big|\; T=1\right] \\
&= \mathbb{E}_{t^*}\left[\mathbb{E}_{t^*}\left\{\frac{A\mathbb{1}\{T=0\}}{\mathbb{P}(T=0|X)}S_{A|T,X}(A, T, X)\Big|X\right\} \;\Big|\; T=1\right] \\
&= \int\int\int \frac{A\mathbb{1}\{T=0\}}{\mathbb{P}(T=0|X=x)}S_{A|T,x}(r, T, X)f(r|T, x)f(T|x)f(x|T=1)\mathrm{d}r\mathrm{d}x\mathrm{d}T \\
&= \int\int\int \frac{A\mathbb{1}\{T=0\}\mathbb{P}(T=1|X=x)}{\mathbb{P}(T=1)\mathbb{P}(T=0|X=x)}S_{A|T,X}(r, T, X)f(r|T, X)f(T|x)f(x)\mathrm{d}r\mathrm{d}x\mathrm{d}T \\
&= \mathbb{E}\Big[A\mathbb{1}\{T=0\}\pi^\star(X)S_{A|T,X}(r, T, X)\Big]/\mathbb{P}(T=1) \\
&= \mathbb{E}\Big[(A-\mathbb{E}(A|T=0, X))\mathbb{1}\{T=0\}\pi^\star(X)S_{A|T,X}(R, T, X)\Big]/\mathbb{P}(T=1) \\
&= \mathbb{E}\Big[\big(\mathbb{1}\{R\le\theta\}-m^\star(t, X)\big)\mathbb{1}\{T=0\}\pi^\star(X)\big\{\mathbb{1}\{T=0\}S_{A|T,X}(R, T, X)+S_{T,X}(T, X)\big\}\Big]/\mathbb{P}(T=1) \\
&= \mathbb{E}\Big[\big(\mathbb{1}\{R\le\theta\}-m^\star(t, X)\big)\mathbb{1}\{T=0\}\pi^\star(X)S(Z)\Big]/\mathbb{P}(T=1).
\end{aligned}
$$

Combining the two terms together gives us

$$
\begin{aligned}
\frac{\partial}{\partial t}\theta(t)\Big|_{t^*} \propto \mathbb{E}_{t^*}\Big[&\Big\{\mathbb{1}\{T=1\}\Big(m^\star(\theta, X)-(1-\alpha)\Big) \\
&+\mathbb{1}\{T=0\}\pi^\star(x)\Big(\mathbb{1}\{R\le\theta\}-m^\star(\theta, X)\Big)\Big\}S(Z))\Big].
\end{aligned}
$$

We therefore conclude that the following function $\phi(\cdot)$ is proportional to a mean zero gradient of $\theta(t)$ at t^*,

$$
\phi(\theta, X, R, T; \pi^\star, m^\star) = \mathbb{1}\{T=0\}\pi^\star(X)\Big[\mathbb{1}\{R\le\theta\}-m^\star(\theta, X)\Big]+\mathbb{1}\{T=1\}\Big[m^\star(\theta, X)-(1-\alpha)\Big].
$$

S.4 Proof of Lemma 2

Proof of Lemma 2. Throughout the proof, we will write R instead of $R(X,Y)$ for convenience. Firstly, note that

$$\begin{aligned}
\mathbb{P}_{(X,Y)\sim Q_X\times P_{Y|X}}(R(X,Y)\le\theta\big|\theta) &= \mathbb{E}_{X\sim Q_X}\left[\mathbb{P}(R\le\theta\mid X,\theta)\big|\theta\right] \\
&= \mathbb{E}_{X\sim Q_X}\left[\mathbb{P}(R\le\theta\mid X,T=1,\theta)\big|\theta\right] \\
&= \int_{\mathcal{X}}\int_{-\infty}^{\theta} p(x|T=1)p(r|x,T=1)\mathrm{d}r\mathrm{d}x \\
&= \int_{\mathcal{X}}\int_{-\infty}^{\theta} p(x|T=1)p(r|x,T=0)\mathrm{d}r\mathrm{d}x \text{ (Using (7))} \\
&= \int_{\mathcal{X}}\int_{-\infty}^{\theta} \frac{p(x|T=1)}{p(x|T=0)}p(r|x,T=0)p(x|T=0)\mathrm{d}r\mathrm{d}x \\
&= \mathbb{E}\left[\frac{p(X|T=1)}{p(X|T=0)}\mathbb{1}\{R\le\theta\}\,\middle|\,T=0,\theta\right].
\end{aligned} \tag{E.4}$$

On the other hand, we can prove that

$$P[\mathrm{IF}(\theta,X,R,T;\pi,m)] = \mathbb{P}(T=1)\left\{\mathbb{E}\left[\frac{p(X|T=1)}{p(X|T=0)}\mathbb{1}\{R\le\theta\}\,\middle|\,T=0,\theta\right]-(1-\alpha)\right\}. \tag{E.5}$$

This combined with (E.4) completes the proof of (14). We will prove (E.5) in two steps.

$$P[\mathrm{IF}(\theta,X,R,T;\pi,m)] = \mathbb{E}\Big[\mathbb{P}(T=1|X)\big\{\mathbb{P}(R\le\theta|X,\theta)-(1-\alpha)\big\}\big|\theta\Big], \tag{Step 1}$$

and

$$\begin{aligned}
&\mathbb{E}\big[\mathbb{P}(T=1|X)\big\{\mathbb{P}(R\le\theta|X,\theta)-(1-\alpha)\big\}\big|\theta\big] \\
&\quad= \mathbb{P}(T=1)\left\{\mathbb{E}\left[\frac{p(X|T=1)}{p(X|T=0)}\mathbb{1}\{R\le\theta\}\,\middle|\,T=0,\theta\right]-(1-\alpha)\right\}.
\end{aligned} \tag{Step 2}$$

In the proof of (Step 1), we will use the fact that either $\pi(\cdot)$ or $m(\cdot,\cdot)$ represents the correct density ratio or the correct conditional distribution function. The proof of (Step 2) follows essentially from Bayes rule.

Proof of (Step 1). If $\pi(x)=\mathbb{P}(T=1|X=x)/\mathbb{P}(T=0|X=x)$ for all x (i.e., density ratio is correct), then we have

$$\mathbb{E}\left[\mathbb{1}\{T=0\}\pi(X)\big|X=x,R\right] = \mathbb{P}(T=0|X=x,R)\frac{\mathbb{P}(T=1|X=x)}{\mathbb{P}(T=0|X=x)} = \mathbb{P}(T=1|X=x).$$

The second equality here follows because T is independent of R given X. This implies that

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\{T=0\}\pi(X)\big\{\mathbb{1}\{R\le\theta\}-m(\theta,X)\big\}\big|\theta\right] \\
&\quad= \mathbb{E}\Big[\mathbb{P}(T=1|X)\big\{\mathbb{1}\{R\le\theta\}-m(\theta,X)\big\}\big|\theta\Big] \\
&\quad= \mathbb{E}\Big[\mathbb{P}(T=1|X)\big\{\mathbb{P}(R\le\theta|X,\theta)-m(\theta,X)\big\}\big|\theta\Big].
\end{aligned}$$

Similarly,

$$\mathbb{E}\Big[\mathbb{1}\{T=1\}\left\{m(\theta,X)-(1-\alpha)\right\}\big|\theta\Big] = \mathbb{E}\Big[\mathbb{P}(T=1|X)\big\{m(\theta,X)-(1-\alpha)\big\}\big|\theta\Big].$$

Hence, if $\pi(\cdot)$ is the true density ratio, then

$$P[\mathrm{IF}(\theta, X, R, T; \pi, m)] = \mathbb{E}\Big[\mathbb{P}(T=1|X)\big\{\mathbb{P}(R \le \theta|X,\theta) - (1-\alpha)\big\}\big|\theta\Big].$$

This completes the proof of (Step 1) when $\pi(\cdot)$ is the true density ratio.

If $m(\gamma, x) = \mathbb{E}[\mathbb{1}\{R \le \gamma\}|X = x]$ for all $\gamma \in \mathbb{R}$, $x \in \mathcal{X}$ (i.e., the conditional mean is correct), then using the conditional independence of R and T given X, we have

$$\mathbb{E}\Big[\mathbb{1}\{T=0\}\pi(X)\big\{\mathbb{1}\{R \le \theta\} - m(\theta, X)\big\}\big|\theta\Big] = 0.$$

Hence,

$$\begin{aligned}
P[\mathrm{IF}(\theta, X, R, T; \pi, m)] &= \mathbb{E}\Big[\mathbb{1}\{T=1\}\big\{m(\theta,X) - (1-\alpha)\big\}\big|\theta\Big] \\
&= \mathbb{E}\Big[\mathbb{P}(T=1|X)\big\{m(\theta,X) - (1-\alpha)\big\}\big|\theta\Big] \\
&= \mathbb{E}\Big[\mathbb{P}(T=1|X)\big\{\mathbb{P}(R \le \theta|X,\theta) - (1-\alpha)\big\}\big|\theta\Big].
\end{aligned}$$

This completes the proof of (Step 1) if $m(\cdot,\cdot)$ is the true conditional mean function.

Proof of (Step 2)**.**

$$\begin{aligned}
&\mathbb{E}\big[\mathbb{P}(T=1|X)\big\{\mathbb{P}(R \le \theta|X,\theta) - (1-\alpha)\big\}\big|\theta\big] \\
&= \mathbb{E}\big[\mathbb{P}(T=1|X)\mathbb{P}(R \le \theta|X,\theta)\big|\theta\big] - \mathbb{P}(T=1)(1-\alpha) \\
&= \mathbb{E}\left[\mathbb{1}\{T=0\}\frac{\mathbb{P}(T=1|X)}{\mathbb{P}(T=0|X)}\mathbb{P}(R \le \theta|X,\theta)\big|\theta\right] - \mathbb{P}(T=1)(1-\alpha) \\
&= \mathbb{E}\left[\mathbb{1}\{T=0\}\frac{\mathbb{P}(T=1|X)}{\mathbb{P}(T=0|X)}\mathbb{1}\{R \le \theta\}\big|\theta\right] - \mathbb{P}(T=1)(1-\alpha) \\
&\overset{(b)}{=} \frac{\mathbb{P}(T=1)}{\mathbb{P}(T=0)}\mathbb{E}\left[\mathbb{1}\{T=0\}\frac{p(X|T=1)}{p(X|T=0)}\mathbb{1}\{R \le \theta\}\big|\theta\right] - \mathbb{P}(T=1)(1-\alpha) \\
&= \frac{\mathbb{P}(T=1)}{\mathbb{P}(T=0)}\mathbb{E}_T\left\{\mathbb{E}\left[\mathbb{1}\{T=0\}\frac{p(X|T=1)}{p(X|T=0)}\mathbb{1}\{R \le \theta\}\Big|T,\theta\right]\right\} - \mathbb{P}(T=1)(1-\alpha) \\
&= \frac{\mathbb{P}(T=1)}{\mathbb{P}(T=0)}\mathbb{P}(T=0)\mathbb{E}\left[\frac{p(X|T=1)}{p(X|T=0)}\mathbb{1}\{R \le \theta\}\Big|T=0,\theta\right] - \mathbb{P}(T=1)(1-\alpha) \\
&= \mathbb{P}(T=1)\left\{\mathbb{E}\left[\frac{p(X|T=1)}{p(X|T=0)}\mathbb{1}\{R \le \theta\}\,\Big|\,T=0,\theta\right] - (1-\alpha)\right\},
\end{aligned}$$

where equality (b) comes from Bayes rule. This completes the proof of (Step 2). □

S.5 Proof of Theorem 2

Proof. By definition of the IF function, it holds $\forall \gamma \in \mathbb{R}$,

$$
\begin{aligned}
P\big[\mathrm{IF}(\gamma, X, R, \widehat{\pi}, \widehat{m})\big] &= P\Big[\mathbb{P}(T=0|X)\widehat{\pi}(X)\big\{\mathbb{P}(R \le \gamma|X) - \widehat{m}(\gamma, X)\big\} \\
&\quad + \mathbb{P}(T=1|X)\big\{\widehat{m}(\gamma, X) - (1-\alpha)\big\}\Big] \\
&= P\Big[\mathbb{P}(T=0|X)\big\{\widehat{\pi}(X) - \pi^\star(X)\big\}\big\{m^\star(\gamma, X) - \widehat{m}(\gamma, X)\big\}\Big] \\
&\quad + P\Big[\mathbb{P}(T=0|X)\pi^\star(X)\big\{m^\star(\gamma, X) - \widehat{m}(\gamma, X))\big\} \\
&\quad + \mathbb{P}(T=1|X)\big\{\widehat{m}(\gamma, X) - (1-\alpha)\big\}\Big] \\
&= P\Big[\mathbb{P}(T=0|X)\big\{\widehat{\pi}(X) - \pi^\star(X)\big\}\big\{m^\star(\gamma, X) - \widehat{m}(\gamma, X)\big\}\Big] \\
&\quad + P\Big[\mathbb{P}(T=1|X)\big\{m^\star(\gamma, X) - \widehat{m}(\gamma, X))\big\} \\
&\quad + \mathbb{P}(T=1|X)\big\{\widehat{m}(\gamma, X) - (1-\alpha)\big\}\Big] \\
&= P\Big[\mathbb{P}(T=0|X)\big\{\widehat{\pi}(X) - \pi^\star(X)\big\}\big\{m^\star(\gamma, X) - \widehat{m}(\gamma, X)\big\}\Big] \\
&\quad + P\Big[\mathbb{P}(T=1|X)\big\{m^\star(\gamma, X) - (1-\alpha)\big\}\Big].
\end{aligned}
\tag{E.6}
$$

Repeating the same calculation in (E.6) with either $\widehat{\pi}$ replaced by $\pi^\star$ and $\widehat{m}$ replaced by $m^\star$ yields $\forall \gamma \in \mathbb{R}$,

$$
P\big[\mathrm{IF}(\gamma, X, R, \pi, m)\big] = P\Big[\mathbb{P}(T=1|X)\big\{m^\star(\gamma, X) - (1-\alpha)\big\}\Big].
$$

Therefore,

$$
\begin{aligned}
&\sup_{\gamma \in \mathbb{R}} \Big| P\big[\mathrm{IF}(\gamma, X, R, \widehat{\pi}, \widehat{m}) - \mathrm{IF}(\gamma, X, R, \pi^\star, m^\star)\big] \Big| \\
&\quad = \sup_{\gamma \in \mathbb{R}} \Big| P\big[\mathbb{P}(T=0|X)\big\{\pi^\star(X) - \widehat{\pi}(X)\big\}\big\{m^\star(\gamma, X) - \widehat{m}(\gamma, X)\big\}\big] \Big| \\
&\quad \le \|\widehat{\pi} - \pi^\star\|_2 \sup_{\gamma} \|\widehat{m}(\gamma, \cdot) - m^\star(\gamma, \cdot)\|_2.
\end{aligned}
$$

The last inequality here follows from Cauchy–Schwarz inequality. □

S.6 Proof of Theorem 3

Proof. Without loss of generality assume the indexes in $\mathcal{I}_2$ is $1, \ldots, n$ with $n := |\mathcal{I}_2|$, and we expand $\mathbb{P}_{\mathcal{I}_2}\big[\mathrm{IF}(\theta, \mathcal{D}_2; \widehat{\pi}, \widehat{m})\big] - P\big[\mathrm{IF}(\theta, X, R, T; \widehat{\pi}, \widehat{m})\big]$ into three parts,

$$
\begin{aligned}
&\mathbb{P}_{\mathcal{I}_2}\big[\mathrm{IF}(\theta, \mathcal{D}_2; \widehat{\pi}, \widehat{m})\big] - P\big[\mathrm{IF}(\theta, X, R, T; \widehat{\pi}, \widehat{m})\big] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{t_i = 0\}\widehat{\pi}(x_i)\mathbb{1}\{r_i \le \theta\} - P\big[\mathbb{1}\{t_i = 0\}\widehat{\pi}(x_i)\mathbb{1}\{r_i \le \theta\}\big] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(\theta, x_i)\big(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}\big) - P\big[\widehat{m}(\theta, x_i)(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\})\big] \\
&\quad - (1-\alpha)\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{t_i = 1\} - \mathbb{P}(T=1)\right] \\
&=: \mathcal{R}_1(\theta) + \mathcal{R}_2(\theta) + \mathcal{R}_3,
\end{aligned}
$$

where the three terms will be controlled separately. In particular, $\sup_\theta |\mathcal{R}_1(\theta)|$ and $\sup_\theta |\mathcal{R}_2(\theta)|$ will be bounded using tools from the empirical processes theory. Also notice that conditional on training data $\mathcal{D}_1$, $\widehat{\pi}$ and $\widehat{m}$ are non-random functions and for ease of notation, we treat these functions as non-random and omit the conditioning part from this point onwards.

For $W_i = (X_i, Y_i, T_i), i \in \mathcal{I}_2$ and any function $f : \mathbb{R}^{d+2} \to \mathbb{R}$, for notation simplicity we define

$$\mathbb{G}_n f \;:=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big\{ f(W_i) - \mathbb{E}[f(W_i)] \big\}.$$

Bound on $\sup_\theta |\mathcal{R}_1(\theta)|$: We have a class of functions

$$\mathcal{F} = \big\{ f : f_{\widehat{\pi},\theta}(w) = \mathbb{1}\{t=1\}\widehat{\pi}(x)\mathbb{1}\{R(x,y) \le \theta\}, \forall \widehat{\pi}(\cdot) \in \mathcal{F}_\pi \big\}.$$

Notice that $\forall \theta \in \mathbb{R}$ and $\widehat{\pi}(\cdot) \in \mathcal{F}_\pi$, we have $|f_\theta(w)| \le \pi_0 \mathbb{1}\{t=1\}$. Therefore, $F(w) := \pi_0 \mathbb{1}\{t=1\}$ is an envelope function of $\{f_\theta(\cdot) : \theta \in \mathbb{R}\}$. Let $\|\cdot\|_{\mathcal{F}}$ denote the supremum norm $\|z\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |z(f)|$.

Applying Lemma 4 with $s(t,x) = \mathbb{1}\{t=0\}\widehat{\pi}(x)$ and $h(x,y) = R(x,y)$ gives us

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} \le \mathfrak{C}\pi_0, \tag{E.7}$$

where $\mathfrak{C}$ is a universal constant. Applying McDiarmid's inequality gives us

$$\mathbb{P}\big(\|\mathbb{G}_n\|_{\mathcal{F}} - \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} \ge u\big) \le \exp\bigg(-\frac{2u^2}{\sum_{i=1}^n c_i^2}\bigg) \le \exp\bigg(-\frac{2u^2}{\sum_{i=1}^n 4\pi_0^2/n}\bigg) = \exp(-u^2/2\pi_0^2), \tag{E.8}$$

where

$$\begin{aligned} c_i :=& \sup_{\substack{(x_i,y_i,t_i),(x_i',y_i',t_i'),\\ (x_1,y_1,t_1),\ldots,(x_n,y_n,t_n)}} \sup_\theta \sqrt{n}\bigg| \frac{1}{n}\sum_{j=1}^n \mathbb{1}\{t_j=0\}\widehat{\pi}(x_j)\mathbb{1}\{r_j \le \theta\} - \frac{1}{n}\sum_{j=1, j\ne i}^n \mathbb{1}\{t_j=0\}\widehat{\pi}(x_j)\mathbb{1}\{r_j\le\theta\} \\ &\qquad - \frac{1}{n}\mathbb{1}\{t_i'=0\}\widehat{\pi}(x_i')\mathbb{1}\{r_i'\le\theta\}\bigg| \\ \le& \sup_{(x_i,y_i,t_i),(x_i',y_i',t_i')} \sup_\theta \sqrt{n}\bigg|\frac{1}{n}\mathbb{1}\{t_i=0\}\widehat{\pi}(x_i)\mathbb{1}\{r_i\le\theta\} - \frac{1}{n}\mathbb{1}\{t_i'=0\}\widehat{\pi}(x_i')\mathbb{1}\{r_i'\le\theta\}\bigg| \le \frac{2\pi_0}{\sqrt{n}}. \end{aligned}$$

Substituting the expectation bound (E.7) in (E.8) and setting the right hand side of (E.8) to δ yields for another absolute constant $\mathfrak{C}'$

$$\mathbb{P}\bigg(\|\mathbb{G}_n\|_{\mathcal{F}} \ge \mathfrak{C}'\sqrt{\pi_0^2 + \pi_0^2\log\big(\tfrac{1}{\delta}\big)}\bigg) \le \mathbb{P}\bigg[\|\mathbb{G}_n\|_{\mathcal{F}} \ge \mathfrak{C}\bigg\{\pi_0 + \pi_0\sqrt{2\log\big(\tfrac{1}{\delta}\big)}\bigg\}\bigg] \le \delta. \tag{E.9}$$

Bound on $\sup_\theta |\mathcal{R}_2(\theta)|$: We define the class of functions $\mathcal{F} = \big\{f : f_\theta(w) = \widehat{m}(\theta, x_i)(\mathbb{1}\{t_i=1\} - \mathbb{1}\{t_i=0\})\big\}$ with $F(w) = m_0(\mathbb{1}\{t_i=1\} - \mathbb{1}\{t_i=0\})$ as its envelope.

Note that

$$\begin{aligned} \sup_\theta |\mathbb{G}_n f| &= \sup_\theta \bigg|\mathbb{G}_n\Big[(\mathbb{1}\{t_i=1\} - \mathbb{1}\{t_i=0\})\widehat{m}(\theta,x_i)\Big]\bigg| \\ &= \sup_\theta \bigg|\mathbb{G}_n\Big[(\mathbb{1}\{t_i=1\} - \mathbb{1}\{t_i=0\})\int_0^{m_0} \mathbb{1}\{\widehat{m}(\theta,x_i)\ge u\}\mathrm{d}u\Big]\bigg| \\ &= \sup_\theta \bigg|\int_0^{m_0} \mathbb{G}_n\Big[(\mathbb{1}\{t_i=1\} - \mathbb{1}\{t_i=0\})\mathbb{1}\{\widehat{m}(\theta,x_i)\ge u\}\mathrm{d}u\Big]\bigg| \\ &\overset{(a)}{=} \sup_\theta \bigg|\int_0^{m_0} \mathbb{G}_n\Big[(\mathbb{1}\{t_i=1\} - \mathbb{1}\{t_i=0\})\mathbb{1}\{h(x_i,u)\le\theta\}\mathrm{d}u\Big]\bigg|, \text{ for some function } h \\ &\le \int_0^{m_0} \sup_\theta \bigg|\mathbb{G}_n\Big[(\mathbb{1}\{t_i=1\} - \mathbb{1}\{t_i=0\})\mathbb{1}\{h(x_i,u)\le\theta\}\Big]\bigg|\mathrm{d}u, \end{aligned}$$

where equality (a) is from the monotonicity of $\widehat{m}(\theta, x)$ in θ. Taking the expectation on both sides gives us

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} \leq \int_0^{m_0} \mathbb{E}\Big[\sup_\theta \Big|\mathbb{G}_n\big[(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\})\mathbb{1}\{h(x_i, u) \leq \theta\}\big]\Big|\Big] \mathrm{d}u. \tag{E.10}$$

Applying Lemma 4 for $s(t, x) = \mathbb{1}\{t = 1\} - \mathbb{1}\{t = 0\}$ gives us for any fixed u,

$$\mathbb{E}\sup_\theta \Big|\mathbb{G}_n\big[(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\})\mathbb{1}\{h(x_i, u) \leq \theta\}\big]\Big| \leq \mathfrak{C},$$

where $\mathfrak{C}$ is a universal constant. Plugging this back to (E.10) gives us

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} \leq \mathfrak{C} m_0. \tag{E.11}$$

Using McDiarmid's inequality we have

$$\mathbb{P}(\|\mathbb{G}_n\|_{\mathcal{F}} - \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}} \geq u) \leq \exp\bigg(-\frac{2u^2}{\sum_{i=1}^n c_i^2}\bigg) \leq \exp\bigg(-\frac{2u^2}{\sum_{i=1}^n 4m_0^2/n}\bigg) = \exp(-u^2/2m_0^2), \tag{E.12}$$

where

$$\begin{aligned} c_i := \sup_{\substack{(x_i,y_i,t_i),(x_i',y_i',t_i'),\\ (x_1,y_1,t_1),\ldots,(x_n,y_n,t_n)}} \sup_\theta \sqrt{n}\bigg|&\frac{1}{n}\sum_{j=1}^n \widehat{m}(\theta, x_j)(\mathbb{1}\{t_j = 1\} - \mathbb{1}\{t_j = 0\}) \\ &- \frac{1}{n}\sum_{j=1, j\neq i}^n \widehat{m}(\theta, x_j)(\mathbb{1}\{t_j = 1\} - \mathbb{1}\{t_j = 0\}) - \frac{1}{n}\widehat{m}(\theta, x_i')(\mathbb{1}\{t_i' = 1\} - \mathbb{1}\{t_i' = 0\})\bigg| \\ \leq \sup_{(x_i,y_i,t_i),(x_i',y_i',t_i')} \sup_\theta \frac{1}{\sqrt{n}}\Big|&\widehat{m}(\theta, x_i)(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}) - \widehat{m}(\theta, x_i')(\mathbb{1}\{t_i' = 1\} - \mathbb{1}\{t_i' = 0\})\Big| \leq \frac{2m_0}{\sqrt{n}}. \end{aligned}$$

Substituting the expectation bound (E.11) in (E.12) and setting the right hand side of (E.12) to δ yields for another absolute constant $\mathfrak{C}'$,

$$\mathbb{P}\bigg(\|\mathbb{G}_n\|_{\mathcal{F}} \geq \mathfrak{C}' m_0 \sqrt{1 + \log\big(\tfrac{1}{\delta}\big)}\bigg) \leq \mathbb{P}\bigg[\|\mathbb{G}_n\|_{\mathcal{F}} \geq \mathfrak{C} m_0 \bigg\{1 + \sqrt{2\log\big(\tfrac{1}{\delta}\big)}\bigg\}\bigg] \leq \delta. \tag{E.13}$$

Bound on $\mathcal{R}_3$: Because the random variables in the averaging of $\mathcal{R}_3$ are i.i.d, applying Hoeffding's inequality yields

$$\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{t_i = 1\} - \mathbb{P}(T = 1) \geq t\bigg\} \leq \exp\bigg(-\frac{2t^2}{n}\bigg).$$

And this leads to

$$\mathbb{P}\bigg(R_3 \geq (1 - \alpha)\sqrt{\frac{1}{2n}\log\big(\tfrac{1}{\delta}\big)}\bigg) \leq \delta. \tag{E.14}$$

Combining (E.9), (E.13) and (E.14) together using the union bound gives the result that for a universal constant $\mathfrak{C}$,

$$\mathbb{P}\bigg\{\sup_\theta \big|\mathcal{R}_1(\theta) + \mathcal{R}_2(\theta) + \mathcal{R}_3\big| \geq \mathfrak{C}\sqrt{\frac{(m_0 + \pi_0 + 1 - \alpha)^2 \log\big(\frac{1}{\delta}\big) + (m_0 + \pi_0)^2}{n}}\bigg\} \leq \delta.$$

□

S.7 Proof of Theorem 5

Proof. Without loss of generality assume the indexes in $\mathcal{D}^{\text{tr}}$ is $1,\dots,N$ with $N := |\mathcal{D}^{\text{tr}}|$, and we expand $\mathbb{P}_N\big[\text{IF}(\theta,\mathcal{D}^{\text{tr}};\widehat{\pi},\widehat{m})\big] - P\big[\text{IF}(\theta,X,R,T,\widehat{\pi},\widehat{m})\big]$ into three parts,

$$\begin{aligned}
&\mathbb{P}_N\big[\text{IF}(\theta,\mathcal{D}^{\text{tr}};\widehat{\pi},\widehat{m})\big] - P\big[\text{IF}(\theta,X,R,T;\widehat{\pi},\widehat{m})\big] \\
=\ &\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{t_i=0\}\widehat{\pi}(x_i)\mathbb{1}\{r_i\le\theta\} - \mathbb{E}\big[\mathbb{1}\{t_i=0\}\widehat{\pi}(x_i)\mathbb{1}\{r_i\le\theta\}\big] \\
&+\frac{1}{N}\sum_{i=1}^{N}\widehat{m}(\theta,x_i)\big(\mathbb{1}\{t_i=1\}-\mathbb{1}\{t_i=0\}\big) - \mathbb{E}\big[\widehat{m}(\theta,x_i)(\mathbb{1}\{t_i=1\}-\mathbb{1}\{t_i=0\})\big] \\
&-(1-\alpha)\bigg[\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{t_i=1\} - \mathbb{P}(T=1)\bigg] \\
=:\ &\mathcal{R}_1(\widehat{\pi},\theta)+\mathcal{R}_2(\widehat{m},\theta)+\mathcal{R}_3,
\end{aligned}$$

where the three terms will be controlled separately. In particular, $\sup_\theta \mathcal{R}_1(\widehat{\pi},\theta)$ and $\sup_\theta \mathcal{R}_2(\widehat{m},\theta)$ will be bounded using tools from empirical processes.

Bound on $\sup_\theta \mathcal{R}_1(\widehat{\pi},\theta)$**:** We have a class of functions $\mathcal{F} = \big\{f : f_{\widehat{\pi},\theta}(w) = \mathbb{1}\{t=1\}\widehat{\pi}(x)\mathbb{1}\{R(x,y)\le\theta\}, \forall\widehat{\pi}(\cdot)\in\mathcal{F}_\pi, \theta\in\mathbb{R}\big\}$. Notice that $\forall\theta\in\mathbb{R}$ and $\widehat{\pi}(\cdot)\in\mathcal{F}_\pi$, we have $|f_\theta(w)|\le\pi_0\mathbb{1}\{t=1\}$. Therefore, $F(w) := \pi_0\mathbb{1}\{t=1\}$ is an envelope function of $\{f_{\widehat{\pi},\theta}(\cdot) : \widehat{\pi}(\cdot)\in\mathcal{F}_\pi, \theta\in\mathbb{R}\}$.

Note that because $\widehat{\pi}(\cdot)$ depends on the evaluation data, $\mathcal{R}_1(\widehat{\pi},\theta)$ is not an average of N i.i.d random variables. However we can still bound this term by $\sup_{f\in\mathcal{F}}|\mathbb{G}_N f|$ where for any fixed $f\in\mathcal{F}, \mathbb{G}_N f/\sqrt{N}$ is an average of N i.i.d. random variables with mean 0.

For any pair of functions $f_{\widehat{\pi}_1,\theta_1}, f_{\widehat{\pi}_2,\theta_2}\in\mathcal{F}$,

$$\begin{aligned}
\|f_{\widehat{\pi}_1,\theta_1} - f_{\widehat{\pi}_2,\theta_2}\|_Q &= \Big[\sum_{i=1}^{N}\mathbb{1}^2\{t_i=1\}\big(\widehat{\pi}_1(x_i)\mathbb{1}\{r_i\le\theta_1\}-\widehat{\pi}_2(x_i)\mathbb{1}\{r_i\le\theta_2\}\big)^2 Q(w_i)\Big]^{1/2} \\
&= \frac{\Big[\sum_{i=1}^{N}\mathbb{1}\{t_i=1\}\big(\widehat{\pi}_1(x_i)\mathbb{1}\{r_i\le\theta_1\}-\widehat{\pi}_2(x_i)\mathbb{1}\{r_i\le\theta_2\}\big)^2 Q(w_i)\Big]^{1/2}}{\big[\sum_{i=1}^{N}\mathbb{1}^2\{t_i=1\}\ Q(w_i)\big]^{1/2}}\times \\
&\qquad\qquad \big[\sum_{i=1}^{N}\mathbb{1}^2\{t_i=1\}]Q(w_i)\big]^{1/2} \\
&= \big\|\widehat{\pi}_1(x_i)\mathbb{1}\{r_i\le\theta_1\}-\widehat{\pi}_2(x_i)\mathbb{1}\{r_i\le\theta_2\}\big\|_{\widetilde{Q}}\cdot\big[\sum_{i=1}^{N}\mathbb{1}\{t_i=1\}Q(w_i)\big]^{1/2}, \qquad \text{(E.15)}
\end{aligned}$$

where $\widetilde{Q}$ is the new probability measure defined by

$$\widetilde{Q}(w_i) := \mathbb{1}\{t_i=1\}Q(w_i)/\sum_{i=1}^{N}\mathbb{1}\{t_i=1\}Q(w_i), i=1,\dots,N. \qquad \text{(E.16)}$$

Using $F(w)=\pi_0\mathbb{1}\{t=1\}$ as the envelope function of the class $\mathcal{F}$, (E.15) becomes

$$\|f_{\widehat{\pi}_1,\theta_1} - f_{\widehat{\pi}_2,\theta_2}\|_Q = \big\|\widehat{\pi}_1(x_i)\mathbb{1}\{r_i\le\theta_1\}-\widehat{\pi}_2(x_i)\mathbb{1}\{r_i\le\theta_2\}\big\|_{\widetilde{Q}}\cdot\|F\|_Q/\pi_0. \qquad \text{(E.17)}$$

Define a new class of functions $\mathcal{F}_1 := \big\{f : f_{\widehat{\pi},\theta}(z) = \widehat{\pi}(x)\mathbb{1}\{R(x,y)\le\theta\}\big\}$. Then (E.17) gives the relationship between the covering numbers of the two classes $\mathcal{F}$ and $\mathcal{F}_1$,

$$N\big(\varepsilon\|F\|_Q,\mathcal{F},L_2(Q)\big) \le N\big(\pi_0\varepsilon,\mathcal{F}_1,L_2(\widetilde{Q})\big). \qquad \text{(E.18)}$$

We take the envelope function of $\mathcal{F}_1$ to be $F_1 \equiv \pi_0$ and use Theorem 2.6.7 of van der Vaart and Wellner (1996) with $r = 2$ to get a bound on the covering number bound for $\mathcal{G} = \{g_\theta : g_\theta(z) = \mathbb{1}\{R(x,y) \le \theta\}, \forall \theta \in \mathbb{R}\}$. Specifically, for any probability measure Q, there exists a universal constant $\mathfrak{C}$ such that

$$N\big(\varepsilon, \mathcal{G}, L_2(Q)\big) \le \frac{\mathfrak{C}}{\varepsilon}.$$

Let $\mathcal{H} := \{h : h = \widehat{\pi}(x), \forall \widehat{\pi} \in \mathcal{F}_\pi\}$. The relationship between covering and bracketing numbers gives us

$$N(\varepsilon, \mathcal{H}, L_2(Q)) \le N_{[]}(2\varepsilon, \mathcal{H}, L_2(Q)) \le \exp\big(C(2\pi_0\varepsilon)^{-\alpha_\pi}\big).$$

Applying Lemma 5 with $\mathcal{F}_1 = \mathcal{G}$, $\mathcal{F}_2 = \mathcal{H}$ and $C_1 = 1, C_2 = \pi_0$ gives

$$N\big(\pi_0\varepsilon, \mathcal{F}_1, L_2(Q)\big) \le N\big(\varepsilon/2, \mathcal{G}, L_2(Q)\big) N\big(\pi_0\varepsilon/2, \mathcal{H}, L_2(Q)\big) \le \frac{2\mathfrak{C}}{\varepsilon} \exp\big(C(\pi_0\varepsilon)^{-\alpha_\pi}\big).$$

Therefore, it holds for some constant $\mathfrak{C}'$ such that,

$$\log N\big(\varepsilon \|F_1\|_Q, \mathcal{F}_1, L_2(Q)\big) \le C(\pi_0\varepsilon)^{-\alpha_\pi} - \log(\varepsilon) + \mathfrak{C} \le \mathfrak{C}'(\pi_0\varepsilon)^{-\alpha_\pi}. \tag{E.19}$$

Because (E.19) holds for any probability measure Q, we choose Q to be $\widetilde{Q}$ defined in (E.16) to bound the right hand side of (E.18),

$$\log N\big(\varepsilon \|F\|_Q, \mathcal{F}, L_2(Q)\big) \le \log N\big(\pi_0\varepsilon, \mathcal{F}_1, L_2(\widetilde{Q})\big) \le \mathfrak{C}'(\pi_0\varepsilon)^{-\alpha_\pi}.$$

We apply Lemma A.1 of Srebro et al. (2012) so that

$$\begin{aligned}
\mathbb{E}\|\mathbb{G}_N\|_{\mathcal{F}}/\pi_0 &\lesssim \inf_\eta \left\{ N^{1/2}\eta + \int_\eta^1 \log_+^{1/2} N_{[]}\left(\pi_0\varepsilon, \mathcal{F}, \|\cdot\|_{L_2(P)}\right) \mathrm{d}\varepsilon \right\} + N^{-1/2} \log_+ N_{[]}\left(\pi_0, \mathcal{F}, \|\cdot\|_{L_2(P)}\right) \\
&\le \inf_\eta \left\{ N^{1/2}\eta + \mathfrak{C} \int_\eta^1 (\pi_0\varepsilon)^{-\alpha_\pi/2} \mathrm{d}\varepsilon \right\} + N^{-1/2}\mathfrak{C}\pi_0^{-\alpha_\pi} \\
&\le \frac{1}{\pi_0} N^{1/2-1/\alpha_\pi}\big(1 + \mathbb{1}\{\alpha_\pi = 2\}\log N\big) + N^{-1/2}\mathfrak{C}\pi_0^{-\alpha_\pi} \\
&\le \frac{\mathfrak{C}}{\pi_0} N^{(1/2-1/\alpha_\pi)_+}\big(1 + \mathbb{1}\{\alpha_\pi = 2\}\log N\big), \text{ when } N \text{ is large.}
\end{aligned} \tag{E.20}$$

Using McDiarmid's inequality we have

$$\mathbb{P}\big(\|\mathbb{G}_N\|_{\mathcal{F}} - \mathbb{E}\|\mathbb{G}_N\|_{\mathcal{F}} \ge u\big) \le \exp\left(-\frac{2u^2}{\sum_{i=1}^N c_i^2}\right) \le \exp\left(-\frac{2u^2}{\sum_{i=1}^N 4\pi_0^2/N}\right) = \exp(-u^2/2\pi_0^2), \tag{E.21}$$

where

$$\begin{aligned}
c_i := \sup_{\substack{(x_i,y_i,t_i),(x_i',y_i',t_i'),\\ (x_1,y_1,t_1),\ldots,(x_N,y_N,t_N)}} \sup_{\widehat{\pi},\theta} \sqrt{N} \Bigg| & \frac{1}{N}\sum_{j=1}^N \mathbb{1}\{t_j = 0\}\widehat{\pi}(x_j)\mathbb{1}\{r_j \le \theta\} - \frac{1}{N}\sum_{j=1, j\ne i}^N \mathbb{1}\{t_j = 0\}\widehat{\pi}(x_j)\mathbb{1}\{r_j \le \theta\} \\
& - \frac{1}{N}\mathbb{1}\{t_i' = 0\}\widehat{\pi}(x_i')\mathbb{1}\{r_i' \le \theta\}\Bigg| \\
\le \sup_{(x_i,y_i,t_i),(x_i',y_i',t_i')} \sup_{\widehat{\pi},\theta} \sqrt{N} & \left| \frac{1}{N}\mathbb{1}\{t_i = 0\}\widehat{\pi}(x_i)\mathbb{1}\{r_i \le \theta\} - \frac{1}{N}\mathbb{1}\{t_i' = 0\}\widehat{\pi}(x_i')\mathbb{1}\{r_i' \le \theta\}\right| \le \frac{2\pi_0}{\sqrt{N}}.
\end{aligned}$$

Substituting the expectation bound (E.20) in (E.21) and setting the right hand side of (E.21) to δ yields for another absolute constant $\mathfrak{C}'$,

$$\mathbb{P}\left[\|\mathbb{G}_n\|_{\mathcal{F}} \ge \mathfrak{C}'\left\{N^{(1/2-1/\alpha_\pi)_+}\big(1 + \mathbb{1}\{\alpha_\pi = 2\}\log N\big) + \pi_0\sqrt{2\log\big(\tfrac{1}{\delta}\big)}\right\}\right] \le \delta. \tag{E.22}$$

Bound on $\sup_\theta \mathcal{R}_2(\widehat{m},\theta)$**:** We define the class of functions $\mathcal{F} = \{f : f_{\widehat{m},\theta}(w) = \widehat{m}(\theta, x_i)(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}, \forall \widehat{m}(\cdot,\cdot) \in \mathcal{F}_m \text{ and } \theta \in \mathbb{R}\}$ with $F(w) = m_0(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\})$ as its envelope. Similar as before we bound this term by $\sup_{f\in\mathcal{F}} |\mathbb{G}_N f|$ and for any fixed $f \in \mathcal{F}, |\mathbb{G}_N f|/\sqrt{N}$ is an average of i.i.d. random variables.

For any pair of functions $f_{\widehat{m}_1,\theta_1}, f_{\widehat{m}_2,\theta_2} \in \mathcal{F}$,

$$\begin{aligned}
\|f_{\widehat{m}_1,\theta_1} - f_{\widehat{m}_2,\theta_2}\|_Q &= \Big[\sum_{i=1}^N \big(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}\big)^2 \big(\widehat{m}_1(\theta_1, x_i) - \widehat{m}_2(\theta_2, x_i)\big)^2 Q(w_i)\Big]^{1/2} \\
&= \frac{\Big[\sum_{i=1}^N \big(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}\big)^2 \big(\widehat{m}_1(\theta_1, x_i) - \widehat{m}_2(\theta_2, x_i)\big)^2 Q(w_i)\Big]^{1/2}}{\big[\sum_{i=1}^N \big(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}\big)^2 Q(w_i)\big]^{1/2}} \times \\
&\qquad\qquad \big[\sum_{i=1}^N \big(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}\big)^2 Q(w_i)\big]^{1/2} \\
&= \big\|\big(\widehat{m}_1(\theta_1, x_i) - \widehat{m}_2(\theta_2, x_i)\big\|_{\widetilde{Q}} \cdot \big[\sum_{i=1}^n \big(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}\big)^2 Q(w_i)\big]^{1/2}, \qquad \text{(E.23)}
\end{aligned}$$

where $\widetilde{Q}$ is the new probability measure defined by

$$\widetilde{Q}(w_i) := \big(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}\big)^2 Q(w_i) / \sum_{i=1}^N \big(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}\big)^2 Q(w_i), i = 1, \ldots, N.$$

Using $F(w) = m_0(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\})$ as the envelope function of the class $\mathcal{F}$, (E.23) becomes

$$\|f_{\widehat{m}_1,\theta_1} - f_{\widehat{m}_2,\theta_2}\|_Q = \big\|\widehat{m}_1(\theta_1, x_i) - \widehat{m}_2(\theta_2, x_i)\big\|_{\widetilde{Q}} \cdot \|F\|_Q / m_0. \qquad \text{(E.24)}$$

Recall the class of functions $\mathcal{F}_m := \big\{f : f_{\widehat{m},\theta}(z) = \widehat{m}(\theta, x)\big\}$ with its envelope $F_1 \equiv m_0$. Then (E.24) gives the relationship between the covering numbers of the two classes $\mathcal{F}$ and $\mathcal{F}_m$. Along with the bracketing number assumption on $\mathcal{F}_m$ it follows that

$$\log N\big(\varepsilon\|F\|_Q, \mathcal{F}, L_2(Q)\big) \le \log N\big(m_0\varepsilon, \mathcal{F}_m, L_2(\widetilde{Q})\big) \le \log N_{[]}\big(2m_0\varepsilon, \mathcal{F}_m, L_2(\widetilde{Q})\big) \le C(2m_0\varepsilon)^{-\alpha_m}.$$

Applying the same technique as we used in (E.20) when bounding $\sup_\theta \mathcal{R}_1$ gives us $\forall \alpha_m \ge 0$,

$$\mathbb{E}\|\mathbb{G}_N\|_{\mathcal{F}} \le \mathfrak{C} N^{(1/2-1/\alpha_m)_+}\big(1 + \mathbb{1}\{\alpha_m = 2\}\log N\big). \qquad \text{(E.25)}$$

Using McDiarmid's inequality we have

$$\mathbb{P}(\|\mathbb{G}_N\|_{\mathcal{F}} - \mathbb{E}\|\mathbb{G}_N\|_{\mathcal{F}} \ge u) \le \exp\Big(-\frac{2u^2}{\sum_{i=1}^N c_i^2}\Big) \le \exp\Big(-\frac{2u^2}{\sum_{i=1}^N 4m_0^2/N}\Big) = \exp(-u^2/2m_0^2), \qquad \text{(E.26)}$$

where

$$\begin{aligned}
c_i := \sup_{\substack{(x_i,y_i,t_i),(x_i',y_i',t_i'),\\ (x_1,y_1,t_1),\ldots,(x_N,y_N,t_N)}} \sup_{\widehat{m},\theta} \sqrt{N}\Big|&\frac{1}{n}\sum_{j=1}^N \widehat{m}(\theta, x_j)(\mathbb{1}\{t_j = 1\} - \mathbb{1}\{t_j = 0\}) \\
&- \frac{1}{N}\sum_{j=1, j\ne i}^N \widehat{m}(\theta, x_j)(\mathbb{1}\{t_j = 1\} - \mathbb{1}\{t_j = 0\}) - \frac{1}{N}\widehat{m}(\theta, x_i')(\mathbb{1}\{t_i' = 1\} - \mathbb{1}\{t_i' = 0\})\Big| \\
\le \sup_{(x_i,y_i,t_i),(x_i',y_i',t_i')} \sup_{\widehat{m},\theta} \frac{1}{\sqrt{N}}\Big|&\widehat{m}(\theta, x_i)(\mathbb{1}\{t_i = 1\} - \mathbb{1}\{t_i = 0\}) - \widehat{m}(\theta, x_i')(\mathbb{1}\{t_i' = 1\} - \mathbb{1}\{t_i' = 0\})\Big| \le \frac{2m_0}{\sqrt{N}}.
\end{aligned}$$

Substituting the expectation bound (E.25) in (E.26) and setting the right hand side of (E.26) to δ yields for another absolute constant $\mathfrak{C}'$,

$$\mathbb{P}\Big[\|\mathbb{G}_N\|_{\mathcal{F}} \ge \mathfrak{C}\Big\{N^{(1/2-1/\alpha_m)_+}\big(1 + \mathbb{1}\{\alpha_m = 2\}\log N\big) + m_0\sqrt{2\log\big(\frac{1}{\delta}\big)}\Big\}\Big] \le \delta. \qquad \text{(E.27)}$$

Bound on $\mathcal{R}_3$: Because the random variables in the averaging of $\mathcal{R}_3$ are i.i.d, applying Hoeffding's inequality yields

$$\mathbb{P}\bigg\{\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{t_i=1\}-\mathbb{P}(T=1)\geq t\bigg\}\leq\exp\bigg(-\frac{2t^2}{N}\bigg).$$

And this leads to

$$\mathbb{P}\bigg(R_3\geq(1-\alpha)\sqrt{\frac{1}{2N}\log\big(\frac{1}{\delta}\big)}\bigg)\leq\delta. \tag{E.28}$$

Combining (E.22), (E.27) and (E.28) together using the union bound gives the result that for a universal constant $\mathfrak{C}$,

$$\mathbb{P}\bigg\{\sup_{\theta}\big(\mathcal{R}_1(\theta)+\mathcal{R}_2(\theta)+\mathcal{R}_3\big)\geq\mathfrak{C}\bigg(N^{-1/(\alpha_m\vee 2)}\big(1+\mathbb{1}\{\alpha_m=2\}\log N\big)+N^{-1/(\alpha_\pi\vee 2)}\big(1+\mathbb{1}\{\alpha_\pi=2\}\log N\big)$$
$$+\sqrt{\frac{(m_0+\pi_0+1-\alpha)^2\log\big(\frac{1}{\delta}\big)}{N}}\bigg)\bigg\}\leq\delta.$$

□

S.8 Proof of Theorem 1

The first half of the Theorem, specifically (11), is adapted from Lemma S1 of Qiu et al. (2022). We modify that proof a bit in our setting in the following and first state a Lemma that is adapted from Theorem 2 and Remark 4 from Shah and Peters (2020), which will contribute to the proof of Theorem 1.

Lemma 3. *(No-free-lunch for conditional independence testing). Assume that X is continuous, given any $n\in\mathbb{N},\alpha\in(0,1),M\in(0,\infty]$, and any potentially randomised test η_n that has valid level α for the null hypothesis $\mathcal{P}_{0,M}$, we have that $\mathbb{P}_Q(\eta_n=1)\leq\alpha$ for all $Q\in\mathcal{Q}_{0,M}$. Thus η_n cannot have power against any alternative.*

Let $\bar{\mathcal{E}}\supseteq\bar{\mathcal{P}}^0$ be the space of distributions for the full data point $\bar{O}$ with the distribution of $(X,Y)\mid T=0,1$ both dominated by the Lebesgue measure. These distributions may or may not satisfy the MAR assumption that $Y\perp T\mid X$. For any $x\in\mathcal{X}$, the prediction set $\widehat{C}(x)$ is the short hand notation for $C(x;O_1,\ldots,O_n)$. This notation is helpful to clarify the dependence of $\widehat{C}$ on the observed training data $(O_1,\ldots,O_n)$. Let $\bar{O}_{n+1}$ to denote the full data point from a future draw.

The subtle difference between $\bar{O}$ and O is that $\bar{O}$ represents the full but unobserved data (X,Y,T) whereas O corresponds to the observed data $(X,T,(1-T)Y)$.

Define the randomized test $\eta\left(\bar{O}_1,\ldots,\bar{O}_{n+1}\right)$ as follows:

$$\eta\left(\bar{O}_1,\ldots,\bar{O}_{n+1}\right)=\begin{cases}\mathbb{1}\big\{Y_{n+1}\notin C(X_{n+1};O_1,\ldots,O_n)\big\}, & \text{if } T_{n+1}=1;\\ \begin{cases}1, \text{ w.p. } \alpha,\\ 0, \text{ w.p. } 1-\alpha,\end{cases} & \text{if } T_{n+1}=0.\end{cases}$$

We note that although η is a function of $n+1$ full data points $\left(\bar{O}_1,\ldots,\bar{O}_{n+1}\right)$, it only relies on n observed training data points $(O_1,\ldots,O_n)$ and one future full data point $\bar{O}_{n+1}$. Furthermore, we can see that

$$\begin{aligned}\mathbb{P}_{\bar{P}^0}\left(\eta(\bar{O}_1,\ldots,\bar{O}_{n+1})=1\right)&=\mathbb{P}_{\bar{P}^0}\left(\eta(\bar{O}_1,\ldots,\bar{O}_{n+1})=1|T_{n+1}=1\right)\mathbb{P}(T_{n+1}=1)\\&\quad+\mathbb{P}_{\bar{P}^0}\left(\eta(\bar{O}_1,\ldots,\bar{O}_{n+1})=1|T_{n+1}=1\right)\mathbb{P}(T_{n+1}=1)\\&=\mathbb{P}_{\bar{P}^0}\big(Y\notin\widehat{C}(X)|T=1\big)\mathbb{P}(T=1)+\mathbb{P}_{\bar{P}^0}\big(\eta=1|T=0\big)\mathbb{P}(T=0)\\&\leq\alpha\mathbb{P}(T=1)+\alpha\mathbb{P}(T=0),\text{ by (10) and the definition of }\eta\\&\leq\alpha.\end{aligned}$$

Therefore, η can be viewed as a test with level α for the null hypothesis $Y \perp T \mid X$. Because Theorem 3 states that the power of η against the alternative hypothesis is at most α, we have that

$$\mathbb{P}_{\bar{Q}}\left(\eta\left(\bar{O}_1, \ldots, \bar{O}_{n+1}\right) = 1\right) \leq \alpha \quad \text{for any distribution } \bar{Q} \in \bar{\mathcal{E}}.$$

Using the fact that η is a completely randomized test when when $T_{n+1}) = 0$, this gives $\mathbb{P}(\eta = 1 | T_{n+1} = 0) = \alpha$. Therefore,

$$\begin{aligned}\alpha \geq \mathbb{P}_{\bar{Q}}\left(\eta\left(\bar{O}_1, \ldots, \bar{O}_{n+1}\right) = 1\right) &= \mathbb{P}_{\bar{Q}}(\eta = 1 | T_{n+1} = 0)\mathbb{P}(T_{n+1} = 0) + \mathbb{P}_{\bar{Q}}(\eta = 1 | T_{n+1} = 1)\mathbb{P}(T_{n+1} = 1) \\ &= \alpha\mathbb{P}(T_{n+1} = 0) + \mathbb{P}_{\bar{Q}}(\eta = 1 | T_{n+1} = 1)\mathbb{P}(T_{n+1} = 1).\end{aligned}$$

Therefore,

$$\alpha \geq \mathbb{P}_{\bar{Q}}(\eta = 1 | T_{n+1} = 1) = \mathbb{P}_{\bar{Q}}\big(Y_{n+1} \notin C(X_{n+1}; O_1, \ldots, O_n) \mid T_{n+1} = 1\big). \tag{E.29}$$

For any $x \in \mathcal{X}$, let $D_x \subseteq \mathcal{Y}$ be any Lebesgue measurable set with nonzero finite measure and U_x be the uniform distribution on D_x. Take $\bar{Q}$ to be a distribution such that

(i) the distribution of T is an arbitrary Bernoulli distribution with success probability in $(0, 1)$;

(ii) the distributions of $X \mid T = 0, 1$ satisfy the dominance condition (8), which is that the source distribution $(X|T = 0)$ dominates the target distribution $(X|T = 1)$, and are arbitrary in all other aspects

(iii) the distribution of $Y \mid X = x, T = 0$ is arbitrary and the distribution of $Y \mid X = x, T = 1$ is U_x.

We take $\bar{P}^0$ to be the distribution that is identical to $\bar{Q}$, except that the distribution of $Y \mid X = x, T = 1$ is identical to $Y \mid X = x, T = 0$ rather than U_x under $\bar{P}^0$. Note that $\bar{P}^0 \in \bar{\mathcal{P}}^0$. Since $\widehat{C}$ is trained only on observed data $(O_1, \ldots, O_n)$ and $\bar{P}^0$ and $\bar{Q}$ imply the same distribution of the observed data point $O = (X, T, (1 - T)Y)$, following (E.29), we have that

$$\begin{aligned}&\mathbb{P}_{\bar{Q}}\big(Y_{n+1} \notin C(X_{n+1}; O_1, \ldots, O_n) \mid T_{n+1} = 1\big) \\ &= \int_y \mathbb{P}_{\bar{P}_0}\big(y \notin C(X_{n+1}; O_1, \ldots, O_n) \mid T_{n+1} = 1\big) U_{X_{n+1}}(\mathrm{d}y) \leq \alpha,\end{aligned}$$

where the probability is over training data and possible exogenous randomness in C. Since D_x is arbitrary, it follows that the integrand is bounded by α, namely

$$\mathbb{P}_{\bar{P}_0}\Big(y \notin C(X_{n+1}; O_1, \ldots, O_n) \mid T_{n+1} = 1\Big) \leq \alpha \tag{E.30}$$

for a.e. $y \in \mathcal{Y}$. The desired result (11) follows by replacing the notations X_{n+1} and T_{n+1} with X and T, respectively, and noting that $C\left(x; O_1, \ldots, O_n\right) = \widehat{C}(x)$ by definition.

Note that (E.30) implies

$$\inf_{y \in \mathcal{Y}} \mathbb{P}_{\bar{P}_0}(y \in \widehat{C}(X) | T = 1) \geq 1 - \alpha,$$

for any prediction set $\widehat{C}$ that satisfies (10). Define $L(X) = \inf\{t : t \in \widehat{C}(X)\}$ and $U(X) = \sup\{t : t \in \widehat{C}(X)\}$. Because $\widehat{C}(X) \subseteq [L(X), \infty)$, we get

$$\mathbb{P}_{\bar{P}_0}\Big(\inf_{y \in \mathcal{Y}} y \geq L(X) \;\Big|\; T = 1\Big) = \inf_{y \in \mathcal{Y}} \mathbb{P}_{\bar{P}_0}(y \geq L(X) \;\big|\; T = 1) = \inf_{y \in \mathcal{Y}} \mathbb{P}_{\bar{P}_0}\big(y \in [L(X), \infty) \;\big|\; T = 1\big) \geq 1 - \alpha.$$

Similarly, because $\widehat{C}(X) \subseteq (-\infty, U(X)]$,

$$\mathbb{P}_{\bar{P}_0}\Big(U(X) \geq \sup_{y \in \mathcal{Y}} y \;\Big|\; T = 1\Big) = \inf_{y \in \mathcal{Y}} \mathbb{P}_{\bar{P}_0}\big(y \leq U(X) \;\big|\; T = 1\big) = \inf_{y \in \mathcal{Y}} \mathbb{P}_{\bar{P}_0}\big(y \in (-\infty, U(X)] \;\big|\; T = 1\big) \geq 1 - \alpha.$$

Hence with $\mathcal{Y} = \mathbb{R}$ here, then any prediction set with valid coverage must have at least one of the end points ∞ in absolute value with probability at least $1 - \alpha$ and hence must have an infinite diameter with probability at least $1 - \alpha$.

S.9 Some useful propositions and lemmas

Proposition 1. *(Chernozhukov et al. (2009), Proposition 2) Let the target function $f_0 : \mathcal{X}^d \to K$ be weakly increasing and measurable in x. Let $\widehat{f} : \mathcal{X}^d \to K$ be a measurable function that is an initial estimate of f_0.*

1. *For each ordering π of $1, \ldots, d$, the π-rearranged estimate $\widehat{f}^*_\pi$ is weakly increasing. Moreover, $\widehat{f}^*$, an average of π-rearranged estimates, is weakly increasing.*
2. *A π-rearranged estimate $\widehat{f}^*_\pi$ of $\widehat{f}$ weakly reduces the estimation error of $\widehat{f}$:*

$$\left\{\int_{\mathcal{X}^d} \left|\widehat{f}^*_\pi(x) - f_0(x)\right|^p dx\right\}^{1/p} \leq \left\{\int_{\mathcal{X}^d} \left|\widehat{f}(x) - f_0(x)\right|^p \mathrm{d}x\right\}^{1/p}.$$

S.9.1 Lemma 4 and its proof

Lemma 4. *There exists a universal constant $\mathfrak{C} < \infty$ such that for any functions $s(t,x) \in [-\kappa_0, \kappa_0]$ and $h(x,y)$,*

$$\mathbb{E}\Big[\sup_\theta |\mathbb{G}_n[s(t,x)\mathbb{1}\{h(x,y) \leq \theta\}]|\Big] \leq \mathfrak{C}\kappa_0.$$

Proof. We have a class of functions $\mathcal{F} = \big\{f : f_\theta(w) = s(t,x)\mathbb{1}\{h(x,y) \leq \theta\}\big\}$. Notice that $\forall \theta \in \mathbb{R}$, we have $|f_\theta(w)| \leq |s(t,x)|$. Therefore, $F(w) := |s(t,x)|$ is an envelope function of $\{f_\theta(\cdot) : \theta \in \mathbb{R}\}$. For any discrete probability measure Q, let $\|f(\cdot)\|_Q$ denote the empirical $L_2(Q)$ norm where $\|f(\cdot)\|_Q := \big(\sum_{i=1}^n f^2(x_i)Q(x_i)\big)^{1/2}$.

Let h_i denote $h_i := h(x_i, y_i)$. For any function $f \in \mathcal{F}$ and $\theta_1, \theta_2 \in \mathbb{R}$,

$$\begin{aligned}
\|f_{\theta_1} - f_{\theta_2}\|_Q &= \Big[\sum_{i=1}^n s(t_i,x_i)^2\big(\mathbb{1}\{h_i \leq \theta_1\} - \mathbb{1}\{h_i \leq \theta_2\}\big)^2 Q(w_i)\Big]^{1/2} \\
&= \frac{\Big[\sum_{i=1}^n s(t_i,x_i)^2\big(\mathbb{1}\{h_i \leq \theta_1\} - \mathbb{1}\{h_i \leq \theta_2\}\big)^2 Q(w_i)\Big]^{1/2}}{\big[\sum_{i=1}^n s(t_i,x_i)^2 Q(w_i)\big]^{1/2}} \times \\
&\qquad\qquad \big[\sum_{i=1}^n s(t_i,x_i)^2 Q(w_i)\big]^{1/2} \\
&= \big\|\mathbb{1}\{h_i \leq \theta_1\} - \mathbb{1}\{h_i \leq \theta_2\}\big\|_{\widetilde{Q}} \cdot \big[\sum_{i=1}^n s(t_i,x_i)^2 Q(w_i)\big]^{1/2}, \qquad \text{(E.31)}
\end{aligned}$$

where $\widetilde{Q}$ is the new probability measure defined by

$$\widetilde{Q}(w_i) := s(t_i,x_i)^2 Q(w_i) / \sum_{i=1}^n s(t_i,x_i)^2 Q(w_i). \qquad \text{(E.32)}$$

Using the definition of $F(w)$ as the envelope function of the class $\mathcal{F}$, (E.31) becomes

$$\|f_{\theta_1} - f_{\theta_2}\|_Q = \big\|\mathbb{1}\{h_i \leq \theta_1\} - \mathbb{1}\{h_i \leq \theta_2\}\big\|_{\widetilde{Q}} \cdot \|F\|_Q. \qquad \text{(E.33)}$$

Define a new class of functions $\mathcal{F}_1 := \big\{f : f_\theta(z) = \mathbb{1}\{h(x,y) \leq \theta\}\big\}$ with its envelope function $F_1 \equiv 1$. Then (E.33) gives a relationship between the covering numbers of the two classes $\mathcal{F}$ and $\mathcal{F}_2$,

$$N\big(\varepsilon\|F\|_Q, \mathcal{F}, L_2(Q)\big) \leq N\big(\varepsilon, \mathcal{F}_1, L_2(\widetilde{Q})\big). \qquad \text{(E.34)}$$

From Chapter 2.6 of van der Vaart and Wellner (1996), we know that the VC dimension of the class $\mathcal{F}_1$ is 1. We use Theorem 2.6.7 of van der Vaart and Wellner (1996) with $r = 2$ to get a bound on the covering numbers. Specifically, for any probability measure Q, there exists a universal constant $\mathfrak{C}$ such that

$$N\big(\varepsilon, \mathcal{F}_1, L_2(Q)\big) = N\big(\varepsilon \|F_1\|_Q, \mathcal{F}, L_2(Q)\big) \leq \frac{\mathfrak{C}}{\varepsilon}. \tag{E.35}$$

Because (E.35) holds for any probability measure Q, we choose Q to be $\widetilde{Q}$ defined in (E.32) in order to bound the right hand side of (E.34),

$$N\big(\varepsilon \|F\|_Q, \mathcal{F}, L_2(Q)\big) \leq N\big(\varepsilon, \mathcal{F}_1, L_2(\widetilde{Q})\big) \leq \frac{\mathfrak{C}}{\varepsilon}. \tag{E.36}$$

Next, we obtain an upper bound of the uniform-entropy

$$J(\delta, \mathcal{F}) := \sup_Q \int_0^\delta \sqrt{1 + \log N\big(\varepsilon \|F\|_Q, \mathcal{F}, L_2(Q)\big)} d\varepsilon,$$

where the supremum is taken over all discrete probability measures Q with $\|F\|_Q > 0$. Applying $p = 1$ to Theorem 2.14.1 of van der Vaart and Wellner (1996) gives us

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\mathbb{G}_n f| \;\leq\; J(1, \mathcal{F}) \|F\|_Q \;\lesssim\; \|F\|_Q \int_0^1 \mathfrak{C}^{1/2} \sqrt{\log(1/\varepsilon)} d\varepsilon \;\leq\; \mathfrak{C}_1 \kappa_0,$$

where the second inequality is from (E.36) and $\mathfrak{C}_1$ is a universal constant. □

S.9.2 Lemma 5 and its proof

Lemma 5. *For any two bounded function classes $\mathcal{F}_1$ and $\mathcal{F}_2$ with $\sup_{f_1 \in \mathcal{F}_1} \|f_1\|_\infty \leq C_1$ and $\sup_{f_2 \in \mathcal{F}_2} \|f_2\|_\infty \leq C_2$, the following holds for the covering number of the class of functions $\mathcal{F}_1 \times \mathcal{F}_2 := \{f = f_1 f_2, f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2\}$,*

$$N\big(\varepsilon, \mathcal{F}_1 \times \mathcal{F}_2, L_2(Q)\big) \leq N\big(\frac{\varepsilon}{2C_2}, \mathcal{F}_1, L_2(Q)\big) N\big(\frac{\varepsilon}{2C_1}, \mathcal{F}_2, L_2(Q)\big), \tag{E.37}$$

where Q is some probability measure.

Proof. For any two positive numbers ϵ_1 and ϵ_2, let $\mathcal{G} = \{g_1, \ldots, g_K\} \subseteq \mathcal{F}_1$ be an ε_1-net for $\mathcal{F}_1$ and $\mathcal{H} = \{h_1, \ldots, h_L\} \subseteq \mathcal{F}_2$ be an ε_2-net for $\mathcal{F}_2$, where $K = N(\varepsilon_1, \mathcal{F}_1, L_2(Q))$ and $L = N\big(\varepsilon_2, \mathcal{F}_2, L_2(Q)\big)$. For any function $f \in \mathcal{F}_1 \times \mathcal{F}_2$ with $f = f_1 f_2$, where $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$, let $\widetilde{g}$ and $\widetilde{h}$ be the closest element in $\mathcal{G}$ and $\mathcal{H}$ to f_1 and f_2 respectively. Then,

$$\begin{aligned}
\|f - \widetilde{g}\widetilde{h}\|_Q &= \|f_1 f_2 - \widetilde{g}\widetilde{h}\|_Q \\
&= \|(f_1 - \widetilde{g}) f_2 + \widetilde{g}(f_2 - \widetilde{h})\|_Q \\
&\leq \|(f_1 - \widetilde{g}) f_2\|_Q + \|\widetilde{g}(f_2 - \widetilde{h})\|_Q \\
&\leq C_2 \varepsilon_1 + C_1 \varepsilon_2 \\
&= \varepsilon, \text{ when we take } \varepsilon_1 = \varepsilon/2C_2 \text{ and } \varepsilon_2 = \varepsilon/2C_1.
\end{aligned}$$

This way $\{g_k h_l : 1 \leq k \leq K, 1 \leq l \leq L\}$ is an ε-net of $\mathcal{F}_1 \times \mathcal{F}_2$ and the inequality (E.37) follows. □

S.10 Some more simulation results

S.10.1 Absolute residual score

Figure A.1 and A.2 contain results comparing our doubly robust prediction method with different splits of data, i.e. we are comparing doubly robust prediction with full data described in Algorithm 2 and split data (Algorithm 1) with two splits and three splits, and weighted conformal prediction, all under the absolute residual score as described at the start of Section 7. It can be seen that DRP performs similarly across different splits and they all outperform WCP.

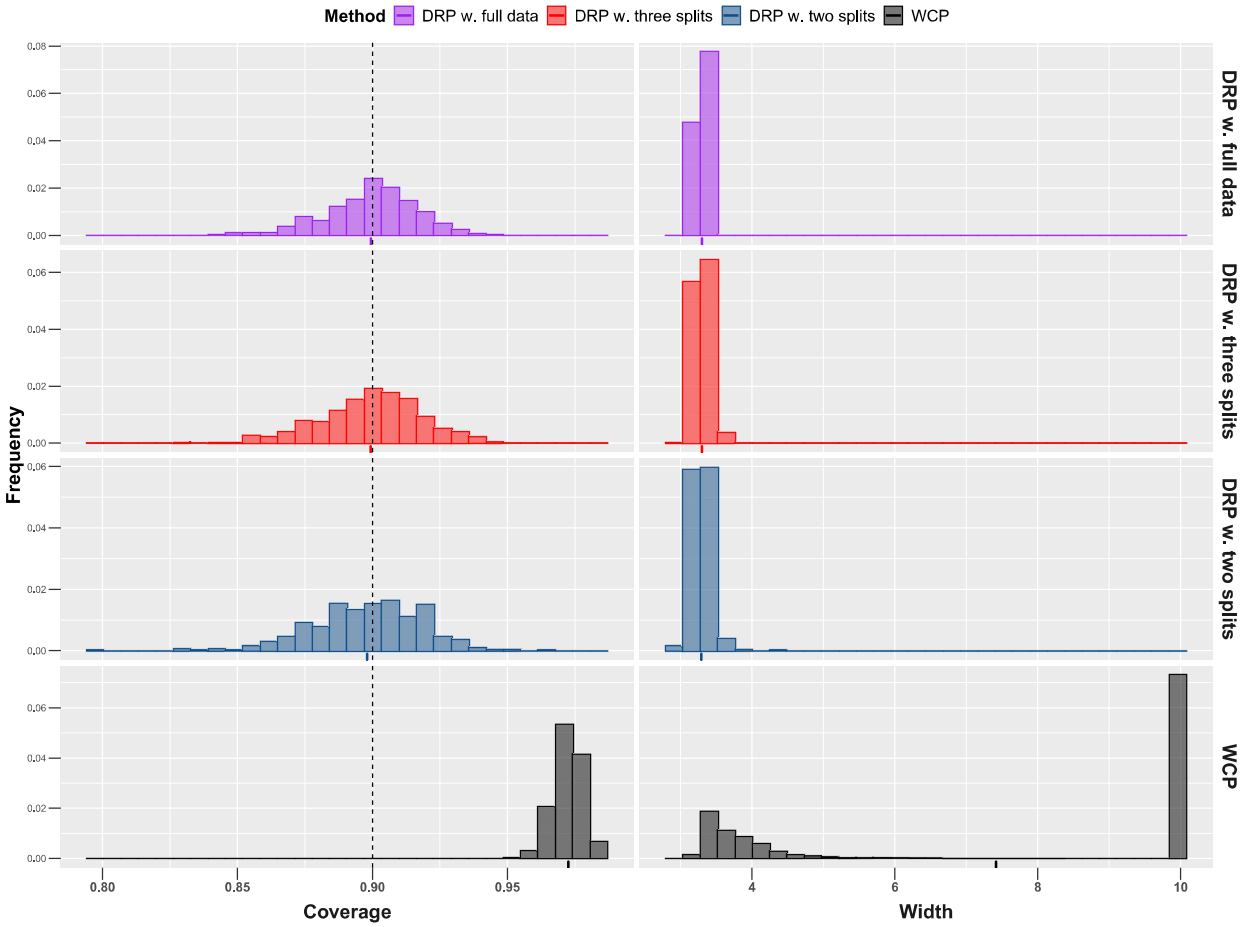

Figure A.1: Coverage and width of Double Robust Prediction (DRP) with full data, two splits, three splits and Weighted Conformal Prediction (WCP) on synthetic data with the absolute residual score. The width is truncated at 10 for WCP.

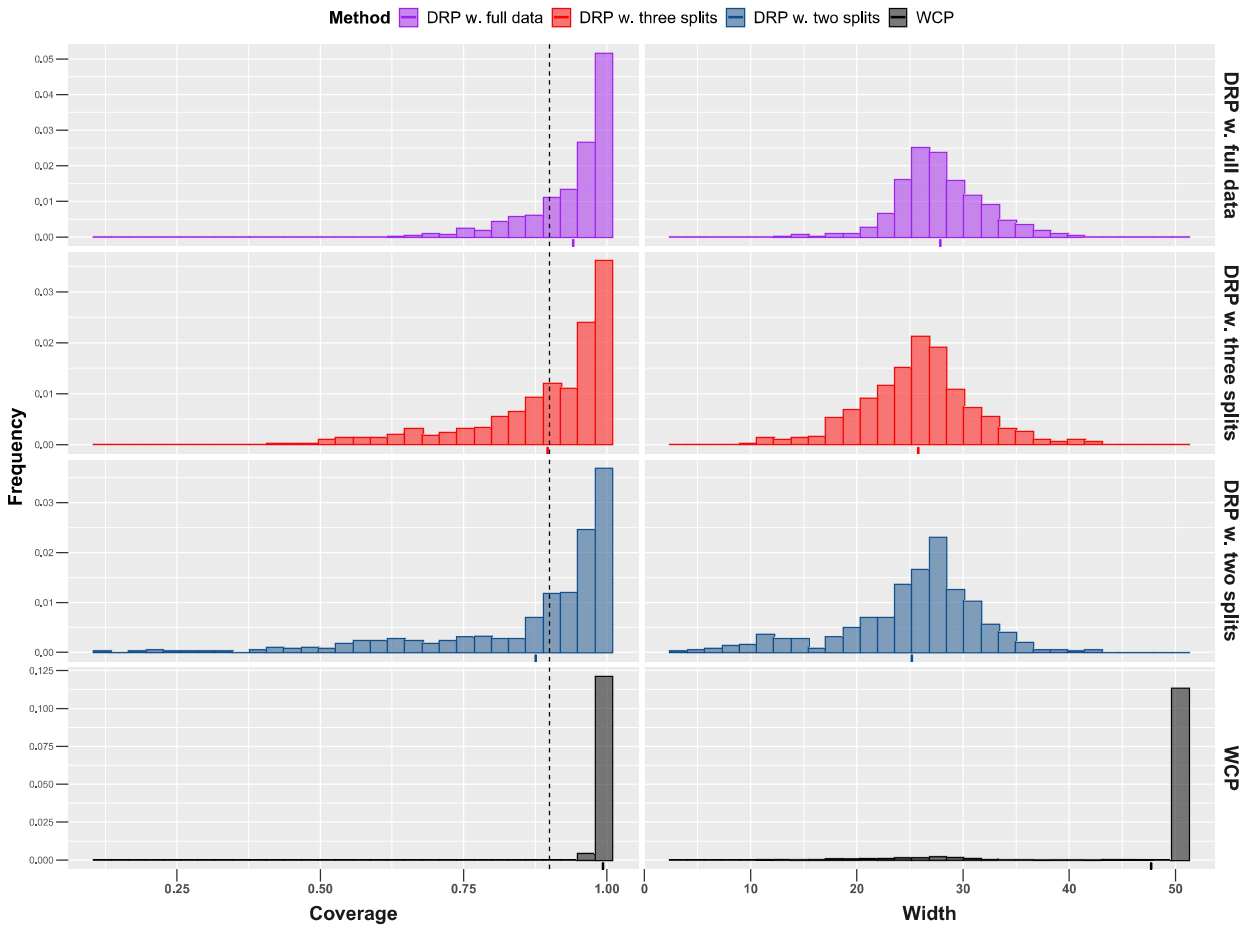

Figure A.2: Coverage and width of Doubly Robust Prediction (DRP) with full data, two splits, three splits and Weighted Conformal Prediction (WCP) on real data with the absolute residual score. The width is truncated at 50 for WCP.

S.10.2 CQR score

Figure A.3 illustrates DRP and WCP's performance on synthetic data (described in Section 7.2) under the CQR score. It is found that all the methods achieve the desired coverage with WCP being overly conservative, i.e. having infinity width over half of the cases.

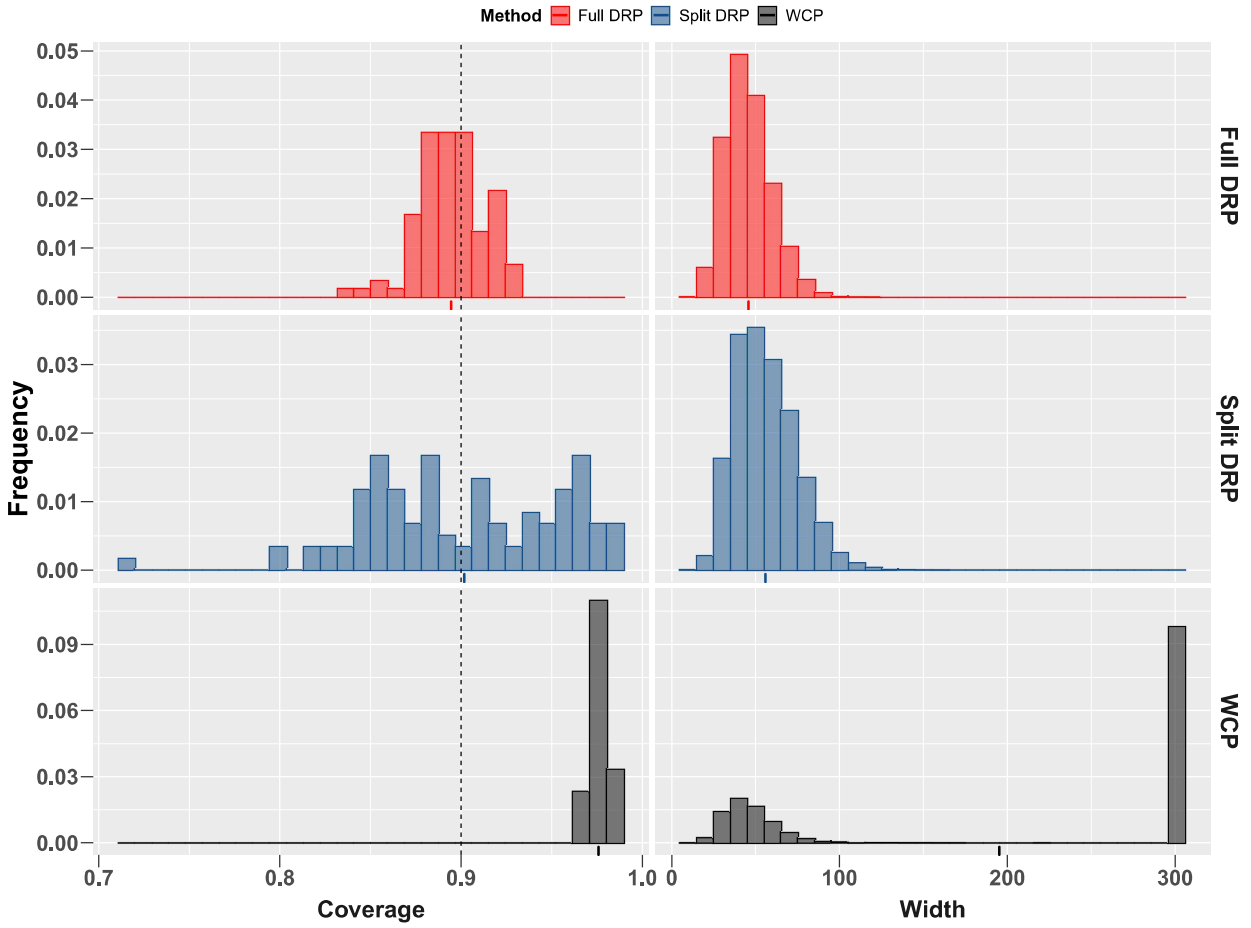

Figure A.3: Histograms of coverage and width of Doubly Robust Prediction (DRP) and Weighted Conformal Prediction (WCP) on synthetic data through CQR score, where the width is truncated at 300 for WCP.

S.11 Proof of Theorem 7

Proof of Theorem 7. By definition of the odds ratio function, we have the following expressions:

$$\log \frac{\mathbb{P}(T=0|X=x,Y=y)}{\mathbb{P}(T=1|X=x,Y=y)} = \log \frac{\mathbb{P}(T=0|X=x,Y=0)}{\mathbb{P}(T=1|X=x,Y=0)} + \gamma^\star(x,y);$$

It follows from Bayes' rule that

$$f(R|X,T=1) = f(R|X,T=0)\frac{\exp^{-\gamma^\star(X,Y)}}{\mathbb{E}\{\exp^{-\gamma^\star(X,Y)|T=0,X}\}}.$$

And this leads to

$$\mathbb{E}\Big(\mathbb{1}\{R\le\theta\}|X,T=1\Big) = \frac{\mathbb{E}[\mathbb{1}\{R\le\theta\}e^{-\gamma^\star(X,Y)}\big|X,T=0]}{\mathbb{E}[e^{-\gamma^\star(X,Y)}\big|X,T=0]}. \tag{E.38}$$

- If $\eta(x)$ is correct: This would imply the density ratio is correct, i.e.

$$\exp\{-\eta(x)-\gamma^\star(x,y)\} = \mathbb{P}(T=1|X=x,Y=y)/\mathbb{P}(T=0|X=x,Y=y),$$

then we have

$$\mathbb{E}\left[\mathbb{1}\{T=0\}\exp\{-\eta(X)-\gamma^\star(X,Y)\}\big|X=x,Y\right]=\mathbb{P}(T=0|X=x,Y)\frac{\mathbb{P}(T=1|X=x,Y)}{\mathbb{P}(T=0|X=x,Y)}=\mathbb{P}(T=1|X=x,Y).$$

This implies that

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}\{T=0\}\exp\{-\eta(X)-\gamma^\star(X,Y)\}\{\mathbb{1}\{R\le\theta\}-m(\theta,X)\}|\theta\right]\\&=\mathbb{E}\Big[\mathbb{P}(T=1|X,Y)\{\mathbb{1}\{R\le\theta\}-m(\theta,X)\}|\theta\Big]\\&=\mathbb{E}\Big[\mathbb{P}(T=1|X,Y)\{\mathbb{P}(R\le\theta|X)-m(\theta,X)\}|\theta\Big].\end{aligned}$$

Similarly,

$$\mathbb{E}\Big[\mathbb{1}\{T=1\}\left\{m(\theta,X)-(1-\alpha)\right\}|\theta\Big]=\mathbb{E}\Big[\mathbb{P}(T=1|X,Y)\{m(\theta,X)-(1-\alpha)\}|\theta\Big].$$

Hence, if $\exp\{-\eta(x)-\gamma^\star(x,y)\}$ is the true density ratio, then

$$\begin{aligned}\mathbb{E}[\mathrm{IF}(\theta,X,Y,R,T;\eta^\star,m,\gamma^\star)]&=\mathbb{E}\Big[\mathbb{P}(T=1|X,Y)\{\mathbb{P}(R\le\theta|X,\theta)-(1-\alpha)\}|\theta\Big]\\&=\mathbb{E}\Big[\mathbb{P}(T=1|X,Y)\{\mathbb{P}(R\le\theta|X,\theta)|\theta\Big]-(1-\alpha)\mathbb{P}(T=1)\\&=\mathbb{E}\Big[\mathbb{E}(\mathbb{1}\{T=1\}|X,Y)\mathbb{1}\{R\le\theta\}|\theta\Big]-(1-\alpha)\mathbb{P}(T=1)\\&=\mathbb{E}\Big[\mathbb{E}(1\{T=1\}\mathbb{1}\{R\le\theta\}|X,Y,\theta)\Big]-(1-\alpha)\mathbb{P}(T=1)\\&=\mathbb{E}\Big[1\{T=1\}\mathbb{1}\{R\le\theta\}|\theta\Big]-(1-\alpha)\mathbb{P}(T=1)\\&=\mathbb{P}(T=1)\Big\{\mathbb{E}\big[\mathbb{1}\{R\le\theta\}|T=1,\theta\big]-(1-\alpha)\Big\}.\end{aligned}$$

Therefore,

$$\mathbb{E}[\mathrm{IF}(r_\alpha,X,Y,R,T;\eta^\star,m,\gamma^\star)]=0.$$

- If m is correct: In this case,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}\{T=0\}\exp\{-\eta(X)-\gamma^\star(X,Y)\}\{\mathbb{1}\{R\le\theta\}-m(\theta,X)\}\right]\\&=\mathbb{E}\Big\{\exp\{-\eta(X)\}\mathbb{E}\Big[\mathbb{1}(R\le\theta)\exp\{-\gamma^\star(X,Y)\}|T=0,X\Big]\Big\}\mathbb{P}(T=0)\\&\quad-\mathbb{E}\Big\{\exp\{-\eta(X)\}\mathbb{P}(R\le\theta|X,T=1)\mathbb{E}\Big[\exp\{-\gamma^\star(X,Y)\}|T=0,X\Big]\Big\}\mathbb{P}(T=0)\\&=0,\ \text{by (E.38)}.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}[\mathrm{IF}(\theta,X,Y,R,T;\eta,m^\star,\gamma^\star)]&=\mathbb{E}\Big[\mathbb{1}\{T=1\}\big\{m^\star(\theta,X)-(1-\alpha)\big\}|\theta\Big]\\&=\mathbb{E}\Big[\mathbb{1}\{T=1\}\big\{\mathbb{E}\big(\mathbb{1}\{R\le\theta\}|X,T=1,\theta\big)-(1-\alpha)\big\}|\theta\Big]\\&=\mathbb{P}(T=1)\Big\{\mathbb{E}\Big[\mathbb{E}\big(\mathbb{1}\{R\le\theta\}|X,T=1,\theta\big)|T=1,\theta\Big]-(1-\alpha)\Big\}\\&=\mathbb{P}(T=1)\Big\{\mathbb{E}\big[\mathbb{1}\{R\le\theta\}|T=1,\theta\big]-(1-\alpha)\Big\}.\end{aligned}$$

Therefore,

$$\mathbb{E}[\mathrm{IF}(r_\alpha,X,Y,R,T;\eta,m^\star,\gamma^\star)]=0.$$

Finally, by the double robustness property of $\mathrm{IF}(\cdots)$, it can be easily verified that $\partial_t \mathbb{E}[\mathrm{IF}(\cdots; \eta_t, m^\star, \gamma^\star)]/\partial \eta_t$ and $\partial_t \mathbb{E}[\mathrm{IF}(\cdots; \eta^\star, m_t, \gamma^\star)]/\partial m_t$ for regular parametric submodels η_t and m_t are both zero at the truth. And this concludes our claim that $\mathrm{IF}(\cdots)$ is the efficient influence function up to a proportionality constant. $\square$