Conformal Inference of Individual Treatment Effects Using Conditional Density Estimates

Baozhen Wang, Xingye Qiao

Department of Mathematics and Statistics, Binghamton University
{bwang62, xqiao}@binghamton.edu

Abstract

In an era where diverse and complex data are increasingly accessible, the precise prediction of individual treatment effects (ITE) becomes crucial across fields such as healthcare, economics, and public policy. Current state-of-the-art approaches, while providing valid prediction intervals through Conformal Quantile Regression (CQR) and related techniques, often yield overly conservative prediction intervals. In this work, we introduce a conformal inference approach to ITE using the conditional density of the outcome given the covariates. We leverage the reference distribution technique to efficiently estimate the conditional densities as the score functions under a two-stage conformal ITE framework. We show that our prediction intervals are not only marginally valid but are narrower than existing methods. Experimental results further validate the usefulness of our method.

1 Introduction

Understanding the effect of interventions on an individual level is important across many domains, such as healthcare, economics, and public policy. Traditional average treatment effect estimates consider all individuals but fail to account for the heterogeneity in individual responses. As diverse data from various fields become more accessible, machine learning plays an increasingly significant role in revealing insights from the data.

In recent years, many research works have focused on machine learning algorithms that provide point estimates of the Conditional Average Treatment Effect (CATE) (Athey and Imbens 2016; Wager and Athey 2018; Künzel et al. 2019; Meng and Qiao 2020), which quantifies the expected difference in treatment outcomes for individuals with specific characteristics. While CATE takes into account the varying covariates of the individuals by averaging effects across similar individuals, it can still sometimes overlook individual-level heterogeneity. Recent works on Individualized Treatment Effect (ITE) have made significant progress in addressing these limitations, offering more accurate predictions tailored to each individual (Hill 2011; Alaa and Van Der Schaar 2017; Shalit, Johansson, and Sontag 2017). They utilize Gaussian processes or Bayesian approaches to provide interval-valued predictions at the individual level. However, these methods could be model-specific, and not always ensure that the predicted confidence intervals achieve the nominal coverage probability, i.e., the proportion of times that the true treatment effect lies within the constructed intervals across different samples. This issue can limit the reliability and generalizability of the treatment effect estimates, particularly when applied to diverse patient populations or under varying clinical conditions.

Conformal prediction (Vovk, Gammerman, and Shafer 2005; Lei, Robins, and Wasserman 2013; Lei et al. 2018) provides a model-agnostic and distribution-free framework that outputs interval predictions with the desired coverage probability. The main idea of (split) conformal prediction is to compute a conformity score by fitting a predictive model on the training set, and then evaluate the conformity score on the calibration set to quantify the uncertainty of future predictions. Although the conformal framework is model-agnostic and always achieves a coverage guarantee, the average length of resulting prediction intervals can highly rely on the choice of the score functions. Lei and Candès (2021) propose a two-stage methodology utilizing weighted conformal prediction (WCP) (Tibshirani et al. 2019) to address the covariate shift problem during counterfactual inference. This method is considered as the state-of-the-art method which provides prediction intervals for ITE problems, with desired coverage guarantee as well as reasonably short intervals. Chen et al. (2024) study a similar approach, where they use the joint density to compute weights in WCP. Alternatively, Alaa, Ahmad, and van der Laan (2023) develop a methodology called conformal meta-learner, which applies conformal prediction directly with imputed pseudo outcomes. Both methods utilize Conformal Quantile Regression (CQR) (Romano, Patterson, and Candes 2019), an approach that integrates the concept of conformal prediction with quantile regression. CQR is capable of adapting to any heteroscedasticity in the data, but may not guarantee the shortest prediction intervals. Additionally, the two-stage design inherent in both methods tends to produce conservative results. Experimental studies consistently show that these methods yield conservative prediction intervals with a coverage much greater than the desired level (Lei and Candès 2021; Alaa, Ahmad, and van der Laan 2023). Motivated by the need for more precise prediction interval for ITE, this work aims to refine these methodologies to achieve shorter prediction intervals while

maintaining the coverage guarantee.

Inspired by the conformal predictions under the classification setting (Lei 2014; Sadinle, Lei, and Wasserman 2019), one can show that directly using the conditional density of the outcome Y given the covariates X as the score function in conformal prediction will lead to the shortest prediction interval (see discussion in Section 3.1). While a substantial amount of research has focused on estimating the regression function $E(y|x)$, the task of estimating the full conditional density $f(y|x)$, particularly in scenarios where x is high-dimensional, has received considerably less attention. Izbicki, Shimizu, and Stern (2022) proposed a framework that utilizes conditional densities under regression setting. However, their focus is on achieving local conditional coverage and they employ a non-parametric smoothing technique to estimate the conditional densities, which is less efficient.

In this paper, we introduce a novel approach to conformal inference of individual treatment effects (ITE) using conditional densities as the score function. To address the computational difficulties associated with estimating full conditional densities, we employ a reference distribution technique to alleviate the problem. We theoretically show that the proposed method achieves shorter prediction intervals as well as maintaining the desired coverage guarantee. Empirical studies, including both simulations and semi-synthetic benchmarks, strongly indicate that our proposed method surpasses existing state-of-the-art methods in prediction length.

The remainder of this article is outlined as follows. We begin by reviewing the background knowledge of the potential outcome framework and conformal prediction in Section 2. We introduce our methodologies along with the discussion of theoretical guarantee in Section 3. In Section 4, we present supporting simulation and semi-synthetic experiments. Section 5 concludes by summarizing our contributions and the practical implications. Proofs and additional discussions can be found in the supplementary material.

2 Background

In this section, we first introduce the standard potential outcome framework. Then we review the conformal prediction in Section 2.2 and the weighted conformal prediction under covariate shift in Section 2.3.

2.1 Potential Outcome Framework and Objective Statement

We focus on the standard potential outcome framework (Neyman 1923; Rubin 1974) with a binary treatment. Denote by $A = \{0, 1\}$ the binary treatment indicator, by $X \in \mathcal{X} \subseteq \mathbb{R}^d$ the covariates, by $Y \in \mathcal{Y} \subseteq \mathbb{R}$ the observed outcome. For each subject i, let $Y_i(1)$ and $Y_i(0)$ be the pair of potential outcomes under $A = 1$ and $A = 0$ respectively. The fundamental problem of causal inference is that we can only observe one potential outcome out of $Y_i(1)$ and $Y_i(0)$ for each subject (Holland 1986). Let

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

The following assumptions are often considered: (1) *Unconfoundedness (strong ignorability)*: $(Y(1), Y(0)) \perp A|X$, which allows us to interpret the differences in outcomes as causal effects, rather than being confounded by other factors. Under unconfoundedness, the conditional distributions of a potential outcome are invariant across treatment groups: $P(Y|X, A = a) = P(Y(a)|X)$. (2) *Stable unit treatment value assumption* (SUTVA): $Y_i = Y_i(A)$, which ensures individual treatment effects are not influenced by other units. (3) *Positivity*: $0 < P(A = 1|X = x) < 1$, which ensures that every individual has a nonzero probability of receiving each treatment condition.

Existing methods mostly focused on conditional average treatment effects (CATE) (Athey and Imbens 2016; Wager and Athey 2018; Künzel et al. 2019), defined as $\tau(x) = E(Y(1) - Y(0)|X = x)$. In this work, our primary focus is on individual treatment effects (ITE), defined as $Y_i(1) - Y_i(0)$ for subject i without knowing the treatment assignment, and to construct a prediction interval of the ITE. Given the observations $(X_i, Y_i, A_i), i = 1, \ldots, n$, our goal is to construct a predictive interval $\hat{C}(x)$ that covers the true ITE for a new test individual $n+1$ with covariate X_{n+1} with high probability, i.e.

$$\mathbb{P}\left(Y_{n+1}(1) - Y_{n+1}(0) \in \hat{C}(X_{n+1})\right) \geq 1 - \alpha, \quad (1)$$

for a pre-specified level $\alpha \in (0, 1)$. Typically, α is a small value, such as 0.05.

2.2 Conformal Prediction

Conformal prediction (Vovk, Gammerman, and Shafer 2005; Lei, Robins, and Wasserman 2013; Lei et al. 2018) provides a means to a prediction set that with a predetermined probability covers the true value for future individuals based on a finite sample. Below we describe the construction of the original conformal prediction set. We first choose a score function $S(\cdot, \cdot)$, whose arguments consist of a point (x, y), and some dataset D. By convention, a low value of $S((x, y), D)$ indicates that the point (x, y) "conforms" to D, whereas a high value indicates that (x, y) is atypical relative to the points in D. For convenience of the method development below in this article, we choose S to be a conformity, instead of nonconformity, score; that is, a high value of S indicates that (x, y) "conforms" to D.

Given a training data set $(X_i, Y_i), i = 1, ..., n$, and a fixed $x \in \mathbb{R}^d$, we obtain $\hat{C}(x) \subseteq \mathbb{R}$, the conformal prediction set, by repeating the following procedure for each $y_{\text{trial}} \in \mathbb{R}$: we first calculate the conformity scores $V_i^{(x,y_{\text{trial}})} = S((X_i, Y_i), \bigcup_{i=1}^n (X_i, Y_i) \cup \{(x, y_{\text{trial}})\})$, for $i = 1, \ldots, n$, and $V_{n+1}^{(x,y_{\text{trial}})} = S((x, y_{\text{trial}}), \bigcup_{i=1}^n (X_i, Y_i))$, and then we include y_{trial} in $\hat{C}(x)$ if $V_{n+1}^{(x,y_{\text{trial}})} \geq \text{Quantile}(\alpha; V_{1:n}^{(x,y_{\text{trial}})} \cup \{\infty\})$, that is, if no less than $\alpha(n + 1)$ many of $V_i^{(x,y_{\text{trial}})}$'s are no greater than $V_{n+1}^{(x,y_{\text{trial}})}$. Importantly, the symmetry in the construction of the conformity scores guarantees a satisfactory coverage rate in finite samples (Lei et al. 2018):

$$\mathbb{P}(Y \in \hat{C}(X)) \geq 1 - \alpha, \quad (2)$$

where $\mathbb{P}$ is taken over the $n + 1$ i.i.d. draws of training samples and the test point.

The above original version of conformal prediction provides a finite-sample guarantee of the coverage rate. However, it can be computationally expensive, especially if S has to be computed through an expensive machine-learning method. A more popular alternative is the split-conformal method (Papadopoulos et al. 2002; Vovk, Gammerman, and Shafer 2005), where the entire training data is split into two parts. The first part is used to estimate the score function S, which is then evaluated on the second part of the data. The splitting process not only alleviates the computational burden of the full conformal prediction but also mitigates the risk of overfitting, as the score function is calibrated on a separate set from where it was trained.

2.3 Weighted Conformal Prediction under Covariate Shift

Both the original and the split version of conformal prediction assume that the distributions of the target data and the training data are the same. Tibshirani et al. (2019) generalized conformal prediction for regression to WCP under covariate shift assumptions. Covariate shift (Shimodaira 2000; Sugiyama et al. 2007) refers to the scenario where the marginal distribution of covariates differs between the training and target data set, denoted as P_X and Q_X respectively, while the conditional distributions of the outcome given the covariates remain the same, denoted as $P_{Y|X}$. Lei and Candès (2021) studied the counterfactuals inference utilizing WCP under covariate shift, where

$$\text{Training: } (X,Y) \sim P_X \times P_{Y|X};$$

$$\text{Target: } (X,Y) \sim Q_X \times P_{Y|X}.$$

Here we briefly go over the algorithm. Assume that the probability measure of the target data covariates is absolutely continuous with respect to that of the training data covariates, we consider using the Radon-Nikodym derivative $w(x) = dQ_X(x)/dP_X(x)$ to address the shift in covariate distribution. Define $p_i(x) = w(x_i)/[\sum_{i'=1}^{n} w(x_{i'}) + w(x)]$, for $i = 1, \ldots, n$ and $p_{n+1}(x) = w(x)/[\sum_{i'=1}^{n} w(x_{i'}) + w(x)]$. We can use weighted quantile of the scores computed in the calibration data as the cutoff value, with $p_i(x)$ as the weight, to obtain the prediction set:

$$\hat{C}(x) = \Bigg\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \geq \text{Quantile}\Bigg(\alpha; \sum_{i=1}^{n} p_i(x)\delta_{V_i^{(x,y)}} + p_{n+1}(x)\delta_{\infty} \Bigg) \Bigg\},$$

where δ_c is a Dirac measure placing a point mass at c. Assume we have the true value of $w(x)$, Tibshirani et al. (2019) showed that $\hat{C}(x)$ satisfies:

$$\mathbb{P}_{(X,Y)\sim Q_X \times P_{Y|X}}(Y \in \hat{C}(X)) \geq 1 - \alpha.$$

Under the potential outcome framework, the covariate distribution of the training data is a mixture of $P_{X|A=1}$ and $P_{X|A=0}$. WCP can not directly handle training data of a mixed type due to computational challenges associated with weighted calculations. Lei and Candès (2021) propose a two-stage framework to overcome this issue. On the first stage, one can use training data from the treatment group $P_{X|A=1} \times P_{Y|X}$, to produce interval estimates for those from control group $P_{X|A=0} \times P_{Y|X}$ via WCP and vice versa. Then, in the second stage, one can integrate the interval outcomes from both groups using a secondary conformal prediction procedure or a naive Bonferroni correction. In Section 3.3, we will explore this two-stage framework, and propose another less conservative alternative using the concept of X-learner (Künzel et al. 2019).

3 Methodology

In this section, we begin by demonstrating how using the conditional density as the score function can optimize the length of the conformal prediction interval. In Section 3.2, we introduce a reference distribution technique for efficient estimation of conditional densities. In Section 3.3, we adopt the two-stage framework proposed by Lei and Candès (2021) to develop algorithms that compute shorter prediction intervals for the Individual Treatment Effect (ITE) of new subjects, ensuring desired coverage guarantees.

3.1 Conditional Density as the Score Function

Let $\mathbb{P}$ denote the joint distribution of (X, Y) and f denote the density of $\mathbb{P}$ with respect to Lebesgue measure. Throughout the article, we denote $f(y|x) = f(Y = y|X = x)$ as the conditional density of Y equaling y given X equals x.

We define $C : \mathbb{R}^d \to \mathcal{M}(\mathbb{R})$, where $\mathcal{M}(\mathbb{R})$ represents the set of all measurable intervals over $\mathbb{R}$. The function C serves as a confidence interval predictor, which provides an interval $C(x)$ intended to contain the response variable y, based on input x. Consider the following optimization problem that minimizes the expected length of these intervals while ensuring that the probability of y falling within $C(x)$ is at least the predefined level $1 - \alpha$:

$$\min_{C} \mathbb{E}\left\{|C(X)|\right\} \quad \text{subject to} \quad \mathbb{P}\left\{y \in C(X)\right\} \geq 1-\alpha \quad (3)$$

Theorem 1. *Let t_α denote the α quantile of $f(Y|X = x)$. The solution that optimizes* (3) *is given by $C_\alpha^* = \{(x, y) : f(y|x) \geq t_\alpha\}$. And the optimal predictor can be written as*

$$C_\alpha^*(x) = \{y : f(y|x) \geq t_\alpha\}. \quad (4)$$

The proof can be found in the supplementary material. A similar problem of (3) has been explored under classification contexts by Lei (2014) and Sadinle, Lei, and Wasserman (2019). According to Theorem 1, if we can consistently estimate $f(y|x)$ and apply it as a score function in conformal prediction, as detailed in Algorithm 1, we can optimize the length of the prediction interval while ensuring a coverage guarantee of at least $1 - \alpha$.

3.2 Estimate Conditional Density Using Reference Distribution Technique

Using $f(y|x)$ as the score function in conformal prediction can be optimal; however, estimating the full conditional density $f(y|x)$ presents significant challenges. Traditional methods like the non-parametric kernel density estimator

(De Gooijer and Zerom 2003) must address each dimension of $x \in \mathbb{R}^d$ and often struggle due to the curse of dimensionality. Contemporary approaches, such as tree-based (Holmes, Gray, and Isbell 2012) and neural network methods (Rothfuss et al. 2019), involve complex computations and may prove less efficient for large-scale applications. To effectively address this issue, we adapt a reference distribution technique originally intended for unconditional density estimation (Hastie et al. 2009). The details of this adaptation are described as follows.

Suppose we have n i.i.d. random samples drawn from the joint density $f(x, y) = f(y|x)h(x)$, denoted as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We use a reference probability density function $f_0(y)$, from which a sample of size n independent of $h(x)$ is drawn using Monte Carlo methods, denoted as $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n$. We then combine a duplicate of $x_1, x_2, \ldots, x_n$ with these $\tilde{y}$ to form a joint reference distribution $f_0(x, y) = f_0(y)h(x)$. By assigning $Z = 1$ to each data point from $f(x, y)$ and $Z = 0$ to those from $f_0(x, y)$, we estimate $\mu(x, y) := E(Z|x, y)$ by supervised learning using the aggregated dataset:

$$\mu(x, y) = \frac{f(x, y)}{f(x, y) + f_0(x, y)} = \frac{f/f_0}{1 + f/f_0}, \tag{5}$$

The resulting estimate, $\hat{\mu}(x, y)$, can be inverted to provide an estimate for the joint density

$$\hat{f}(x, y) = f_0(x, y) \cdot \frac{\hat{\mu}(x, y)}{1 - \hat{\mu}(x, y)}.$$

Dividing $h(x)$ on both sides, we obtain an estimator of conditional density:

$$\hat{f}(y|x) = f_0(y) \cdot \frac{\hat{\mu}(x, y)}{1 - \hat{\mu}(x, y)}. \tag{6}$$

Techniques such as logistic regression and random forests, which efficiently estimate log-odds $\log(f/f_0)$, are natural choices for this procedure. Generally, many reference density can be used for $f_0(y)$, provided that the support of the reference density covers that of the original one. However, in practice, the accuracy of $\hat{f}(y|x)$ can be influenced by the choice of $f_0(y)$. We recommend using a Gaussian distribution with the same mean and a slightly larger variance than the original y's to alleviate the extreme case of none overlapping.

3.3 Weighted Conformal Inference Using Conditional Density Estimates

By leveraging the reference distribution technique to efficiently estimate the conditional density, we can apply the weighted conformal prediction (Tibshirani et al. 2019) to derive an estimate of (6). As introduced in Section 2.3, Lei and Candès (2021) propose a two-stage framework to overcome the challenges of WCP when training data exhibit a mixed distribution of covariates under the potential outcome framework. A two-stage framework seems a common and necessary choice to compensate for never observing the potential outcome. Alaa, Ahmad, and van der Laan (2023) also utilize a two-stage framework, where they first impute a pseudo outcome and then apply conformalized quantile regression (CQR) on these pseudo outcomes along with the covariates. A central motivation of this paper is to reduce the length of the prediction interval as much as possible, due to the inherent conservativeness of prediction intervals under the potential outcome framework.

Algorithm 1 below outlines the procedure of the first stage, where we implement WCP using the conditional density estimate $\hat{f}(Y|X)$ as the score function. Within the first stage of our framework, Algorithm 1 is implemented twice: once using training data from the treatment group to obtain interval estimates for the control group, and conversely, using control group data to estimate intervals for the treatment group. In the former scenario, the weight function $w(x)$ is calculated as:

$$\frac{dP_{X|A=0}(x)}{dP_{X|A=1}(x)} \propto \frac{P(A = 0|X = x)}{P(A = 1|X = x)} = \frac{1 - \pi(x)}{\pi(x)} \tag{7}$$

where $\pi(x) := P(A = 1|X = x)$ is the propensity score (Rosenbaum and Rubin 1983). This score captures the treatment assignment mechanism under given covariate conditions.

Similarly, when using data from the control group to estimate intervals for the treatment group, the weight function is:

$$w(x) = \frac{dP_{X|A=1}(x)}{dP_{X|A=0}(x)} \propto \frac{\pi(x)}{1 - \pi(x)} \tag{8}$$

Upon implementing Algorithm 1 in both scenarios, we first partition the training data into two splits, indexed by $\mathcal{I}_1$ and $\mathcal{I}_2$. The first split is used to estimate the weight function $\hat{w}(x)$ and the conditional density estimate $\hat{f}(Y|X)$ using the reference distribution technique. We then evaluate both $\hat{w}(x)$ and $\hat{f}(Y|X)$ on the second split, then compute the threshold $\hat{t}_\alpha$ as the α quantile of the weighted conformity scores.

Theorem 2. *Under Algorithm 1, if the non-conformity scores V_i have no ties almost surely, Q_X is absolutely continuous with respect to P_X, and $\mathbb{E}[\hat{w}(X)] < \infty$, then, given the true weights, i.e., $\hat{w}(\cdot) = w(\cdot)$:*

$$P_{(X,Y)\sim Q_X \times P_{Y|X}}(Y_{n+1} \in \hat{C}(X_{n+1})) \geq 1 - \alpha.$$

In general, if $\hat{w}(\cdot) \neq w(\cdot)$, define $\Delta w = \frac{1}{2}\mathbb{E}_{X\sim P_X}|\hat{w}(X) - w(X)|$. In this case, coverage is lower bounded by $1 - \alpha - \Delta w$.

Theorem 2 establishes that for any choice of target covariate distribution Q_X, it is possible to obtain a prediction interval for the outcome with a desired coverage guarantee. The first part of Theorem 2 is a split version of Theorem 2 from (Tibshirani et al. 2019) and the second part adapts from Theorem 3 in (Lei and Candès 2021). We provide proofs in the supplementary material for completeness.

In stage one, we implement Algorithm 1 twice: once using training data from the treatment group in $\mathcal{I}_1$ with weights defined in (7) to compute $\hat{C}_i(X_i)$ for $i \in \mathcal{I}_2$ where $A_i = 0$, and once using data from the control group in $\mathcal{I}_1$ with weights in (8) to compute $\hat{C}_i(X_i)$ for $i \in \mathcal{I}_2$ where $A_i = 1$.

Algorithm 1: Weighted Split-Conformal using Conditional Density Estimate (CD)

Input: Level α, data (X_i, Y_i) from $P_X \times P_{Y|X}$ where $i \in \mathcal{I}$, and a test point X_{n+1} from Q_X

Output: A prediction set $\hat{C}(x)$

1: Split $\mathcal{I}$ into two equal sized subsets $\mathcal{I}_1$ and $\mathcal{I}_2$.
2: Estimate the weight function $\hat{w}(x)$ using $\mathcal{I}_1$.
3: For each $i \in \mathcal{I}_1$, generate $\tilde{Y}_i$ from a normal distribution with the same mean and slightly larger variance of data in $\mathcal{I}_1$. Assign $Z = 1$ to (X_i, Y_i) and $Z = 0$ to $(X_i, \tilde{Y}_i)$, then fit a classification algorithm $\hat{\mu}$ to obtain $\hat{f}(Y|X)$ according to Section 3.2.
4: For each $i \in \mathcal{I}_2$, compute the score $V_i = \hat{f}(Y_i|X_i)$ and the weight $\hat{w}(X_i)$
5: Compute the normalized weights $\hat{p}_i(x) = \hat{w}(X_i)/\big[\sum_{i\in\mathcal{I}_2} \hat{w}(X_i) + \hat{w}(X_{n+1})\big]$ and $\hat{p}_{n+1}(x) = \hat{w}(X_{n+1})/\big[\sum_{i\in\mathcal{I}_2} \hat{w}(X_i) + \hat{w}(X_{n+1})\big]$
6: Compute $\hat{t}_\alpha$ as the α-th quantile of the distribution $\sum_{i\in\mathcal{I}_2} \hat{p}_i(x)\delta_{V_i} + \hat{p}_\infty(x)\delta_\infty$
7: **return** $\hat{C}(X_{n+1}) = \big\{y : \hat{f}(y|X_{n+1}) \geq \hat{t}_\alpha\big\}$

We obtain prediction sets $\hat{C}_i(X_i)$ for each point in $\mathcal{I}_2$, which satisfy

$$\mathbb{P}(Y_i(1-j) \in \hat{C}_i(X_i)|A_i = j) \geq 1 - \alpha. \quad (9)$$

Recall for each individual i with $A_i = j$, we can observe its factual outcome $Y_i = Y_i(j)$. That means we can obtain a confidence interval of $\text{ITE}_i = Y_i(1) - Y_i(0)$ for each $i \in \mathcal{I}_2$ by subtraction. Specifically, for those in the treatment group, i.e., $A_i = 1$,

$$\hat{C}_i = \big[Y_i(1) - \max(\hat{C}(X_i)), Y_i(1) - \min(\hat{C}(X_i))\big], \quad (10)$$

and for those in the control group, i.e., $A_i = 0$,

$$\hat{C}_i = \big[\min(\hat{C}(X_i)) - Y_i(0), \max(\hat{C}(X_i)) - Y_i(0)\big]. \quad (11)$$

Eq. (10) and (11) are useful, but they depend on knowing the value of A_i, which is not available for future data.

In stage two, we apply a secondary procedure to the new training data pairs $(X_i, \hat{C}_i)$ to eliminate the dependency on the treatment assignment A. Algorithm 2 below sketches the *Exact* method, where we apply another split Conformal on $(X_i, \hat{C}_i)$. Here we denote $\hat{C}_i = \big(\hat{C}_i^L, \hat{C}_i^U\big)$ for simplicity.

Lemma 1 (Lei and Candès (2021))**.** *Assume $(X_i, \hat{C}_i)$ are i.i.d. from (X, C). Then, for a test point X_{n+1} under Algorithm 1 and 2, both with miscoverage level $\alpha/2$,*

$$\mathbb{P}(Y_{n+1}(1) - Y_{n+1}(0) \in \hat{C}_{ITE}(X_{n+1}) \geq 1 - \alpha.$$

Lemma 1 can be directly proved using Theorem 2 above and Theorem 2 in (Lei and Candès 2021). Lemma 1 shows that by using an exact method, we can construct a prediction interval that covers the true ITE with a desired coverage level as in (1).

Algorithm 2: CD-Exact

Input: Level γ, data $(X_i, \hat{C}_i), i \in \mathcal{I}_2$, where $\hat{C}_i$ are obtained from Algorithm 1 using control and treatment group respectively, as shown in (10) and (11), and a test point X_{n+1}

Output: A prediction interval $\hat{C}_{\text{ITE}}(X_{n+1})$

1: Split $\mathcal{I}$ into two equal sized subsets $\mathcal{I}_{tr}$ and $\mathcal{I}_{ca}$
2: On $\mathcal{I}_{tr}$, fit conditional mean $\hat{\tau}_L$ and $\hat{\tau}_U$ of $\hat{C}_i^L$ and $\hat{C}_i^U$ given X
3: For each $i \in \mathcal{I}_{ca}$, compute score $V_i = \max\big\{\hat{\tau}_L(X_i) - \hat{C}_i^L, \hat{C}_i^U - \hat{\tau}_U(X_i)\big\}$.
4: Compute η as the $1-\gamma$ quantile over scores V_i.
5: **return** $\hat{C}_{\text{ITE}}(X_{n+1}) = \big[\hat{\tau}_L(X_{n+1}) - \eta, \hat{\tau}_U(X_{n+1}) + \eta\big]$

Another method with theoretical guarantees is the *Naive* method, which employs a straightforward approach using a naive Bonferroni correction, designed as follows:

$$\hat{C}_{\text{ITE}}(x) = \big[\min(\hat{C}^1(x)) - \max(\hat{C}^0(x)), \max(\hat{C}^1(x)) - \min(\hat{C}^0(x))\big]$$

Here $\hat{C}^1$ and $\hat{C}^0$ denote the prediction intervals computed for the treatment group and the control group, respectively. By setting the miscoverage level in Algorithm 1 to be $1 - \alpha/2$, the Naive method achieves the desired coverage probability of $1 - \alpha$ for the ITE, as specified in (1).

In practice, while the Exact and Naive methods tend to be overly conservative, utilizing the conditional density estimate as the score function in Algorithm 1 helps to mitigate this issue, as demonstrated in our empirical results in Section 4.2. A practical and favorable alternative is the *Inexact* method, which yields much shorter prediction intervals, albeit without theoretical guarantees. To implement the Inexact method, we fit plug-in estimates for the 40% conditional quantiles of $\hat{C}_i^L$ and 60% conditional quantiles of $\hat{C}_i^U$, respectively. For a new test point, these quantiles are then used to straightforwardly compute the prediction interval.

Inspired by X-learner (Künzel et al. 2019), we propose a fourth, less conservative alternative method named *CD-X*. We fit four plug-in estimates for the conditional means of $\hat{C}_i^L$ and $\hat{C}_i^U$ for both $A_i = 0$ and $A_i = 1$, denoted as $\tilde{C}_0^L(x), \tilde{C}_0^U(x), \tilde{C}_1^L(x), \tilde{C}_1^U(x)$. In Algorithm 1, we also estimate the propensity scores $\hat{\pi}(x)$ using $\mathcal{I}_1$. Then, for a new test individual, the prediction interval can be computed using the formula

$$\hat{C}_{\text{ITE}}(x) = \big[\hat{\pi}(x)\tilde{C}_1^L(x) + (1 - \hat{\pi}(x))\tilde{C}_0^L(x), \hat{\pi}(x)\tilde{C}_1^U(x) + (1 - \hat{\pi}(x))\tilde{C}_0^U(x)\big]$$

Although CD-X does not offer theoretical guarantees, it performs well in most cases within our numerical experiments, achieving the desired coverage meanwhile producing the shortest prediction intervals. In the next section, we will examine the empirical performance of methods ensemble with CD (Algorithm 1) and WCP.

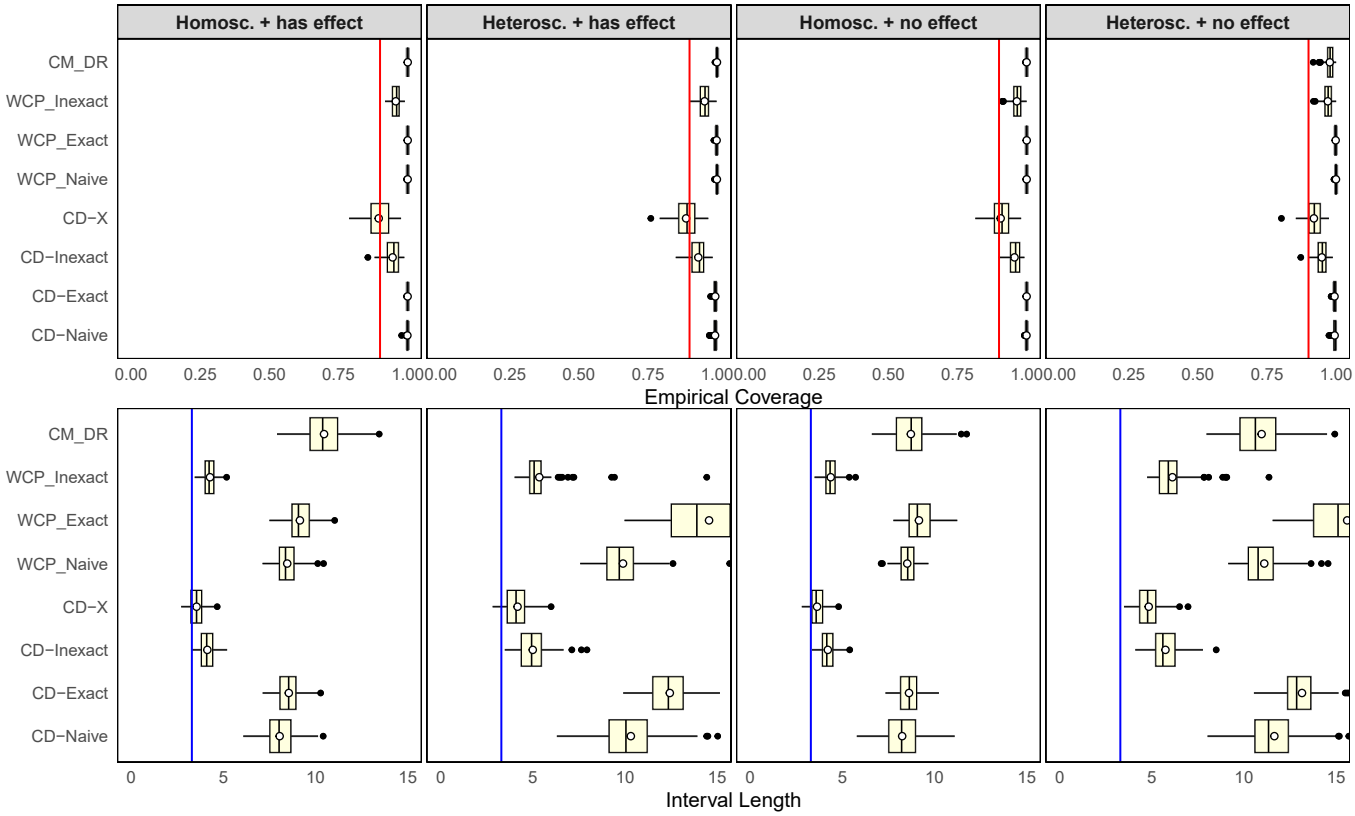

Figure 1: Performance of all baselines in four simulation scenarios described in Section 4.2. The red vertical lines correspond to target coverage, and the blue vertical lines correspond to the optimal interval width. Here CD stands for Algorithm 1, WCP stands for weighted conformal prediction, and CM stands for conformal meta-learners.

4 Experimental Studies

4.1 Experimental Setup

The nature of potential outcomes limits our observations to factual outcomes, excluding counterfactuals. This characteristic necessitates the validation of ITE estimation primarily through simulations and semi-synthetic data. In this section, we will explore the performance of our methods across one simulation featuring four different settings, and three semi-synthetic benchmarks.

We consider the following baselines known to produce valid prediction intervals for ITEs:

- **WCP-Inexact**, **WCP-Exact**, and **WCP-Naive**: We consider state-of-the-art (SOTA) methods based on WCP, each integrating the Inexact, Exact, and Naive approaches.
- **CM-DR**: We also evaluate the Conformal Meta-learners (Alaa, Ahmad, and van der Laan 2023) with the doubly-robust learner. This method is one of the top performers among three methods proposed in (Alaa, Ahmad, and van der Laan 2023) and also offers theoretical guarantees. Throughout our experimental studies, we estimate the propensity score in CM-DR rather than assuming it is known, as done in the original study by (Alaa, Ahmad, and van der Laan 2023), to ensure a fair comparison.

4.2 Simulation Studies

In the simulation studies, we combine the data-generation processes described in Lei and Candès (2021) and Alaa, Ahmad, and van der Laan (2023), which were originally proposed by Wager and Athey (2018). The data are generated as follows: Covariates X are sampled from $\text{Unif}([0,1]^d)$. The propensity score is generated based on $\pi(x) = \frac{1}{4}[1+\beta_{2,4}(x_1)]$, and the treatment assignment $A|X$ is generated from $\text{Bern}(\pi(x))$. The potential outcomes, $Y(j)$ for $j = 0, 1$, are modeled based on the function $g(x) = 2/\{1+\exp[-12(x-0.5)]\}$, where $\mathbb{E}[Y(1)|X] = g(x_1)g(x_2)$ and $\mathbb{E}[Y(0)|X] = \gamma g(x_1)g(x_2)$. Here γ controls the treatment effect. These outcomes are then generated from the model $\mathbb{E}[Y(j)|X] + \sigma(X)\epsilon$, where $\epsilon \sim N(0,1)$. We consider four scenarios, derived from a 2 x 2 factorial design: homoscedastic ($\sigma(x) = 1$) and heteroscedastic ($\sigma(x) = -\log(x_1)$) errors, and treatment has no effect ($\gamma = 1$) and the effects are heterogeneous ($\gamma = 0$).

4.3 Semi-synthetic Datasets

We also explore the performance of our approaches on three semi-synthetic datasets, which feature real covariates combined with simulated outcomes. These datasets include the National Study of Learning Mindsets (NSLM) (Yeager et al. 2019), the 2016 Atlantic Causal Inference Conference Com-

Method	NSLM		ACIC		IHDP		With guarantee
	Coverage	Avg. len.	Coverage	Avg. len.	Coverage	Avg. len.	
CM-DR	99.9 (0.00)	6.45 (0.14)	99.9 (0.02)	52.9 (1.80)	99.9 (0.03)	82.1 (5.31)	✔
WCP-Inexact	93.6 (0.29)	2.20 (0.03)	96.1 (0.48)	16.6 (0.40)	83.2 (0.87)	8.86 (0.72)	✘
WCP-Exact	99.9 (0.01)	4.48 (0.04)	99.8 (0.04)	40.7 (2.11)	99.5 (0.10)	34.9 (4.60)	✔
WCP-naive	99.9 (0.01)	4.23 (0.03)	99.8 (0.04)	30.7 (1.38)	99.3 (0.13)	21.4 (2.09)	✔
CD-X	86.3 (0.37)	1.78 (0.02)	93.9 (0.42)	**12.2** (0.30)	79.9 (0.83)	6.40 (0.51)	✘
CD-Inexact	92.0 (0.29)	**2.09** (0.02)	94.7 (0.49)	14.4 (0.37)	80.3 (0.93)	8.31 (0.69)	✘
CD-Exact	99.9 (0.01)	4.13 (0.03)	99.5 (0.07)	30.4 (0.63)	99.4 (0.10)	25.6 (2.95)	✔
CD-Naive	99.8 (0.02)	4.02 (0.04)	99.2 (0.13)	29.3 (0.67)	96.6 (0.39)	**20.2** (2.04)	✔

Table 1: Performance of all methods on semi-synthetic datasets. Empirical coverage (in percentages) and average interval lengths are shown, with standard errors in parentheses. Bold numbers highlight the best performance. "With guarantee" indicates methods that provide theoretical coverage guarantees.

petition (ACIC) (Dorie et al. 2019), and the Infant Health and Development Program (IHDP) datasets (Hill 2011). Detailed descriptions of all datasets are available in the supplementary material.

4.4 Results and Discussion

As our focus is on the prediction interval for ITEs, we use two commonly used metrics across all experimental studies: empirical coverage for true ITEs and average prediction interval length. The empirical coverage is defined as the empirical probability that the true ITE falls within the predicted interval. The average prediction interval length is measured as the mean of the widths of these intervals across all test instances.

Figure 1 illustrates the outcomes of the simulation studies, averaging results over 100 replications for each scenario as described in Section 4.2. CD-X generally excels, achieving the shortest and near-optimal interval lengths while maintaining the desired coverage levels, except in the scenario with heteroscedastic errors and treatment effect where it undercovers by 0.01. When comparing methods incorporating CD with those using WCP, CD methods consistently provide shorter intervals. Specifically, CD-Inexact is on average 0.26 shorter than WCP-Inexact, CD-Exact is 1.43 shorter than WCP-Exact. CD-Naive and WCP-Naive achieve nearly the same interval lengths on average.

The semi-synthetic experiments further demonstrate the superiority of the proposed CD methods. Across all three benchmarks, CD methods consistently outperform others. Although CD-X and CD-Inexact lack coverage guarantees, they excel in the ACIC and NSLM datasets, respectively, with the shortest average length and the coverage guarantee. For each pair of matched CD and WCP methods, CD consistently provides shorter intervals in *all* cases. Regarding the IHDP results, Alaa and Van Der Schaar (2017) reported that CM-DR performed well with an average interval length of 16.7, assuming a known propensity score. We emphasize that in our studies, we estimate the propensity score for CM-DR to ensure a fair comparison. This is particularly critical given the imbalanced setting of the IHDP dataset, which includes only 747 samples (139 treated and 608 control), making accurate propensity score estimation challenging. In this context, CD-Naive outperforms all other methods.

Although our primary goal is to mitigate the excessive conservatism of prediction intervals, the inherent challenges posed by unobserved counterfactuals naturally lead to conservative outcomes. In our experimental studies, the CD-Exact method consistently demonstrated coverage rates exceeding 99%, despite a targeted desired coverage of 90%. Our two-stage framework for predictive inference of ITE addresses the reduction of prediction interval length for counterfactuals in Stage 1 by efficiently using conditional density as the score function. However, further reducing conservatism in Stage 2 remains an area for future research, aiming to find an optimal balance between the less conservative Inexact methods and the overly conservative Exact methods.

5 Conclusions

In this paper, we developed a two-stage framework to provide interval estimates for Individual Treatment Effects (ITEs), a task inherently challenging due to the nature of unobservable potential outcomes. We successfully leveraged the reference distribution technique to efficiently estimate the optimal conformal scores—the conditional densities. Both theoretical and experimental results demonstrate that our framework outperforms existing state-of-the-art Weighted Conformal Prediction (WCP) methods. The practical implications of our work can be substantial, particularly given the difficulty inherent in estimating ITEs. Our method's success in reducing the average length of prediction intervals enhances their usability in real-world scenarios, particularly in decision-making processes requiring precise estimates, such as in clinical and policy-making fields.

References

Alaa, A.; Ahmad, Z.; and van der Laan, M. 2023. Conformal Meta-learners for Predictive Inference of Individual Treatment Effects. *arXiv preprint arXiv:2308.14895*.

Alaa, A. M.; and Van Der Schaar, M. 2017. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30.

Athey, S.; and Imbens, G. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27): 7353–7360.

Chen, Z.; Guo, R.; Ton, J.-F.; and Liu, Y. 2024. Conformal counterfactual inference under hidden confounding. In *Pro-*

ceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 397–408.

De Gooijer, J. G.; and Zerom, D. 2003. On conditional density estimation. *Statistica Neerlandica*, 57(2): 159–176.

Dorie, V.; Hill, J.; Shalit, U.; Scott, M.; and Cervone, D. 2019. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1): 43–68.

Hastie, T.; Tibshirani, R.; Friedman, J. H.; and Friedman, J. H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hill, J. L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1): 217–240.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960.

Holmes, M. P.; Gray, A. G.; and Isbell, C. L. 2012. Fast nonparametric conditional density estimation. *arXiv preprint arXiv:1206.5278*.

Izbicki, R.; Shimizu, G.; and Stern, R. B. 2022. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87): 1–32.

Künzel, S. R.; Sekhon, J. S.; Bickel, P. J.; and Yu, B. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10): 4156–4165.

Lei, J. 2014. Classification with confidence. *Biometrika*, 101(4): 755–769.

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

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

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

Meng, H.; and Qiao, X. 2020. Doubly robust direct learning for estimating conditional average treatment effect. *arXiv preprint arXiv:2004.10108*.

Neyman, J. 1923. On the application of probability theory to agricultural experiments. Essay on principles. *Ann. Agricultural Sciences*, 1–51.

Papadopoulos, H.; Proedrou, K.; Vovk, V.; and Gammerman, A. 2002. Inductive confidence machines for regression. In *European Conference on Machine Learning*, 345–356. Springer.

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

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Rothfuss, J.; Ferreira, F.; Walther, S.; and Ulrich, M. 2019. Conditional density estimation with neural networks: Best practices and benchmarks. *arXiv preprint arXiv:1903.00954*.

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

Sadinle, M.; Lei, J.; and Wasserman, L. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, 3076–3085. PMLR.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2): 227–244.

Sugiyama, M.; Nakajima, S.; Kashima, H.; Buenau, P.; and Kawanabe, M. 2007. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20.

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

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.

Wager, S.; and Athey, S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242.

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