Conditional Distributional Treatment Effect with Kernel Conditional Mean Embeddings and U-Statistic Regression

Junhyung Park [1] **Uri Shalit** [2] **Bernhard Schölkopf** [1] **Krikamol Muandet** [1]

Abstract

We propose to analyse the conditional distributional treatment effect (CoDiTE), which, in contrast to the more common conditional average treatment effect (CATE), is designed to encode a treatment's distributional aspects beyond the mean. We first introduce a formal definition of the CoDiTE associated with a distance function between probability measures. Then we discuss the CoDiTE associated with the maximum mean discrepancy via kernel conditional mean embeddings, which, coupled with a hypothesis test, tells us whether there is any conditional distributional effect of the treatment. Finally, we investigate what kind of conditional distributional effect the treatment has, both in an exploratory manner via the conditional witness function, and in a quantitative manner via U-statistic regression, generalising the CATE to higher-order moments. Experiments on synthetic, semi-synthetic and real datasets demonstrate the merits of our approach.

1. Introduction

Analysing the effect of a treatment (medical drug, economic programme, etc.) has long been a problem of great importance, and has attracted researchers from diverse domains, including econometrics (Imbens & Wooldridge, 2009), political sciences (Künzel et al., 2019), healthcare (Foster et al., 2011) and social sciences (Imbens & Rubin, 2015). The field has naturally received much attention of statisticians over the years (Rosenbaum, 2002; Rubin, 2005; Imbens & Rubin, 2015), and in the past few years, the machine learning community has started applying its own armoury to this problem – see Section 1.2 for a succinct review.

Traditional methods for treatment effect evaluation focus on the analysis of the average treatment effect (ATE), such as an increase or decrease in average income, inequality or poverty, aggregated over the population. However, the ATE is not informative about the individual responses to the intervention and how the treatment impact varies across individuals (known as *treatment effect heterogeneity*). The study of conditional average treatment effect (CATE)[1] has been proposed to analyse such heterogeneity in the mean treatment effect. Although sufficient in many cases, the CATE is still an average. As such, it fails to capture information about distributional aspects of the treatment beyond the mean. A significant amount of interest exists for developing methods that can analyse distributional treatment effects conditioned on the covariates (Chang et al., 2015; Bitler et al., 2017; Shen, 2019; Chernozhukov et al., 2020; Hohberg et al., 2020; Briseño Sanchez et al., 2020).

Our contributions are as follows. Firstly, we formally define the *conditional distributional treatment effect* (CoDiTE) associated with a chosen distance function between distributions. Then we use kernel conditional mean embeddings (Song et al., 2013; Park & Muandet, 2020a) to analyse the CoDiTE associated with the *maximum mean discrepancy* (Gretton et al., 2012). Coupled with a statistical hypothesis test, this can determine *whether* there exists any effect of the treatment, conditioned on a set of covariates. Finally, we use *conditional witness functions* and *U-statistic regression* to investigate *what kind* of effect the treatment has.

1.1. Problem Set-Up: Potential Outcomes Framework

Throughout this paper, we take $(\Omega, \mathcal{F}, P)$ as the underlying probability space, $\mathcal{X}$ as the input space and $\mathcal{Y} \subseteq \mathbb{R}$ as the output space. Let $Z : \Omega \to \{0, 1\}$, $X : \Omega \to \mathcal{X}$ and $Y_0, Y_1, Y : \Omega \to \mathcal{Y}$ be random variables representing, respectively, the treatment assignment, covariates, the potential outcomes under control and treatment, and the observed outcome, i.e. $Y = Y_0(1 - Z) + Y_1 Z$. For example, Z may indicate whether a subject is administered a medical treatment ($Z = 1$) or not ($Z = 0$). The potential outcomes Y_1, Y_0 respectively correspond to subject's responses had they received treatment or not. The covariates X corre-

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]Technion, Israel Institute of Technology. Correspondence to: Junhyung Park <junhyung.park@tuebingen.mpg.de>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]See Section 1.1 for the definitions of the ATE and CATE.

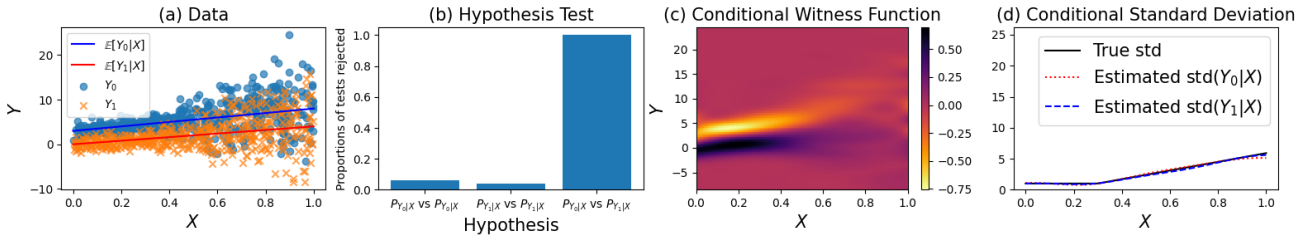

Figure 1. **Toy illustration of higher-order heterogeneity that cannot be captured by CATE. (a) Data.** $X \sim \text{Uniform}[0,1]$, $Y_0 = 3 + 5X + \mathbf{1}_{X<0.3} N + 7\mathbf{1}_{X\geq 0.3}(1 + (X - 0.3))N$ and $Y_1 = 4X + \mathbf{1}_{X<0.3} N + 7\mathbf{1}_{X\geq 0.3}(1 + (X - 0.3))N$, where $N \sim \mathcal{N}(0,1)$; in particular, the CATE is increasing with X. **(b) Hypothesis test** (Section 4.2) Each of the hypotheses $P_{Y_0|X} \equiv P_{Y_0|X}$, $P_{Y_1|X} \equiv P_{Y_1|X}$ and $P_{Y_0|X} \equiv P_{Y_1|X}$ are tested 100 times. The last (false) hypothesis is rejected in most tests, while the first two (true) hypotheses are not rejected in most tests, meaning that both type I and type II errors are low. **(c) Conditional witness function** (Section 5.1). The conditional witness function is close to zero for all Y at $X \geq 0.5$, demonstrating that $P_{Y_0|X}$ and $P_{Y_1|X}$ are similar in this region of $\mathcal{X}$. For $X < 0.4$, the witness function is positive in regions where the density of Y_1 is higher than that of Y_0, and negative in regions where the density of Y_0 is higher than that of Y_1. **(d) U-statistic regression** (Section 5.2). True conditional standard deviation (in black) is estimated (in red and blue for control and treatment groups respectively) as a function of X via U-statistic regression (since variance is a U-statistic) and the square-root operation. We see that the standard deviation increases linearly for $X \geq 0.3$.

spond to subject's characteristics such as age, gender, race that could influence both the potential outcomes and the choice of treatment. We denote the distributions of random variables by subscripting P, e.g. P_X for the distribution of X. Throughout, we impose the mild condition that conditional distribution $P(\cdot \mid X)$ admits a *regular version* (Çınlar, 2011, p.150, Definition 2.4, Proposition 2.5).

Each unit $i = 1, ..., n$ is associated with an independent copy $(X_i, Z_i, Y_{0i}, Y_{1i})$ of (X, Z, Y_0, Y_1). However, for each $i = 1, ..., n$, we observe either Y_{0i} or Y_{1i}; this missing value problem is known as the *fundamental problem of causal inference* (Holland, 1986), preventing us from directly computing the difference in the outcomes under treatment and control for each unit. As a result, we only have access to samples $\{(x_i, z_i, y_i)\}_{i=1}^n$ of (X, Z, Y). We write $n_0 = \sum_{i=1}^n \mathbf{1}_{z_i=0}$ and $n_1 = \sum_{i=1}^n \mathbf{1}_{z_i=1}$ for the control and treatment sample sizes, and denote the control and treatment samples by $\{(x_i^0, y_i^0)\}_{i=1}^{n_0}$ and $\{(x_i^1, y_i^1)\}_{i=1}^{n_1}$.

We assume *strong ignorability* (Rosenbaum & Rubin, 1983):

unconfoundedness $Z \perp\!\!\!\perp (Y_0, Y_1) \mid X$; and

overlap $0 < e(X) = P(Z = 1 \mid X) = \mathbb{E}[Z \mid X] < 1$.

Causal treatment effects are then identifiable from observational data, since $P_{Y_0|X} = P_{Y_0|X,Z=0} = P_{Y|X,Z=0}$, and similarly for $P_{Y_1|X}$. The quantity $e(X)$ is the *propensity score*. In a *randomised experiment*, $e(X)$ is known and controlled (Imbens & Rubin, 2015, p.40, Definition 3.10).

The usual objects of interest in the treatment effect literature are the *average treatment effect* (ATE), $\mathbb{E}[Y_1 - Y_0]$, and the *conditional average treatment effect* (CATE), $T(x) = \mathbb{E}[Y_1 - Y_0 \mid X = x]$. In this paper, we propose to extend the analysis to compare other aspects of the conditional distributions, $P_{Y_0|X}$ and $P_{Y_1|X}$. One compelling reason to do this is that estimating CATE is inherently a problem of *comparing two means*, and as such, is only meaningful if the corresponding variances are given. Consider the toy example in Figure 1. The CATE is constructed to be increasing with X, but taking into account the variance, the treatment effect is clearly more pronounced for small values of X. For example, the probability of Y_1 being greater than Y_0 is much higher for smaller values of X.

Beyond the mean and variance, researchers may also be interested in other higher-moment treatment effect heterogeneity, such as Gini's mean difference or skewness, or indeed how the entire conditional densities of the control and treatment groups differ given the covariates, in an exploratory fashion. Panels (b), (c) and (d) in Figure 1 demonstrate each of the steps we propose in this paper applied to this toy dataset: hypothesis testing of equality of conditional distributions, the conditional witness function and U-statistic regression (variance, in this instance), respectively.

1.2. Related Work & Summary of Contributions

In the past few years the machine learning community has focused much effort on models for estimating the CATE function. Some approaches include Gaussian processes (Alaa & van der Schaar, 2017; 2018), Bayesian regression trees (Hill, 2011; Hahn et al., 2020), random forests (Wager & Athey, 2018), neural networks (Johansson et al., 2016; Shalit et al., 2017; Louizos et al., 2017; Atan et al., 2018; Shi et al., 2019), GANs (Yoon et al., 2018), boosting and adaptive regression splines (Powers et al., 2018) and kernel mean embeddings (Singh et al., 2020).

Distributional extensions of the ATE have been considered by many authors. Abadie (2002) tested the hypotheses of

equality and stochastic dominance of the marginal outcome distributions P_{Y_0} and P_{Y_1}, whereas Kim et al. (2018); Muandet et al. (2018) focus on estimating P_{Y_0} and P_{Y_1}, or some distance between them. These works do not consider treatment effect heterogeneity. Singh et al. (2020, Appendix C) consider CATE as well as distributional treatment effect, and while it seems that the ideas can straightforwardly be extended to conditional distributional treatment effect, it is not explicitly considered in the paper.

The CoDiTE incorporates both distributional considerations of treatment effects *and* treatment effect heterogeneity. Interest has been growing, especially in the econometrics literature, for such analyses – indeed, Bitler et al. (2017) provided concrete evidence that in some settings, the CATE does not suffice. Existing works that analyse the CoDiTE can be split into three categories, depending on how distributions are characterised: (i) quantiles, (ii) cumulative distributional functions, and (iii) specific distributional parameters, such as the mean, variance, skewness, etc. In category (i), quantile regression is a powerful tool (Koenker, 2005); however, in order to get a distributional picture via quantiles, one needs to estimate a large number of quantiles, and issues of crossing quantiles arise, whereby estimated quantiles are non-monotone. In category (ii), Chernozhukov et al. (2013; 2020) propose splitting $\mathcal{Y}$ into a grid and regressing for the cumulative distribution function at each point in the grid, but this also brings issues of non-monotonicity of the cumulative distribution function, similar to crossing quantiles. Shen (2019) estimates the cumulative distribution functions $P(Y_0 < y^*)$ and $P(Y_1 < y^*)$ for each $y^* \in \mathcal{Y}$ given each value of $X = x$ by essentially applying the Nadaraya-Watson conditional U-statistic of Stute (1991) to the U-kernel $h(y) = \mathbf{1}(y \leq y^*)$. In category (iii), generalised additive models for location, scale and shape (GAMLSS) (Stasinopoulos et al., 2017) have been applied for CoDiTE analysis (Hohberg et al., 2020; Briseño Sanchez et al., 2020), but being a parametric model, despite its flexibility, the researcher has to choose a model beforehand to proceed, and issues of model misspecification are unavoidable.

Interest has also always existed for hypothesis tests in the context of treatment effect analysis, especially in econometrics (Imbens & Wooldridge, 2009, Sections 3.3 and 5.12). Abadie (2002) tested the equality between the marginal distributions of Y_0 and Y_1, while Crump et al. (2008) tested for the equality of $\mathbb{E}[Y_1|X]$ and $\mathbb{E}[Y_0|X]$. Lee & Whang (2009); Lee (2009); Chang et al. (2015); Shen (2019) were interested, among others, in the hypothesis of the equality of $P_{Y_1|X}$ and $P_{Y_0|X}$, which we consider in Section 4.2.

Summary of Contributions We characterise distributions in two ways – first as elements in a reproducing kernel Hilbert space via kernel conditional mean embeddings, which, to the best of our knowledge, is a novel attempt in the treatment effect literature, and secondly via specific distributional parameters, as in category (iii). The former characterisation gives us a novel way of testing for the equality of conditional distributions, as well as an exploratory tool for density comparison between the groups via conditional witness functions. For the latter characterisation, we provide, to the best of our knowledge, a novel U-statistic regression technique by generalising kernel ridge regression, which, in contrast to GAMLSS, is fully nonparametric. Neither characterisation requires the estimation of a large number of quantities, unlike characterisations via quantiles or cumulative distribution functions.

2. Preliminaries

In this section, we briefly review reproducing kernel Hilbert space embeddings and U-statistics. A more complete introduction can be found in Appendix A.

2.1. Reproducing Kernel Hilbert Space Embeddings

Let $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a (scalar) positive definite kernel on $\mathcal{Y}$ with *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ (Berlinet & Thomas-Agnan, 2004, p.7, Definition 1). Given a random variable Y on $\mathcal{Y}$ satisfying $\mathbb{E}[\sqrt{l(Y,Y)}] < \infty$, the *kernel mean embedding* of Y is defined as $\mu_Y(\cdot) = \mathbb{E}[l(Y,\cdot)]$ (Smola et al., 2007, Eqn. (2a)). Given two random variables Y and Y', the *maximum mean discrepancy* (MMD) between them is defined as $\|\mu_Y - \mu_{Y'}\|_{\mathcal{H}}$ (Gretton et al., 2012, Lemma 4), where $\mu_Y - \mu_{Y'}$ is the (unnormalised) *witness function* (Gretton et al., 2012, Section 2.3; Lloyd & Ghahramani, 2015, Eqn. (3.2)). If the embedding is injective from the space of probability measures on $\mathcal{Y}$ to $\mathcal{H}$, then we say that l is *characteristic* (Fukumizu et al., 2008, Section 2.2), in which case the MMD is a proper metric. Given another random variable X on $\mathcal{X}$, the *conditional mean embedding* (CME) of Y given X is defined as $\mu_{Y|X} = \mathbb{E}[l(Y,\cdot) \mid X]$ (Park & Muandet, 2020a, Definition 3.1)[2].

Denote by $L^2(\mathcal{X}, P_X; \mathcal{H})$ the Hilbert space of (equivalence classes of) measurable functions $F : \mathcal{X} \to \mathcal{H}$ such that $\|F(\cdot)\|^2_{\mathcal{H}}$ is P_X-integrable, with inner product $\langle F_1, F_2 \rangle_2 = \int_{\mathcal{X}} \langle F_1(x), F_2(x) \rangle_{\mathcal{H}} dP_X(x)$. Given an *operator-valued kernel* $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{H})$, where $\mathcal{L}(\mathcal{H})$ is the Banach space of bounded linear operators $\mathcal{H} \to \mathcal{H}$, there exists an associated *vector-valued RKHS* of functions $\mathcal{X} \to \mathcal{H}$ (Carmeli et al., 2006, Definition 2.1, Definition 2.2, Proposition 2.3).

2.2. U-Statistics

Let $Y_1, ..., Y_r$ be independent copies of Y, and let $h : \mathcal{Y}^r \to \mathbb{R}$ be a symmetric function, i.e. for any permutation π

[2] We use the conditional expectation interpretation of the CME. An interpretation of the CME as an operator from an RKHS on $\mathcal{X}$ to $\mathcal{H}$ also exists (Song et al., 2009; 2013; Fukumizu et al., 2013).

of $(1, ..., r)$, $h(y_1, ..., y_r) = h(y_{\pi(1)}, ..., y_{\pi(r)})$, such that $h(Y_1, ..., Y_r)$ is integrable. Given i.i.d. copies $\{Y_i\}_{i=1}^n$ of Y, the *U-statistic* (Hoeffding, 1948; Serfling, 1980, p. 172) for an unbiased estimation of $\theta(P_Y) = \mathbb{E}[h(Y_1, ..., Y_r)]$ is $\hat{\theta}(Y_1, ..., Y_n) = \frac{1}{\binom{n}{r}} \sum h(Y_{i_1}, ..., Y_{i_r})$ where $\binom{n}{r}$ is the binomial coefficient and the summation is over the $\binom{n}{r}$ combinations of r distinct elements $\{i_1, ..., i_r\}$ from $\{1, ..., n\}$.

This has been extended to the conditional case (Stute, 1991). Given another random variable X on $\mathcal{X}$ and independent copies $X_1, ..., X_r$ of it, we can consider the estimation of $\theta(P_{Y|X}) = \mathbb{E}[h(Y_1, ..., Y_r)|X_1, ..., X_r]$. Stute (1991); Derumigny (2019) extend the Nadaraya-Watson regressor (Nadaraya, 1964; Watson, 1964) to estimate $\theta(P_{Y|X})$.

3. Conditional Distributional Treatment Effect

In this section, we generalise the notion of CATE to account for distributional differences between treatment and control groups, rather than just the mean difference.

Definition 3.1. Let D be some distance function between probability measures. We define the *conditional distributional treatment effect* (CoDiTE) associated with D as

$$U_D(x) = D(P_{Y_0|X=x}, P_{Y_1|X=x}).$$

Here, the choice of D depends on what characterisation of distributions is used (c.f. Section 1.2). For example, if $D(P_{Y_0|X=x}, P_{Y_1|X=x}) = \mathbb{E}[Y_1 \mid X = x] - \mathbb{E}[Y_0 \mid X = x]$, we recover the CATE, i.e. $U_D(x) = T(x)$, thereby showing that the CoDiTE is a strict generalisation of the CATE. Different choices of D will require different estimators.

The usual performance metric of a CATE estimator $\hat{T}$ is the *precision of estimating heterogeneous effects* (PEHE) (first proposed in sample form by Hill (2011, Section 4.3); we report the population-level definition, found in, for example, Alaa & Van Der Schaar (2019, Eqn. (5)):

$$\|\hat{T} - T\|_2^2 = \mathbb{E}[|\hat{T}(X) - T(X)|^2].$$

We propose a performance metric of an estimator of the CoDiTE in an exactly analogous manner.

Definition 3.2. Given a distance function D, for an estimator $\hat{U}_D$ of U_D, we define the *precision of estimating heterogeneous distributional effects* (PEHDE) as

$$\psi_D(\hat{U}_D) = \|\hat{U}_D - U_D\|_2^2 = \mathbb{E}[|\hat{U}_D(X) - U_D(X)|^2].$$

Again, if D measures the difference in expectations, then the associated PEHDE ψ_D reduces to the usual PEHE.

Henceforth, we explore different choices of the distance function D, as well as methods of estimating the corresponding CoDiTE U_D, to answer the following questions:

Q1 Are $P_{Y_0|X}$ and $P_{Y_1|X}$ different? In other words, is there any distributional effect of the treatment? (Section 4)

Q2 If so, how does the distribution of the treatment group differ from that of the control group? (Section 5)

4. CoDiTE associated with MMD via CMEs

In this section, we answer Q1, i.e. we investigate whether the treatment has any effect at all. To this end we choose D to be the MMD with the associated kernel l being characteristic. Then writing $\mu_{Y_0|X}$ and $\mu_{Y_1|X}$ for the CMEs of Y_0 and Y_1 given X respectively (c.f. Section 2.1), we have

$$\begin{aligned} U_{\text{MMD}}(x) &= \text{MMD}(P_{Y_0|X=x}, P_{Y_1|X=x}) \\ &= \|\mu_{Y_1|X=x} - \mu_{Y_0|X=x}\|_{\mathcal{H}}. \end{aligned} \tag{1}$$

Since l is characteristic, $P_{Y_0|X=x}$ and $P_{Y_1|X=x}$ are equal if and only if $\text{MMD}(P_{Y_0|X=x}, P_{Y_1|X=x}) = 0$. What makes the MMD a particularly convenient choice is that for each $x \in \mathcal{X}$, $P_{Y_0|X=x}$ and $P_{Y_1|X=x}$ are represented by individual elements $\mu_{Y_0|X=x}$ and $\mu_{Y_1|X=x}$ in the RKHS $\mathcal{H}$, which means that we can estimate the associated CoDiTE simply by performing regression with $\mathcal{X}$ as the input space and $\mathcal{H}$ as the output space, as will be shown in the next section.

4.1. Estimation and Consistency

We now discuss how to obtain empirical estimates of $U_{\text{MMD}}(x)$. Recall that, by the unconfoundedness assumption, we can estimate $\mu_{Y_0|X}$ and $\mu_{Y_1|X}$ separately from control and treatment samples respectively. We perform operator-valued kernel regression (Micchelli & Pontil, 2005; Kadri et al., 2016) in separate vector-valued RKHSs $\mathcal{G}_0$ and $\mathcal{G}_1$, endowed with kernels $\Gamma_0(\cdot, \cdot) = k_0(\cdot, \cdot)\text{Id}$ and $\Gamma_1(\cdot, \cdot) = k_1(\cdot, \cdot)\text{Id}$, where $k_0, k_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are scalar-valued kernel and $\text{Id} : \mathcal{H} \to \mathcal{H}$ is the identity operator. Following Park & Muandet (2020a, Eqn. (4)), the empirical estimates $\hat{\mu}_{Y_0|X}$ and $\hat{\mu}_{Y_1|X}$ of $\mu_{Y_0|X}$ and $\mu_{Y_1|X}$ are constructed, for each $x \in \mathcal{X}$, as

$$\begin{aligned} \hat{\mu}_{Y_0|X=x} &= \boldsymbol{k}_0^T(x)\mathbf{W}_0\boldsymbol{l}_0 \in \mathcal{G}_0 \\ \text{and} \quad \hat{\mu}_{Y_1|X=x} &= \boldsymbol{k}_1^T(x)\mathbf{W}_1\boldsymbol{l}_1 \in \mathcal{G}_1, \quad \text{where} \end{aligned} \tag{2}$$

$\mathbf{W}_0 = (\mathbf{K}_0 + n_0\lambda_{n_0}^0\mathbf{I}_{n_0})^{-1}$, $\mathbf{W}_1 = (\mathbf{K}_1 + n_1\lambda_{n_1}^1\mathbf{I}_{n_1})^{-1}$, $[\mathbf{K}_0]_{1\le i,j\le n_0} = k_0(x_i^0, x_j^0)$, $[\mathbf{K}_1]_{1\le i,j\le n_1} = k_1(x_i^1, x_j^1)$, $\lambda_{n_0}^0, \lambda_{n_1}^1 > 0$ are regularisation parameters, $\mathbf{I}_{n_0}$ and $\mathbf{I}_{n_1}$ are identity matrices, $\boldsymbol{k}_0(x) = (k_0(x_1^0, x), ..., k_0(x_{n_0}^0, x))^T$, $\boldsymbol{k}_1(x) = (k_1(x_1^1, x), ..., k_1(x_{n_1}^1, x))^T$, $\boldsymbol{l}_0 = (l(y_1^0, \cdot), ..., l(y_{n_0}^0, \cdot))^T$ and $\boldsymbol{l}_1 = (l(y_1^1, \cdot), ..., l(y_{n_1}^1, \cdot))^T$.

By plugging in the estimates (2) in the expression (1) for U_{MMD}, we can construct $\hat{U}_{\text{MMD}}$ as

$$\hat{U}_{\text{MMD}}(x) = \|\hat{\mu}_{Y_1|X=x} - \hat{\mu}_{Y_0|X=x}\|_{\mathcal{H}}.$$

The next lemma establishes a closed-form expression for $\hat{U}_{\text{MMD}}$ based on the control and treatment samples.

Algorithm 1 Kernel conditional discrepancy (KCD) test of conditional distributional treatment effect

Input: data $\{(x_i, z_i, y_i)\}_{i=1}^n$, significant level α, kernels k_0, k_1, l, regularisation parameters $\lambda_{n_0}^0, \lambda_{n_1}^1$, no. of permutations m.
Calculate $\hat{t}$ using Lemma 4.4 based on the input data.
KLR of $\{z_i\}_{i=1}^n$ against $\{x_i\}_{i=1}^n$ to obtain $\hat{e}(x_i)$.
for $k = 1$ **to** m **do**
  For each $i = 1, ..., n$, sample $\tilde{z}_i \sim \text{Bernoulli}(\hat{e}(x_i))$.
  Calculate $\hat{t}_k$ from the new dataset $\{x_i, \tilde{z}_i, y_i\}_{i=1}^n$.
end for
Calculate the p-value as $p = \frac{1+\sum_{l=1}^m \mathbf{1}\{\hat{t}_l > \hat{t}\}}{1+m}$.
if $p < \alpha$ **then**
  Reject H_0.
end if

Lemma 4.1. *For each $x \in \mathcal{X}$, we have*

$$\begin{aligned}\hat{U}^2_{\text{MMD}}(x) = \boldsymbol{k}_0^T(x)\mathbf{W}_0\mathbf{L}_0\mathbf{W}_0^T\boldsymbol{k}_0(x) \\ - 2\boldsymbol{k}_0^T(x)\mathbf{W}_0\mathbf{L}\mathbf{W}_1^T\boldsymbol{k}_1(x) \\ + \boldsymbol{k}_1^T(x)\mathbf{W}_1\mathbf{L}_1\mathbf{W}_1^T\boldsymbol{k}_1(x), \quad \textit{where}\end{aligned}$$

$[\mathbf{L}_0]_{1\le i,j\le n_0} = l(y_i^0, y_j^0)$, $[\mathbf{L}]_{1\le i\le n_0, 1\le j\le n_1} = l(y_i^0, y_j^1)$ *and* $[\mathbf{L}_1]_{1\le i,j\le n_1} = l(y_i^1, y_j^1)$.

The proof of this, and all other results, are deferred to Appendix C. The next theorem shows that, using *universal kernels* Γ_0, Γ_1 (Carmeli et al., 2010, Definition 4.1), $\hat{U}_{\text{MMD}}$ is universally consistent with respect to the PEHDE.

Theorem 4.2 (Universal consistency)**.** *Suppose that k_0, k_1 and l are bounded, that Γ_0 and Γ_1 are universal, and that $\lambda_{n_0}^0$ and $\lambda_{n_1}^1$ decay at slower rates than $\mathcal{O}(n_0^{-1/2})$ and $\mathcal{O}(n_1^{-1/2})$ respectively. Then as $n_0, n_1 \to \infty$,*

$$\psi_{\text{MMD}}(\hat{U}_{\text{MMD}}) = \mathbb{E}[(\hat{U}_{\text{MMD}}(X) - U_{\text{MMD}}(X))^2] \xrightarrow{p} 0.$$

4.2. Statistical Hypothesis Testing

We are interested in whether or not the two conditional distributions $P_{Y_0|X}$ and $P_{Y_1|X}$, corresponding to control and treatment, are equal. The hypotheses are then

H_0**:** $P_{Y_0|X=x}(\cdot) = P_{Y_1|X=x}(\cdot)$ P_X-almost everywhere.

H_1**:** There exists $A \subseteq \mathcal{X}$ with positive measure such that $P_{Y_0|X=x}(\cdot) \neq P_{Y_1|X=x}(\cdot)$ for all $x \in A$.

The null hypothesis H_0 means that the treatment has no effect for any of the covariates, whereas the alternative hypothesis H_1 means that the treatment has an effect on *some* of the covariates, where the effect is distributional. For notational simplicity, we write $P_{Y_0|X} \equiv P_{Y_1|X}$ if H_0 holds.

We use the following criterion for $P_{Y_0|X} \equiv P_{Y_1|X}$, which we call the *kernel conditional discrepancy* (KCD):

$$t = \mathbb{E}[\|\mu_{Y_1|X} - \mu_{Y_0|X}\|^2_{\mathcal{H}}].$$

The following lemma tells us that t can indeed be used as a criterion of $P_{Y_0|X} \equiv P_{Y_1|X}$.

Lemma 4.3. *If l is a characteristic kernel, $P_{Y_0|X} \equiv P_{Y_1|X}$ if and only if $t = 0$.*

Next, we define a plug-in estimate $\hat{t}$ of t, which we will use as the test statistic of our hypothesis test:

$$\hat{t} = \frac{1}{n}\sum_{i=1}^n \left\|\hat{\mu}_{Y_1|X=x_i} - \hat{\mu}_{Y_0|X=x_i}\right\|^2_{\mathcal{H}}.$$

Then we have a closed-form expression for $\hat{t}$ as follows.

Lemma 4.4. *We have*

$$\begin{aligned}\hat{t} = \frac{1}{n}\text{Tr}\left(\tilde{\mathbf{K}}_0\mathbf{W}_0\mathbf{L}_0\mathbf{W}_0^T\tilde{\mathbf{K}}_0^T\right) \\ - \frac{2}{n}\text{Tr}\left(\tilde{\mathbf{K}}_0\mathbf{W}_0\mathbf{L}\mathbf{W}_1^T\tilde{\mathbf{K}}_1^T\right) \\ + \frac{1}{n}\text{Tr}\left(\tilde{\mathbf{K}}_1\mathbf{W}_1\mathbf{L}_1\mathbf{W}_1^T\tilde{\mathbf{K}}_1^T\right),\end{aligned}$$

where $\mathbf{L}_0, \mathbf{L}_1$ and $\mathbf{L}$ are as defined in Lemma 4.1 and $[\tilde{\mathbf{K}}_0]_{1\le i\le n, 1\le j\le n_0} = k_0(x_i, x_j^0)$ and $[\tilde{\mathbf{K}}_1]_{1\le i\le n, 1\le j\le n_1} = k_1(x_i, x_j^1)$.

The consistency of $\hat{t}$ in the limit of infinite data is shown in the following theorem.

Theorem 4.5. *Under the same assumptions as in Theorem 4.2, we have $\hat{t} \xrightarrow{p} t$ as $n_0, n_1 \to \infty$.*

Unfortunately, it is extremely difficult to compute the (asymptotic) null distribution of $\hat{t}$ analytically, and so we resort to resampling the treatment labels to simulate the null distribution. To ensure that our resampling scheme respects the control and treatment covariate distributions $P_{X|Z=0}$ and $P_{X|Z=1}$, we follow the conditional resampling scheme of Rosenbaum (1984). We first estimate the propensity score $e(x_i)$ for each datapoint x_i (e.g. using kernel logistic regression (KLR) (Zhu & Hastie, 2005; Marteau-Ferey et al., 2019)), and then resample each data label from this estimated propensity score. By repeating this resampling procedure and computing the test statistic on each resampled dataset, we can simulate from the null distribution of the test statistic. Finally, the test statistic computed from the original dataset is compared to this simulated null distribution, and the null hypothesis is rejected or not rejected accordingly. The exact procedure is summarised in Algorithm 1.

5. Understanding the CoDiTE

After determining *whether* $P_{Y_0|X}$ and $P_{Y_1|X}$ are different via MMD-associated CoDiTE and hypothesis testing, we now turn to Q2, i.e. we investigate *how* they are different.

5.1. Conditional Witness Functions

For two real-valued random variables, the witness function between them is a useful tool for visualising where their densities differ, without explicitly estimating the densities (Gretton et al., 2012, Figure 1; Lloyd & Ghahramani, 2015, Figure 1). We extend this to the conditional case with the (unnormalised) *conditional witness function* $\mu_{Y_1|X} - \mu_{Y_0|X}$.

Let us fix $x \in \mathcal{X}$. The witness function between $P_{Y_1|X=x}$ and $P_{Y_0|X=x}$ is $\mu_{Y_1|X=x} - \mu_{Y_0|X=x} : \mathcal{Y} \to \mathbb{R}$. For $y \in \mathcal{Y}$ in regions where the density of $P_{Y_1|X=x}$ is greater than that of $P_{Y_0|X=x}$, we have $\mu_{Y_1|X=x}(y) - \mu_{Y_0|X=x}(y) > 0$. For y in regions where the converse is true, we similarly have $\mu_{Y_1|X=x}(y) - \mu_{Y_0|X=x}(y) < 0$. The greater the difference in density, the greater the magnitude of the witness function. For each $y \in \mathcal{Y}$, the associated CoDiTE is

$$U_{\text{witness},y}(x) = \mu_{Y_1|X=x}(y) - \mu_{Y_0|X=x}(y).$$

The estimates in (2) can be plugged in to obtain the estimate $\hat{U}_{\text{witness},y} = \hat{\mu}_{Y_1|X=x}(y) - \hat{\mu}_{Y_0|X=x}(y)$. Since convergence in the RKHS norm implies pointwise convergence (Berlinet & Thomas-Agnan, 2004, p.10, Corollary 1), Theorem 4.2 implies the consistency of $\hat{U}_{\text{witness},y}$ with respect to the corresponding PEHDE. Clearly, if X is more than 1-dimensional, heat maps as in Figure 1(c) cannot be plotted; however, fixing a particular $x \in \mathcal{X}$, $\hat{\mu}_{Y_1|X=x} - \hat{\mu}_{Y_0|X=x}$ can be plotted against y, since $Y \subseteq \mathbb{R}$. Such plots will be informative of where the density of $P_{Y_1|X=x}$ is greater than that of $P_{Y_0|X=x}$ and vice versa.

5.2. CoDiTE associated with Specific Distributional Quantities via U-statistic Regression

Next, we consider CoDiTE on specific distributional quantities, such as the mean, variance or skewness, or some function thereof. For example, Briseño Sanchez et al. (2020, Eqn. (2)) were interested, in addition to the CATE, in the treatment effect on the standard deviation $U_D(x) = \text{std}(Y_1|X = x) - \text{std}(Y_0|X = x)$. Our motivating example in Figure 1 could inspire a "standardised" version of the CATE[3]:

$$U_D(x) = \frac{\mathbb{E}[Y_1|X = x] - \mathbb{E}[Y_0|X = x]}{\sqrt{\text{Var}(Y_1|X = x) + \text{Var}(Y_0|X = x)}}. \quad (3)$$

Many of these quantities can be represented as the expectation of a U-kernel, i.e. $\mathbb{E}[h(Y_1, ..., Y_r)]$ (c.f. Section 2.2). For example, $h(y) = y$ gives the mean, $h(y_1, y_2) = \frac{1}{2}(y_1 - y_2)^2$ gives the variance and $h(y_1, y_2) = |y_1 - y_2|$ gives Gini's mean difference. We consider their conditional counterparts, i.e. $\theta(P_{Y_0|X}) = \mathbb{E}[h(Y_{01}, ..., Y_{0r})|X_1, ..., X_r]$ and $\theta(P_{Y_1|X}) = \mathbb{E}[h(Y_{11}, ..., Y_{1r})|X_1, ..., X_r]$ (c.f. Section 2.2). By Çınlar (2011, p.146, Theorem 1.17), there exist functions $F_0, F_1 : \mathcal{X}^r \to \mathbb{R}$ such that $F_0(X_1, ..., X_r) = \theta(P_{Y_0|X})$ and $F_1(X_1, ..., X_r) = \theta(P_{Y_1|X})$.

Estimation of F_0 and F_1 can be done via U-statistic regression, by generalising kernel ridge regression as follows. As in Section 4.1, let $k_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}_0$. Then if we define $k_0^r : \mathcal{X}^r \times \mathcal{X}^r \to \mathbb{R}$ as

$$k_0^r((x_1, ..., x_r), (x'_1, ..., x'_r)) = k_0(x_1, x'_1)...k_0(x_r, x'_r),$$

Berlinet & Thomas-Agnan (2004, p.31, Theorem 13) tells us that k_0^r is a reproducing kernel on $\mathcal{X}^r$ with RKHS $\mathcal{H}_0^r = \mathcal{H}_0 \otimes ... \otimes \mathcal{H}_0$, the r-times tensor product of $\mathcal{H}_0$, whose elements are functions $\mathcal{X}^r \to \mathbb{R}$. We estimate F_0 in $\mathcal{H}_0^r$. Given any $F \in \mathcal{H}_0^r$, the natural least-squares risk is

$$\mathcal{E}(F) = \mathbb{E}[(F(X_1, ..., X_r) - h(Y_{01}, ..., Y_{0r}))^2].$$

Recalling the control sample $\{(x_i^0, y_i^0)\}_{i=1}^{n_0}$, we solve the following regularised least-squares problem:

$$\hat{F}_0 = \underset{F \in \mathcal{H}_0^r}{\arg\min} \left\{ \hat{\mathcal{E}}(F) + \lambda_{n_0}^0 \|F\|_{\mathcal{H}_0^r}^2 \right\} \quad (4)$$

where the empirical least-squares risk $\hat{\mathcal{E}}$ is defined as

$$\hat{\mathcal{E}}(F) = \frac{1}{\binom{n_0}{r}} \sum \left(F(x_{i_1}^0, ..., x_{i_r}^0) - h(y_{i_1}^0, ..., y_{i_r}^0) \right)^2,$$

with the summation over the $\binom{n_0}{r}$ combinations of r distinct elements $\{i_1, ..., i_r\}$ from $\{1, ..., n_0\}$. Note that $\hat{\mathcal{E}}(F)$ is itself a U-statistic for the estimation of $\mathcal{E}(F)$. The following is a representer theorem for the problem in (4).

Theorem 5.1. *The solution $\hat{F}_0$ to the problem in (4) is*

$$\hat{F}_0(x_1, ..., x_r) = \sum_{i_1, ..., i_r}^{n_0} k_0(x_{i_1}^0, x_1)...k_0(x_{i_r}^0, x_r) c_{i_1, ..., i_r}^0$$

where the coefficients $c_{i_1, ..., i_r}^0 \in \mathbb{R}$ are the unique solution of the n^r linear equations,

$$\sum_{j_1, ..., j_r = 1}^{n_0} \left(k_0\left(x_{i_1}^0, x_{j_1}^0\right) ... k_0\left(x_{i_r}^0, x_{j_r}^0\right) + \binom{n_0}{r} \lambda_{n_0}^0 \delta_{i_1 j_1} ... \delta_{i_r j_r} \right) c_{j_1, ..., j_r}^0 = h\left(y_{i_1}^0, ..., y_{i_r}^0\right).$$

Note that if $r = 1$ and $h(y) = y$, we recover the usual kernel ridge regression. The following result shows that this estimation procedure is universally consistent.

[3] In practice, if the CoDiTE involves ratios of estimated quantities, we do not recommend plugging in the estimates directly into the ratio, since, if the denominator is small, then a small error in the estimation of the denominator will result in a large error in the overall CoDiTE estimation. Instead, we recommend that the practitioner estimate the numerator and the denominator separately and interpret the results directly from the raw estimates.

Table 1. Root mean square error in estimating the conditional standard deviation, with standard error from 100 simulations, for GAMLSS (implemented via the R package `gamlss` (Rigby & Stasinopoulos, 2005)) and our U-statistic regression via generalised kernel ridge regression (U-regression KRR; implemented via the Falkon library on Python (Rudi et al., 2017; Meanti et al., 2020)). Lower is better.

Method	**Setting SN**		**Setting LN**		**Setting HN**	
	Control	Treatment	Control	Treatment	Control	Treatment
GAMLSS	0.17 ± 0.031	0.767 ± 0.414	3.3 ± 0.55	15.44 ± 8.128	2.27 ± 0.44	10.91 ± 5.42
U-regression KRR	$\mathbf{0.13 \pm 0.059}$	$\mathbf{0.16 \pm 0.059}$	$\mathbf{1.1 \pm 0.31}$	$\mathbf{2.16 \pm 0.61}$	$\mathbf{0.7 \pm 0.25}$	$\mathbf{1.39 \pm 0.47}$

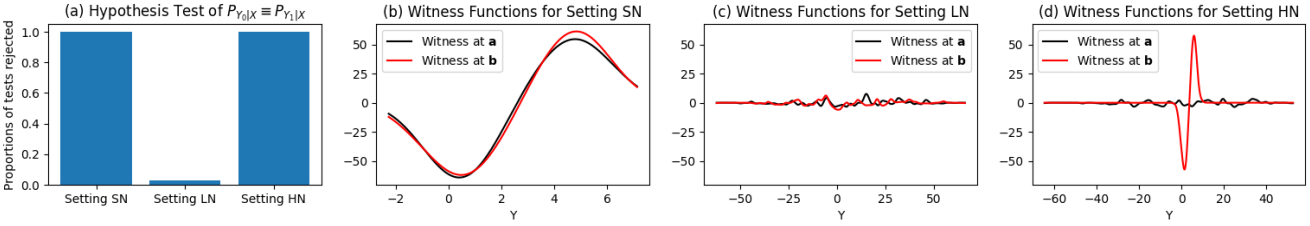

Figure 2. **Hypothesis testing and witness functions on the IHDP dataset.** (a) Hypothesis test is conducted on 100 simulations for each setting, with the bar chart showing proportion of tests rejected for each setting. In setting "LN", where the variance overwhelms the CATE, the test does not reject the hypothesis $P_{Y_0|X} \equiv P_{Y_1|X}$, whereas in the other two settings, the hypothesis is rejected. (b) At both $X = \mathbf{a}$ and $X = \mathbf{b}$, the density of the control group is larger than that of the treatment group around $Y = 0$, and the reverse is true around $Y = 4$, showing the marked effect of the treatment. (c) At both $X = \mathbf{a}$ and $X = \mathbf{b}$, the density of the control and treatment groups are roughly equal for all Y. (d) At $X = \mathbf{a}$, where the variance engulfs the CATE, the density of the control and treatment groups are roughly equal for all Y, whereas at $X = \mathbf{b}$, the witness function clearly shows where the density of one group dominates the other. The juxtaposition of witness functions at different points in the covariate space is an exploratory tool to compare the relative strength of the treatment effect.

Theorem 5.2. *Suppose k_0^r is a bounded and universal kernel and that $\lambda_{n_0}^0$ decays at a slower rate than $\mathcal{O}(n_0^{-1/2})$. Then as $n_0 \to \infty$,*

$$\mathbb{E}\left[\left(\hat{F}_0\left(X_1, ..., X_r\right) - F_0\left(X_1, ..., X_r\right)\right)^2\right] \xrightarrow{p} 0.$$

A consistent estimate $\hat{F}_1$ of F_1 is obtained by exactly the same procedure, using the treatment sample $\{(x_i^1, y_i^1)\}_{i=1}^{n_1}$.

6. Experiments

6.1. Semi-synthetic IHDP Data

We demonstrate the use of our methods on the Infant Health and Development Program (IHDP) dataset (Hill, 2011, Section 4). The covariates are taken from a randomised control trial, from which a non-random portion is removed to imitate an observational study. The reason for its popularity in the CATE literature is that, for each datapoint, the outcome is simulated for both treatment and control, enabling cross-validation and evaluation, which is usually not possible in observational studies due to the missing counterfactuals. Existing works first define the noiseless response surfaces for the control and treatment groups, and generate realisations of the potential outcomes by applying Gaussian noise with constant variance across the whole dataset.

This last assumption of constant variance is somewhat unrealistic, but of little importance in evaluating CATE estimators. In our experiments, we modify the data generating process in three different ways, all of which have the same parallel linear mean response surfaces, with the CATE of 4 ("response surface A" in Hill (2011)). In setting "SN" ("small noise"), the standard deviation of the noise is constant at 1, so that the CATE of 4 translates to a meaningful treatment effect. In setting "LN" ("large noise"), the standard deviation of the noise is constant at 20, meaning that the mean difference in the response surfaces is negligible in comparison. In this case, our test does not reject the hypothesis that the two conditional distributions are the same, and there is no case for further investigation (see middle bar in Figure 2(a)). In setting "HN" ("heterogeneous noise"), the standard deviation is heterogeneous across the dataset, so that the standard deviation is 1 for some data points while others have standard deviation of 20. The exact data generating process is detailed in Appendix B.

In setting "HN", let us consider points $\mathbf{a}, \mathbf{b} \in \mathcal{X}$ with $\mathrm{sd}(Y|X = \mathbf{a}) = 20$ and $\mathrm{sd}(Y|X = \mathbf{b}) = 1$. Then even though the CATE at $\mathbf{a}$ and $\mathbf{b}$ are equal at 4, we have $\mathrm{std}(Y_1 - Y_0|X = \mathbf{a}) \gg \mathrm{std}(Y_1 - Y_0|X = \mathbf{b})$, such that there is a pronounced treatment effect at $\mathbf{b}$, while the variance engulfs the treatment effect at $\mathbf{a}$. The comparative magnitudes of the witness functions conditioned on $\mathbf{a}$ and $\mathbf{b}$ confirm this heterogeneity (see Figure 2(d)). In Table 1, the quality of estimation of the standard deviation via our U-statistic

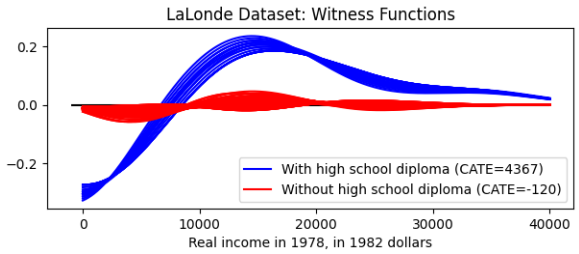

Figure 3. **Witness functions for Black, unmarried participant up to the age of 25, unemployed in both 1974 and 1975.** Each curve (witness function) corresponds to an individual in this subset.

regression is compared with GAMLSS (Stasinopoulos et al., 2017) estimation for each setting.

An immediate benefit is a better understanding of the treatment. Even a perfect CATE estimator cannot capture such heterogeneity in distributional treatment effect (variance, in this case). As argued in Section 1.1, any method that involves comparing mean values (of which CATE is one) should also take into account the variance for it to be meaningful. This will give a clearer picture of the subpopulations on which there is a marked treatment effect, and those on which it is weaker, than relying on the CATE alone. Such knowledge should in turn influence policy decisions, in terms of which subpopulations should be targeted. We note that recently Jesson et al. (2020) considered CATE uncertainty in IHDP in the context of a different task: making or deferring treatment recommendations while using Bayesian neural networks, focusing on cases where overlap fails or under covariate shift; however, distributional considerations can be important even when overlap is satisfied and no covariate shift takes place.

6.2. Real Outcomes: LaLonde Data

In this section, we apply the proposed methods to LaLonde's well-known National Supported Work (NSW) dataset (LaLonde, 1986; Dehejia & Wahba, 1999) which has been used widely to evaluate estimators of treatment effects. The outcome of interest Y is the real earnings in 1978, with treatment Z being the job training. We refer the interested readers to Dehejia & Wahba (1999, Sec. 2.1) for a detailed description of the dataset. As income distributions are known to be skewed to the right, it may be interesting to investigate not only the CATE, but the entire distributions.

The test rejects the hypothesis $P_{Y_0|X} \equiv P_{Y_1|X}$ with p-value of 0.013. As a demonstration of the kind of exploratory analysis that can be conducted using the conditional witness functions, we focus our attention on a subset of the data on which the overlap condition is satisfied – Black, unmarried participants up to the age of 25, who were unemployed in both 1974 and 1975. Figure 3 shows the witness function for each individual in this subset, with the colour of the curve delineating whether the corresponding individual has a high school diploma.

We can see clearly that for those without a high school diploma, the treatment effect is not so pronounced, whereas there is a marked treatment effect for those with it. Negative values of the witness function for small income values mean that we are more likely to get small income values from the control group than the treatment group, whereas larger income values are more likely to come from the treatment group, as indicated by the positive values of the witness functions. In particular, the tail of the blue curves to the right implies a skewness of the density of the treated group relative to the control group, and the treatment group continues to have larger density than the control group for high income values (> 25000), albeit to a lesser extent. Such comparison of densities in different regions of $\mathcal{Y}$ is not possible with the CATE, which is a simple difference of the means between the control and treated groups.

7. Discussion & Conclusion

In this paper, we discussed the analysis of the conditional distributional treatment effect (CoDiTE). We first propose a new kernel-based hypothesis test via kernel conditional mean embeddings to see whether there exists any CoDiTE. Then we proceeded to investigate the nature of the treatment effect via conditional witness functions, revealing where and how much the conditional densities differ, and U-statistic regression, which is informative about the differences in specific conditional distributional quantities.

We foresee that much of the work that has been done by the machine learning community on treatment effect analysis, although cast mostly in the context of CATE, applies for the CoDiTE. Examples include *meta learners* (Künzel et al., 2019), model validation (Alaa & Van Der Schaar, 2019), subgroup analysis (Su et al., 2009; Lee et al., 2020) and covariate balancing (Gretton et al., 2009; Kallus, 2018). overo A major obstacle in any covariate-conditional analysis of treatment effect is this: when the covariate space is high-dimensional, the accuracy and reliability of the estimates deteriorate significantly due to the curse of dimensionality, and we heavily rely on changes to be smooth across the covariate space. This limitation is present not only in methods presented in this paper, but any CATE or CoDiTE analysis. While out of scope for the present paper, it is of interest to investigate how to mitigate this problem.

Last but not least, we argue that the conditional distributional treatment effect can play an important role in making fair and explainable decisions as it provides a more complete picture of the treatment effect. On the one hand, policymakers can use tools that we develop to identify the groups of

individuals for which the outcome distributions differ most through the effect modifiers. On the other hand, the presence of effect modification that is associated with sensitive attributes such as race, ethnicity, and gender creates challenges for decision makers. If they knew that there is effect modification by race, for example, certain groups of individuals may be treated unfairly. In practice, our tools can potentially be used to detect the discrepancy between outcome distributions conditioned on these sensitive attributes, which is also an interesting avenue for future work.

Acknowledgements

We are very grateful to Jonas Kübler at the MPI for Intelligent Systems, Tübingen, for readily providing help with running the codes, and for other insightful discussions. We also thank Alexis Derumigny at Delft University of Technology and Giacomo Meanti at Università degli Studi di Genova for readily answering queries via email and helpful discussions. Finally, we thank Arthur Gretton for his comments on the early draft of our manuscript.

This work was in part supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B, and by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. Uri Shalit was supported by the Israel Science Foundation (grant No. 1950/19).

References

Abadie, A. Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. *Journal of the American statistical Association*, 97(457):284–292, 2002.

Alaa, A. and Schaar, M. Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design. In *International Conference on Machine Learning*, pp. 129–138, 2018.

Alaa, A. and Van Der Schaar, M. Validating Causal Inference Models via Influence Functions. In *International Conference on Machine Learning*, pp. 191–201, 2019.

Alaa, A. M. and van der Schaar, M. Bayesian Inference of Individualized Treatment Effects using Multi-Task Gaussian Processes. In *Advances in Neural Information Processing Systems*, pp. 3424–3432, 2017.

Alaa, A. M. and van der Schaar, M. Bayesian Nonparametric Causal Inference: Information Rates and Learning Algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1031–1046, 2018.

Aronszajn, N. Theory of Reproducing Kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Atan, O., Jordon, J., and van der Schaar, M. Deep-Treat: Learning Optimal Personalized Treatments from Observational Data using Neural Networks. In *AAAI*, pp. 2071–2078, 2018.

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

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment. *Review of Economics and Statistics*, 99(4): 683–697, 2017.

Briseño Sanchez, G., Hohberg, M., Groll, A., and Kneib, T. Flexible Instrumental Variable Distributional Regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1553–1574, 2020.

Carmeli, C., De Vito, E., and Toigo, A. Vector Valued Reproducing Kernel Hilbert Spaces of Integrable Functions and Mercer Theorem. *Analysis and Applications*, 4(04): 377–408, 2006.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

Chang, M., Lee, S., and Whang, Y.-J. Nonparametric Tests of Conditional Treatment Effects with an Application to Single-Sex Schooling on Academic Achievements. *The Econometrics Journal*, 18(3):307–346, 2015.

Chernozhukov, V., Fernández-Val, I., and Melly, B. Inference on Counterfactual Distributions. *Econometrica*, 81 (6):2205–2268, 2013.

Chernozhukov, V., Fernandez-Val, I., and Weidner, M. Network and Panel Quantile Effects via Distribution Regression. *Journal of Econometrics*, 2020.

Çınlar, E. *Probability and Stochastics*, volume 261. Springer Science & Business Media, 2011.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. Nonparametric Tests for Treatment Effect Heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

Dehejia, R. H. and Wahba, S. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

Derumigny, A. Estimation of a Regular Conditional Functional by Conditional U-Statistics Regression. *arXiv preprint arXiv:1903.10914*, 2019.

Dinculeanu, N. *Vector Integration and Stochastic Integration in Banach Spaces*, volume 48. John Wiley & Sons, 2000.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. Subgroup Identification from Randomized Clinical Trial Data. *Statistics in medicine*, 30(24):2867–2880, 2011.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel Measures of Conditional Dependence. In *Advances in neural information processing systems*, pp. 489–496, 2008.

Fukumizu, K., Song, L., and Gretton, A. Kernel Bayes' Rule: Bayesian Inference with Positive Definite Kernels. *The Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate Shift by Kernel Mean Matching. *Dataset shift in machine learning*, 3(4): 5, 2009.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayeisan Regression Tree Models for Causal Inference: Regularisation, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, 15(3):965–1056, 09 2020.

Hill, J. L. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hoeffding, W. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, pp. 293–325, 1948.

Hohberg, M., Pütz, P., and Kneib, T. Treatment Effects Beyond the Mean Using Distributional Regression: Methods and Guidance. *Plos one*, 15(2):e0226514, 2020.

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

Imbens, G. W. and Rubin, D. B. *Causal Inference in Statistics, Social, and Biomedical sciences*. Cambridge University Press, 2015.

Imbens, G. W. and Wooldridge, J. M. Recent Developments in the Econometrics of Program Evaluation. *Journal of economic literature*, 47(1):5–86, 2009.

Jesson, A., Mindermann, S., Shalit, U., and Gal, Y. Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models. *Advances in Neural Information Processing Systems*, 33, 2020.

Johansson, F., Shalit, U., and Sontag, D. Learning Representations for Counterfactual Inference. In *International conference on machine learning*, pp. 3020–3029, 2016.

Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., and Audiffren, J. Operator-Valued Kernels for Learning from Functional Response Data. *The Journal of Machine Learning Research*, 17(1):613–666, 2016.

Kallus, N. Optimal A Priori Balance in the Design of Controlled Experiments. *Journal of the Royal Statistical Society Series B*, 80(1):85–112, 2018.

Kim, K., Kim, J., and Kennedy, E. H. Causal Effects Based on Distributional Distances. *arXiv preprint arXiv:1806.02935*, 2018.

Koenker, R. *Quantile Regression*. Cambridge University Press, 2005.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

Lee, H.-S., Zhang, Y., Zame, W., Shen, C., Lee, J.-W., and van der Schaar, M. Robust Recursive Partitioning for Heterogeneous Treatment Effects with Uncertainty Quantification. *Advances in Neural Information Processing Systems*, 33, 2020.

Lee, M.-J. Non-parametric Tests for Distributional Treatment Effect for Randomly Censored Responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):243–264, 2009.

Lee, S. S. and Whang, Y.-J. Nonparametric Tests of Conditional Treatment Effects. Technical report, Cowles Foundation for Research in Economics, Yale University, 2009.

Lloyd, J. R. and Ghahramani, Z. Statistical Model Criticism using Kernel Two Sample Tests. *Advances in Neural Information Processing Systems*, 28:829–837, 2015.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal Effect Inference with Deep Latent-Variable Models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Marteau-Ferey, U., Bach, F., and Rudi, A. Globally Convergent Newton Methods for Ill-Conditioned Generalized Self-Concordant Losses. In *Advances in Neural Information Processing Systems*, 2019.

Meanti, G., Carratino, L., Rosasco, L., and Rudi, A. Kernel Methods Through the Roof: Handling Billions of Points Efficiently. *Advances in Neural Information Processing Systems*, 33, 2020.

Micchelli, C. A. and Pontil, M. On Learning Vector-Valued Functions. *Neural computation*, 17(1):177–204, 2005.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Muandet, K., Kanagawa, M., Saengkyongam, S., and Marukatat, S. Counterfactual Mean Embedding. *arXiv preprint arXiv:1805.08845*, 2018.

Nadaraya, E. A. On Estimating Regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

Park, J. and Muandet, K. A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings. In *Advances in Neural Information Processing Systems*, 2020a.

Park, J. and Muandet, K. Regularised Least-Squares Regression with Infinite-Dimensional Output Space. *arXiv preprint arXiv:2010.10973*, 2020b.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. Some Methods for Heterogeneous Treatment Effect Estimation in High Dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.

Rigby, R. A. and Stasinopoulos, D. M. Generalized Additive Models for Location, Scale and Shape,(with discussion). *Applied Statistics*, 54:507–554, 2005.

Rosenbaum, P. R. Conditional Permutation Tests and the Propensity Score in Observational Studies. *Journal of the American Statistical Association*, 79(387):565–574, 1984.

Rosenbaum, P. R. *Observational Studies*. Springer Science & Business Media, 2002.

Rosenbaum, P. R. and Rubin, D. B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D. B. Causal Inference using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Rudi, A., Carratino, L., and Rosasco, L. Falkon: An Optimal Large Scale Kernel Method. In *Advances in Neural Information Processing Systems*, pp. 3888–3898, 2017.

Scholkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2001.

Schwabik, S. and Ye, G. *Topics in Banach Space Integration*, volume 10. World Scientific, 2005.

Serfling, R. J. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Shen, S. Estimation and Inference of Distributional Partial Effects: Theory and Application. *Journal of Business & Economic Statistics*, 37(1):54–66, 2019.

Shi, C., Blei, D., and Veitch, V. Adapting Neural Networks for the Estimation of Treatment Effects. In *Advances in Neural Information Processing Systems*, pp. 2507–2517, 2019.

Simon-Gabriel, C.-J. and Schölkopf, B. Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions. *The Journal of Machine Learning Research*, 19(1):1708–1736, 2018.

Singh, R., Xu, L., and Gretton, A. Kernel Methods for Policy Evaluation: Treatment Effects, Mediation Analysis, and Off-Policy Planning. *arXiv preprint arXiv:2010.04855*, 2020.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert Space Embedding for Distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Song, L., Huang, J., Smola, A., and Fukumizu, K. Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 961–968, 2009.

Song, L., Fukumizu, K., and Gretton, A. Kernel Embeddings of Conditional Distributions: A Unified Kernel Framework for Nonparametric Inference in Graphical Models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. *Flexible Regression and Smoothing: Using GAMLSS in R*. CRC Press, 2017.

Stute, W. Conditional U-Statistics. *The Annals of Probability*, 19(2):812–825, 1991.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. Subgroup Analysis via Recursive Partitioning. *Journal of Machine Learning Research*, 10(2), 2009.

Wager, S. and Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018.

Watson, G. S. Smooth Regression Analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372, 1964.

Yoon, J., Jordon, J., and van der Schaar, M. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In *International Conference on Learning Representations*, 2018.

Zhu, J. and Hastie, T. Kernel Logistic Regression and the Import Vector Machine. *Journal of Computational and Graphical Statistics*, 14(1):185–205, 2005.