Conditional Outcome Equivalence: A Quantile Alternative to CATE

Josh Givens
University of Bristol
josh.givens@bristol.ac.uk

Henry W J Reeve
University of Bristol
henry.reeve@bristol.ac.uk

Song Liu
University of Bristol
song.liu@bristol.ac.uk

Katarzyna Reluga
University of Bristol
katarzyna.reluga@bristol.ac.uk

Abstract

The conditional quantile treatment effect (CQTE) can provide insight into the effect of a treatment beyond the conditional average treatment effect (CATE). This ability to provide information over multiple quantiles of the response makes the CQTE especially valuable in cases where the effect of a treatment is not well-modelled by a location shift, even conditionally on the covariates. Nevertheless, the estimation of the CQTE is challenging and often depends upon the smoothness of the individual quantiles as a function of the covariates rather than smoothness of the CQTE itself. This is in stark contrast to the CATE where it is possible to obtain high-quality estimates which have less dependency upon the smoothness of the nuisance parameters when the CATE itself is smooth. Moreover, relative smoothness of the CQTE lacks the interpretability of smoothness of the CATE making it less clear whether it is a reasonable assumption to make. We combine the desirable properties of the CATE and CQTE by considering a new estimand, the conditional quantile comparator (CQC). The CQC not only retains information about the whole treatment distribution, similar to the CQTE, but also having more natural examples of smoothness and is able to leverage simplicity in an auxiliary estimand. We provide finite sample bounds on the error of our estimator, demonstrating its ability to exploit simplicity. We validate our theory in numerical simulations which show that our method produces more accurate estimates than baselines. Finally, we apply our methodology to a study on the effect of employment incentives on earnings across different age groups. We see that our method is able to reveal heterogeneity of the effect across different quantiles.

1 Introduction

In many real world scenarios such as personalised treatment allocation and individual level policy decisions, understanding the effect of a treatment/intervention at an individual level is invaluable in providing bespoke care. The field which aims to understand a treatment's effect given certain covariates is referred to as heterogeneous treatment effect (HTE) estimation and has seen popularity across many applications [11, 10, 24, 20]. Within HTE, the conditional average treatment effect (CATE) has proved itself to be a popular target of study in this area due to its simplicity and interpretability [2, 13, 28]. A key limitation of the CATE however, is that it fails to paint a full picture of the differences between distributions of the two responses. In addition, it can be sensitive to outliers, with extreme values leading to a biased outcome. As such, the conditional quantile treatment effect (CQTE), an estimand which compares the conditional quantiles of the distributions in the treated and untreated populations, has established itself as a popular alternative [1, 5, 26].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3260

While the CQTE offers more information than the CATE and is more robust to outliers, it lacks some of the CATE's desirable estimation properties. Specifically, CQTE estimation involves estimating the quantile functions for the two marginal outcomes. This harms the estimation procedure in cases where estimation of marginal quantile functions is more challenging than estimation of the CQTE itself. An example of this is when the marginal quantile functions are less smooth as a function of the covariates than the CQTE. This aligns with a recurring idea within HTE estimation that the effect of a treatment may be simpler than the marginal outcomes. In contrast to the CQTE, there are many CATE estimators which aim to learn the CATE directly allowing them to exploit its relative simplicity. These include the X-learner [17], R-learner [23], and Doubly Robust (DR) learner [15]. Before estimating the CATE, these procedures require estimation of intermediary estimands (nuisance parameters) which condition on the covariates such as the average marginal outcomes and the propensity score (the probability of being assigned to treatment group). These nuisance parameters are then used to aid the estimation of the CATE. With the DR learner specifically, it has been shown that it can still achieve optimal convergence rates even when estimation of the nuisance parameters is worse than estimation of the CATE itself. This notion is referred to as **double robustness**, as our estimation is robust to sub-optimal estimation of both of the nuisance parameters.

Some attempts have been made to improve CQTE estimation [34, 33] with a key work being that of Kallus and Oprescu [14]. In this they provide an extension of the double robustness property to the CQTE, creating an estimation procedure that can achieve strong convergence even when nuisance parameters are more difficult to estimate. Unfortunately, one of the nuisance parameters which must be estimated is the reciprocal of the conditional densities over the response. These are highly difficult to estimate and risk the errors blowing up in low density regions which could potentially nullify the desirable estimation rates they achieve even with the dependence on the estimation accuracy of these nuisance parameters being less strong. Furthermore it is still unclear how one can interpret relative smoothness in the CQTE compared to the individual quantiles with their being relatively little discussion of this within the literature. In general there is a distinct lack of illustrative examples; which are present for the CATE. To our knowledge no other works specifically aim to tackle this double robustness phenomenon for the CQTE or other quantile based treatment effect estimands.

We introduce a novel estimand called the "conditional quantile comparator" (CQC). The CQC gives the outcome for a treated individual in the same quantile as a given outcome for an untreated individual, conditional on covariates. This relates to the conditional Quantile-Quantile (QQ) plot for the treated and untreated outcomes as demonstrated in Figure 1. Similarly to the CQTE, our new esti-

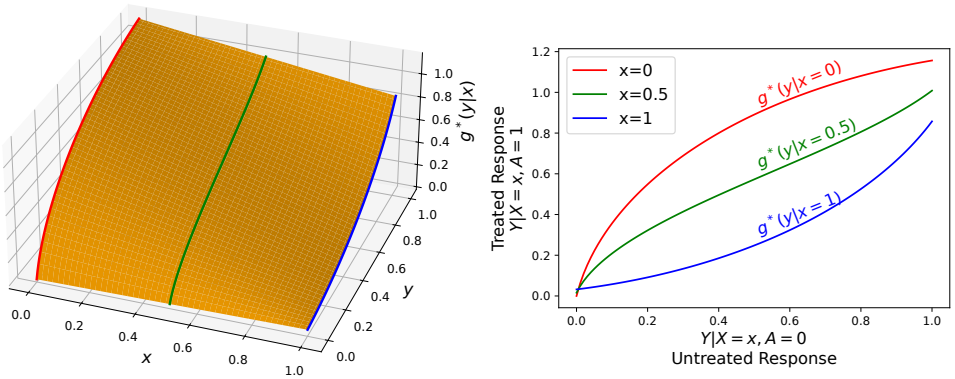


Figure 1: The left plot gives the CQC surface which takes in covariates (x) and an untreated response (y) and returns the treated response of the equivalent quantile ($g^*(y|x)$). The right plot is a QQ-plot of the responses (Y) in the untreated ($A = 0$) vs treated ($A = 1$) population conditional on various covariates ($X = 0, 0.5, 1$). These conditional QQ-plots correspond to "slices" of the CQC surface, as shown by the coloured lines in the left plot. The plot is best viewed in colour.

mand, the CQC, allows us to compare equivalent quantiles while working exclusively in the response landscape, making it a more interpretable tool. This allows us to construct canonical examples of the CQC being smoother than various nuisance parameters, the CQTE, and the CATE; adding to this interpretability. In addition, using the pseudo-outcome framework presented in Kennedy [15], we can leverage CATE estimation procedures to estimate the CQC in a doubly robust way, as mentioned

above. Crucially, the CQC can keep the valuable quantile-level information previously offered by the CQTE while building on much of the CATE literature to acquire its desirable robustness properties and interpretability. Our contributions are as follows:

- Introduce a new estimand for HTE analysis: the conditional quantile comparator (CQC).
- Propose an estimation procedure which we prove to be doubly robust.
- Demonstrate better estimation accuracy especially when the CQC is smooth but individual conditional cumulative distribution functions are not.
- Provide insights into real-world datasets on employment intervention and medical treatment.

2 Set-up

We now introduce the standard HTE set-up in our notation. Let Z denote the random triple (Y, X, A) with Y a random variable (RV) on $\mathcal{Y} \subseteq \mathbb{R}$, X a RV on $\mathcal{X} \subseteq \mathbb{R}^d$, and A a RV on $\{0, 1\}$. We treat Z as representing an individual and interpret the components as

Y : Outcome/Response $\quad$ X : Observed covariates $\quad$ A : Treatment assignment

Remark 1. *We could view our setting as coming from a potential outcome framework [27]. Under this framework we assume there exists RVs Y_0, Y_1 on $\mathcal{Y}$ representing the outcome with and without treatment and that $Y \equiv Y_A$. Y_{1-A} would then be unobserved/unknown for each individual.*

We define the *propensity score* $\pi : \mathcal{X} \to (0, 1)$ by

$$\pi(\boldsymbol{x}) \coloneqq \mathbb{P}(A = 1 | X = \boldsymbol{x}),$$

in other words, π denotes the conditional probability of being assigned to treatment given the covariates. We shall assume that π is continuous and bounded away from 0 and 1. From a potential outcomes perspective, this means that each individual could potentially be assigned to either treatment. We also define

$$F_a(y|\boldsymbol{x}) \coloneqq \mathbb{P}(Y \leq y | X = \boldsymbol{x}, A = a), \tag{1}$$

$$F_a^{-1}(\alpha|\boldsymbol{x}) \coloneqq \inf\{y \in \mathbb{R} | F_a(y|\boldsymbol{x}) \geq \alpha\}, \tag{2}$$

and refer to them as the *conditional cumulative distribution function (CCDF)* and the *quantile function* respectively. We also refer to F_a^{-1} in (2) as the *generalised inverse* of F_a.

We can now define the CATE and CQTE to be given by $\tau : \mathcal{X} \to \mathbb{R}$ and $\tau_q : [0, 1] \times \mathcal{X} \to \mathbb{R}$ with

$$\tau(\boldsymbol{x}) \coloneqq \mathbb{E}[Y | X = \boldsymbol{x}, A = 1] - \mathbb{E}[Y | X = \boldsymbol{x}, A = 0],$$
$$\tau_q(\alpha|\boldsymbol{x}) \coloneqq F_1^{-1}(\alpha|\boldsymbol{x}) - F_0^{-1}(\alpha|\boldsymbol{x}).$$

We let $D \coloneqq \{Z_i\}_{i=1}^{2n} \equiv \{(Y_i, X_i, A_i)\}_{i=1}^{2n}$ for $n \in \mathbb{N}$ be IID copies of Z representing our data sample with i indexing the individual. We assume an even number of samples for notational convenience. For $a \in \{0, 1\}$, we take $I_a \coloneqq \{i | A_i = 1\}$, the indices of individuals on treatment a. We can then define $D_a \coloneqq \{Z_i\}_{\{i \in I_a\}}$ and $n_a \coloneqq |I_a|$ as the dataset and sample size of those on treatment a.

For $n \in \mathbb{N}$, let $[n] \coloneqq \{1, \ldots, n\}$. For a vector $\boldsymbol{w} \in \mathbb{R}^p$ let w_j to represent the j^{th} component of $\boldsymbol{w}$ and let $\|\boldsymbol{w}\|$ be the Euclidean norm unless otherwise specified. We also take $\|\boldsymbol{w}\|_1$ as the 1-norm and $\|\boldsymbol{w}\|_\infty \coloneqq \max_{j \in [p]} |w_j|$. We keep a summary table of all notation used in Appendix A.1.

2.1 Introducing the quantile comparator

Our aim is to find "equivalent quantiles" between the treated and non-treated distributions conditional on the covariates. Specifically, for each $y_0 \in \mathcal{Y}$, $\boldsymbol{x} \in \mathcal{X}$ we aim to find y_1 such that

$$F_1(y_1|\boldsymbol{x}) = F_0(y_0|\boldsymbol{x}).$$

This now allows us to define our primary estimand of interest, the *conditional quantile comparator*.

Definition 1 (Conditional quantile comparator (CQC))**.** *For our triple (Y, X, A), the* conditional quantile comparator *is the measurable function $g^* : \mathcal{Y} \times \mathcal{X} \to \mathcal{Y}$ such that, for all $y \in \mathcal{Y}$, $\boldsymbol{x} \in \mathcal{X}$,*

$$F_1(g^*(y|\boldsymbol{x})|\boldsymbol{x}) = F_0(y|\boldsymbol{x}).$$

 https://doi.org/10.52202/079017-3260

We then simply define y_1 as $g^*(y_0|\boldsymbol{x})$. The name conditional quantile comparator derives from the fact that it returns the value of y_1 in the equivalent quantile of $Y|X=\boldsymbol{x}, A=1$ as the quantile of $Y|X=\boldsymbol{x}, A=0$ that y_0 is in.

Remark 2. *For simplicity and to ensure such a function is well defined, we will assume that $Y|X=\boldsymbol{x}, A=a$ is a continuous RV for any given $\boldsymbol{x}\in\mathcal{X},\ a\in\{0,1\}$ with strictly positive density on its support. We will however allow the support of $Y|X=x, A=a$ to vary in both $\boldsymbol{x}$ and a.*

We now introduce another estimand which will serve as a useful stepping stone in our estimation.

Definition 2 (CCDF contrasting function)**.** *The* CCDF contrasting function *is defined to be $h^*:\mathcal{Y}\times\mathcal{Y}\times\mathcal{X}\to[-1,1]$ given by*

$$h^*(y_0,y_1|\boldsymbol{x}) \coloneqq F_1(y_1|\boldsymbol{x}) - F_0(y_0|\boldsymbol{x}). \tag{3}$$

This estimand allows following alternative definitions for g^* which help its interpretation and estimation (detailed later). We take h^{*-1} representing the inverse of h^* with respect to the 2[nd] argument.

$$g^*(y_0|\boldsymbol{x}) = h^{*-1}(y_0,0|\boldsymbol{x}) = F_1^{-1}(F_0(y_0|\boldsymbol{x})|\boldsymbol{x}). \tag{4}$$

The second equality still holds if we replace the inverses in the above with generalised inverses. The equality (4) falls straight from the definition of each object and shows how we can use h^* to estimate g^*. Moreover, the equality (4) allows us to generalise g^* to discontinuous Y (or pdfs with non-trivial 0 density regions inside the support).

2.2 Exploring the CQC

We specifically focus on the CQC, g^*, as we feel it gives insightful information allowing comparison of the two distributions $(Y|X, A=1)$ and $(Y|X, A=0)$.

The CQC allows us to compare the two distributions beyond simply a single point estimate such as that given by the CATE. This is especially valuable in cases where the two distributions differ beyond just a shift. For example, the effect of some treatments varies greatly between individuals with the same or similar covariates. An example of this is antidepressants, where some patients respond positively while others may have adverse reactions leading to a worse outcome than no treatment whatsoever. Another example is the use of opioids as painkillers where some patients have an increased tolerance making them less effective [12, 22].

As well as being of interest on its own, the quantile comparator relates closely to other estimands of interest. For example, if we take $\Delta^*(y_0|\boldsymbol{x}) \coloneqq g^*(y_0|\boldsymbol{x}) - y_0$ then Δ^* tells us whether the equivalent quantile in the treated distribution is higher or lower. This estimand then serves as a heuristic for whether the treatment is beneficial at that untreated response value.

Furthermore, CQTE can be written as

$$\tau_q(\alpha|\boldsymbol{x}) = \Delta^*\left(F^{-1}_{Y|X,A=0}(\alpha|\boldsymbol{x})\middle|\boldsymbol{x}\right) = g^*\left(F^{-1}_{Y|X,A=0}(\alpha|\boldsymbol{x})\middle|\boldsymbol{x}\right) - F^{-1}_{Y|X,A=0}(\alpha|\boldsymbol{x}), \tag{5}$$

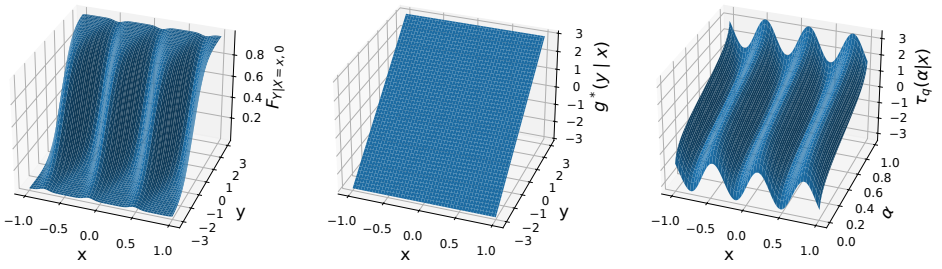
linking the quantile comparator back to the CQTE. This equivalence highlights the perspective that the CQC can be seen as rephrasing the input of the CQTE in terms of the outcome space.

A key idea within CATE literature is the notion that the CATE itself may be a simpler estimand to study than the marginal treatment outcomes ($\mathbb{E}[Y|X=\boldsymbol{x}, A=a]$) may be individually. One can exploit this feature to improve the CATE's estimation. A similar concept exists with the CQC as we will see in the example below.

Example 1 (Illustrative Example)**.** *Suppose that*

$$Y|X=x, A=0 \sim \mathrm{N}(\sin(10x),\ 1^2), \qquad Y|X=x, A=1 \sim \mathrm{N}(2\sin(10x),\ 2^2).$$

Then we have $g^*(y|x)=2y$ which does not depend on x and does not include the sine term present in the individual CDFs. Interestingly, $\mathbb{E}[Y|X, A=1] - \mathbb{E}[Y|X, A=0] = \sin(x)$ hence the CATE is still non-constant in this case (the same also holds for the CQTE). Additionally, we have $\Delta^*(y|x) = g^*(y|x) - y = y$ suggesting the intervention is beneficial for positive y and detrimental for negative y. We now show 3D plots of a CCDF, the CQC, and the CQTE in Figure 2.



(a) $F_{Y|X,0}(y|x)$ (b) $g^*(y|x)$ (c) $\tau_q^*(\alpha|x)$

Figure 2: Surface plots for CCDF (panel (a)), CQC (panel (b)) and CQTE (panel (c)). We can see that CCDF, and CQTE have high-frequency change in x while the CQC does not depend on x.

Example 2 (General Smoothness Case)**.** *Suppose we are in the potential outcomes framework so that Y_0, Y_1 exist with $Y \equiv Y_A$. Now also suppose that $Y_1 = \phi(Y_0, X)$ for some transformation ϕ increasing in Y_0 for each X. Then ϕ gives the CQC (i.e. $\phi = g^*$) meaning that smoothness of the CQC can be seen as smoothness of ϕ. This gives a generalisation of the CATE case where smoothness is present when $Y_1 = Y_0 + \psi(X)$ with ψ smooth*

A specific example could be a treatment which halves all individuals blood pressure. In this case $\phi^*(y, \boldsymbol{x}) = g^*(y|\boldsymbol{x}) = \frac{1}{2}y$ and so the CQC is smooth but the CQTE and CATE would not be if the individual responses are non-smooth.

3 Estimation procedure

We now describe our estimation procedure for the CQC which is motivated by equation (4). At a high level our approach for estimating $g^*(y_0|\boldsymbol{x})$ will be the following:

- Estimate $h^*(y_0, .|\boldsymbol{x})$ using a pseudo-outcome – a specified proxy response computed from (Y, X, A) which we will regress against.
- Find the value $\hat{y}_1$ which makes our estimate of $h^*(y_0, .|\boldsymbol{x})$ closest to 0.

Section 3.1 and Algorithm 1 give our h^* estimation procedure while Section 3.3 and Algorithm 2 give our g^* estimation procedure.

3.1 Estimating the CCDF contrasting function h^*

We focus on estimating h^* primarily for its two nice properties:

1. Similar to g^*, h^* can exhibit smoothness even when the individual CCDFs are not smooth.
2. The estimation of h^* can be re-framed as a CATE problem.

The first property is important as the smoothness of h^* determines the best estimation rate that can be achieved when using non-parametric regression, setting a target for our approach. In particular, smoother functions have better estimation rates. The second property is important as it gives us a method for attaining this target rate. By re-framing the estimation as a CATE problem, we can leverage existing results to build a robust estimator which achieves the target estimation accuracy rate even when the rate of estimating nuisance parameters is sub-optimal. We demonstrate this robustness later using finite sample bounds on the estimation accuracy (Proposition 1 & Theorem 2).

First, we show how the estimation of h^* can be solved using a CATE estimator. Note that

$$h^*(y_0, y_1|\boldsymbol{x}) = \mathbb{E}[\mathbb{1}\{Y \leq y_1\}|X = \boldsymbol{x}, A = 1] - \mathbb{E}[\mathbb{1}\{Y \leq y_0\}|X = \boldsymbol{x}, A = 0].$$

Hence, for a given y_0, y_1, if we define the RV $W_{y_0,y_1} \coloneqq \mathbb{1}\{Y \leq y_A\}$, then estimating $h^*(y_0, y_1|.)$ is equivalent to estimating the CATE with W_{y_0,y_1} replacing Y as the response. To perform this estimation, we turn to a recent method developed by Kennedy [15]. They propose to write the CATE as a conditional expectation of a function of Z called a pseudo-outcome. A robust estimator

 https://doi.org/10.52202/079017-3260

is then obtained by regressing this pseudo-outcome against X. In our setting, the pseudo-outcome with the new response W_{y_0,y_1}, for a sample $(y, \boldsymbol{x}, a)$ is given by

$$\varphi_{y_0,y_1}(y, \boldsymbol{x}, a) := \frac{a - \pi(\boldsymbol{x})}{\pi(\boldsymbol{x})(1 - \pi(\boldsymbol{x}))} \{\mathbb{1}\{y \leq y_a\} - F_a(y_a|\boldsymbol{x})\} + F_1(y_1|\boldsymbol{x}) - F_0(y_0|\boldsymbol{x}) \tag{6}$$
$$= \frac{a - \pi(\boldsymbol{x})}{\pi(\boldsymbol{x})(1 - \pi(\boldsymbol{x}))} \{\mathbb{1}\{y \leq y_a\} - F_a(y_a|\boldsymbol{x})\} + h^*(y_0, y_1|\boldsymbol{x}).$$

Since $h^*(y_0, y_1|\boldsymbol{x}) = \mathbb{E}[\varphi_{y_0,y_1}(Z)|X = \boldsymbol{x}]$ (Proposition 5, Appendix C.1), regressing $\varphi_{y_0,y_1}(Z)$ on X provides an estimate for h^*.

As we do not know the CDFs nor the propensity score, we need to replace them in (6) with estimates. We define $\hat{\pi}, \hat{F}_0, \hat{F}_1$ to be estimates of π, F_0, F_1 respectively. We then construct $\hat{\varphi}_{y_0,y_1}$ in the same way as φ_{y_0,y_1}, but using estimated quantities $\hat{\pi}$, $\hat{F}_a$ instead. $\hat{\varphi}_{y_0,y_1}$ can now serve as the pseudo-outcome in our regression. We also use sample splitting to de-correlate the propensity score and CDF estimates from the h^* estimate. This helps make our estimator doubly robust, as we will see in the following theory.

We are now ready to define our Doubly Robust (DR)-learner to estimate h^* in Algorithm 1.

Algorithm 1 DR estimation procedure for the CCDF contrasting function h^*

Require: $y_0, y_1 \in \mathcal{Y}$, Data D, a regressor (e.g. linear smoother)

1: Define $\mathcal{I} := [n]$, $\mathcal{J} := \{n + 1, \ldots, 2n\}$ and split D into $D_{\mathcal{I}} := \{Z_i\}_{i \in \mathcal{I}}, D_{\mathcal{J}} := \{Z_j\}_{j \in \mathcal{J}}$.
2: Using $D_{\mathcal{I}}$ to estimate $\hat{\pi}, \hat{F}_0(y_0|.), \hat{F}_1(y_1|.)$ by regressing $\mathbb{1}\{A = 1\}, \mathbb{1}\{Y \leq y_0\}, \mathbb{1}\{Y \leq y_1\}$ respectively against X.
3: Use these estimates to obtain $\hat{\varphi}(Z_j)$ for $j \in D_{\mathcal{J}}$ using equation (6)
4: Using $D_{\mathcal{J}}$ to regress $\hat{\varphi}$ against X to obtain estimate $\hat{h}(y_0, y_1|.)$ of $h^*(y_0, y_1|.)$.

Remark 3. *Cross-fitting can be implemented by repeating the procedure with the roles of $\mathcal{I}$ and $\mathcal{J}$ switched and then averaging the two estimates of h^*. We could also perform this procedure multiple times with different random splits of the data to improve our estimator's potential stability.*

Note that our algorithm is not specific on which form of regression to use allowing for any parametric or non-parametric procedure. Further to this, it can also be easily adapted to use other pseudo-outcome procedures such as the R-learner of Nie and Wager [23] or a standard inverse propensity weighting approach which we describe in Appendix A.3.

3.2 Finite sample bound of h^* estimator

In this section, we prove the estimation accuracy of $\hat{h}$, which will play important roles in the following notation and assumptions we need for estimating g^*. These accuracy statements will be made for an arbitrarily fixed $\boldsymbol{x} \in \mathcal{X}$. For our theoretical and experimental results we use linear smoothers for the final regression. This means our estimate $\hat{h}$ and oracle estimate h' are of the form

$$\hat{h}(y_0, y_1|\boldsymbol{x}) = \sum_{j \in \mathcal{J}} w_j \hat{\varphi}_{y_0,y_1}(Z_j) \qquad h'(y_0, y_1|\boldsymbol{x}) = \sum_{j \in \mathcal{J}} w_j \varphi_{y_0,y_1}(Z_j)$$

where the weights $w_j \equiv w_j(\boldsymbol{x}, X_{\mathcal{J}})$ are constructed using $X_{\mathcal{J}} := \{X_j\}_{j \in \mathcal{J}}$ with $w_j \geq 0$ and $\|\boldsymbol{w}\|_1$. Linear smoothers encompass a broad class of estimation techniques used in both low and high-dimensional settings. Examples include k-NN regression [8, 9], kernel ridge regression [29], generalised forests [3], and Mondrian forests [18]. Additionally linear smoothers have been shown to adapt to intrinsic low dimensionality in regression problems in higher dimensions [16], making them an apt estimator for our purposes.

For $\phi : \mathcal{Z} \to \mathbb{R}$ treated as deterministic and $p > 1$, we also define the norms

$$\|\phi\|_{\boldsymbol{w}^s} := \sqrt{\frac{1}{\|\boldsymbol{w}^s\|_1} \sum_{i \in \mathcal{J}} w_j^s \, \mathbb{E}\left[\phi(Z)^2 | X = X_i\right]}$$

and take $\|\phi\|_{\boldsymbol{w}} := \|\phi\|_{\boldsymbol{w}^1}$. Note that this norm is random as the weights depend upon $X_{\mathcal{J}}$.

We now aim to show that we are able to exploit smoothness in h^* even when the CCDFs and propensity score are less smooth. We introduce the notion of smoothness through Hölder functions.

Definition 3 (Hölder functions)**.** *We say that a function $f : \mathcal{X} \to \mathbb{R}$ is (γ, C)-Hölder for $\gamma \in (0, 1], C \geq 1$ if for any $\boldsymbol{x}', \boldsymbol{x}'' \in \mathcal{X}$,*

$$|f(\boldsymbol{x}') - f(\boldsymbol{x}'')| \leq C\|\boldsymbol{x}' - \boldsymbol{x}''\|^{\gamma}.$$

Here, larger γ represents a smoother function which can be estimated at faster rates.

Assumption 1. *For any $y_0, y_1 \in \mathcal{Y},\ \delta > 0, a \in \{0, 1\}$:*

(a) There exists $\xi \in (0, 1/2]$ such that $\pi(\boldsymbol{x}), \hat{\pi}(\boldsymbol{x}') \in [\xi, 1 - \xi]$ for all $\boldsymbol{x}' \in \mathcal{X}$.

(b) With probability at least $1 - \delta$, $\|\hat{\varphi}_{y_0,y_1} - \varphi_{y_0,y_1}\|_{\boldsymbol{w}^2} \leq \varepsilon_{\hat{\varphi}}(n, \delta)$.

(c) With probability at least $1 - \delta$, $\|\hat{\pi} - \pi\|_{\boldsymbol{w}} \leq \varepsilon_{\alpha}(n, \delta)$ and $\|F_a(y_a|.) - \hat{F}_a(y_a|.)\|_{\boldsymbol{w}} \leq \varepsilon_{\beta}(n, \delta)$.

(d) For $\gamma \in (0, 1],\ C \geq 1$, $h^*(y_0, y_1|.)$ is (γ, C)-Hölder.

Assumption 1(a) exists to ensure for any covariate value, neither treatment assignment has too low a probability. Assumption 1(b) controls the convergence of the estimated pseudo-outcome to the true pseudo-outcome. Assumption 1(c) sets up the smoothness of the propensity score and CCDFs alongside the convergence rates of their estimators as $\varepsilon_{\alpha}, \varepsilon_{\beta}$ respectively. Assumption 1(d) sets up the smoothness of h^* which will control the convergence rate of the oracle estimation procedure. Assumptions 1 (c) and (d) control the accuracy of our estimator in the following result.

Proposition 1. *Suppose that Assumption 1 holds and let $\hat{h}$ be a linear smoother estimated as in Algorithm 1. Then for any $y_1, y_0 \in \mathcal{Y},\ \delta \in (0, 2/e]$ and our $\boldsymbol{x} \in \mathcal{X}$, with probability at least $1 - \delta$,*

$$\left|\hat{h}(y_0, y_1|\boldsymbol{x}) - h^*(y_0, y_1|\boldsymbol{x})\right| \leq \varepsilon_h(n, \delta).$$

Here, for each $\delta \in (0, 2/e]$ we have

$$\begin{aligned}
\varepsilon_h(n, \delta) &:= \sqrt{2\log(8/\delta)/n}\,\varepsilon_{\hat{\varphi}}(n, \delta/4) + \varepsilon_{\alpha}(n, \delta/4)\varepsilon_{\beta}(n, \delta/4) + \varepsilon_{\gamma}(n, \delta/4), \\
\varepsilon_{\gamma}(n, \delta) &:= |\mathbb{E}[h'(y_0, y_1|\boldsymbol{x}) - h^*(y_0, y_1|\boldsymbol{x})|X_{\mathcal{J}}, D_{\mathcal{I}}]| \\
&\quad + \sqrt{2\log(2/\delta)/n}\,\|\boldsymbol{w}\|\|\varphi - h(y_0, y_1|.)\|_{\boldsymbol{w}^2} + 2\|\boldsymbol{w}\|_{\infty}\log(2/\delta)/(3\xi).
\end{aligned}$$

We have that ε_{γ} gives an upper bound on the accuracy of the oracle estimation and so acts as a target for our estimation procedure. If $\varepsilon_{\hat{\varphi}}(n, \delta) \to 0$ then the first term in ε_h is guaranteed to be $o(\varepsilon_{\gamma}(n, \delta))$ for fixed δ. Hence the first and last terms converge at oracle rates with respect to n. As the ε_{α} and ε_{β} terms are multiplied together we can obtain better rates than either of them individually have. This is because both ε_{α} and ε_{β} can converge to 0 slower than ε_{γ} while their product $\varepsilon_{\alpha} \cdot \varepsilon_{\beta}$ converges quicker. This provides the desired double robustness as our estimation can converge at oracle rates even when convergence for the nuisance parameters is slower. Now that we have this we can convert our estimate of h^*, into an estimate of g^*.

3.3 Estimating the conditional quantile comparator g^*

In order to obtain an estimate of $g^*(y_0|\boldsymbol{x})$ at a fixed $y_0, \boldsymbol{x}$, we need to obtain estimates of $h^*(y_0, y_1|\boldsymbol{x})$ at various values of y_1. As h^* is monotonic, we would then like to search for a monotonic function which aligns with these estimates. This monotonicity is especially important because it allows us to bound the estimation accuracy of $\hat{h}$ uniformly over all y_1 at a similar rate to our pointwise accuracy. With this, we can easily translate the estimation accuracy in $\hat{h}$ (obtained in proposition 1) into the accuracy in $\hat{g}$. Additionally, it simplifies the process of inverting $\hat{h}$. In general our method in Algorithm 1 will not produce monotonic $\hat{h}$ (this is in contrast to some other approaches such as an IPW pseudo-outcome, see Appendix A.3), or separately estimating the CCDFs). We can however obtain a monotonic estimate of h^* using isotonic projection.

Definition 4 (Isotonic Projection)**.** *We define the* isotonic projection *of $\boldsymbol{\alpha}' \in \mathbb{R}^p$ as follows:*

$$P(\boldsymbol{\alpha}') := \underset{\boldsymbol{\alpha} \in \mathrm{Iso}(p)}{\operatorname{argmin}} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|$$

where $\mathrm{Iso}(p) := \{\boldsymbol{\alpha} \in \mathbb{R}^p | \alpha_j \leq \alpha_{l+1}\ \forall l \in [p-1]\}$, the set of all isotonic vectors in $\mathbb{R}^p$.

https://doi.org/10.52202/079017-3260

Remark 4. *We can use the Pool Adjacent Violators Algorithm (PAVA) [6] which performs isotonic projection and is implemented in the* `IsotonicRegression` *class of sci-kit learn in Python [25].*

Hence, for a fixed value of $(y_0, \boldsymbol{x})$ and a set of predictions $\hat{\alpha}_l = \hat{h}(y_0, y_1^{(l)}|\boldsymbol{x})$ with $y^{(l)} \leq y^{(l+1)}$, we can take $\tilde{\boldsymbol{\alpha}} \coloneqq P_{\text{Iso}(p)}(\hat{\boldsymbol{\alpha}})$ and use these to obtain a new monotonic estimate of h^*. Furthermore, by a result in Yang and Barber [32], $\tilde{\boldsymbol{\alpha}}$ will be at least as accurate as $\hat{\boldsymbol{\alpha}}$ in the worst case. We now describe our approach for estimating the CQC using this projection approach in Algorithm 2.

Algorithm 2 DR estimation procedure for the CQC g^*

Require: Data D; test point $(y_0, \boldsymbol{x})$; sorted evaluation points $\{y^{(l)}\}_{l=1}^p$

1: Apply Algorithm 1 to obtain estimate of h^* given by $\hat{h}$.
2: Define $\hat{\alpha}_l \coloneqq \hat{h}(y_0, y^{(l)}|\boldsymbol{x})$ for $l \in [p]$.
3: Isotonically project $\hat{\boldsymbol{\alpha}}$ using PAVA to obtain $\tilde{\boldsymbol{\alpha}}$ with $\tilde{\alpha}_i \leq \tilde{\alpha}_{i+1}$.
4: Take $\hat{g}(y_0|\boldsymbol{x}) \coloneqq y^{(l^*)}$ with $l^* \coloneqq \text{argmin}_{l \in [p]} |\tilde{\alpha}_l|$.

Remark 5. *For the case where $\hat{h}$ is a step function, these steps can serve as our evaluation points while for continuous $\hat{h}$ one could take these candidate y_1 points at small evenly spaced intervals. Empirically we also find that $\hat{h}$ is already close to isotonic and so step 3. of the algorithm is mostly for the theoretical justification of our approach.*

While it may seem inefficient to be estimating the CQC via the CCDF contrasting function and then inverting, due to the monotonicity of h^*, we actually pay a very small cost in estimation accuracy for having to estimate the CCDF contrasting function over all y_1. We make this notion more explicit in the following section.

3.4 Finite sample bound of the CQC estimator

We now provide the accuracy of our estimate $\hat{g}$ obtained by Algorithm 2 when used in conjunction with linear smoothers. We assume that $\hat{F}_a$ are also fit using linear smoothers of the form

$$\hat{F}_a(y|\boldsymbol{x}', a) = \sum_{i \in \mathcal{I}} w_{F_a;i}(\boldsymbol{x}'; X_{\mathcal{I}}, A_{\mathcal{I}}) \mathbb{1}\{Y_i \leq y\}$$

with $w_j(\boldsymbol{x}') \equiv w_j(\boldsymbol{x}', X_{\mathcal{I}}, A_{\mathcal{I}}) > 0, \|\boldsymbol{w}(\boldsymbol{x}')\|_1 = 1$. We will also require the following assumptions.

Assumption 2. *For our RV X, any $y \in \mathcal{Y}$, $\boldsymbol{x}' \in \mathcal{X}$, $\delta < e^{-1}$:*

(a) There exists some $s, \eta > 0$ such that $F_1(y'|\boldsymbol{x}) \geq \eta$ for all $y' \in B_s(g^*(y|\boldsymbol{x}))$.

(b) W.p. at least $1 - \delta$, $\max_{j \in \mathcal{J}} w_j \leq \varepsilon_{\boldsymbol{w}}(n, \delta)$ and $\max_{i \in \mathcal{I}} w_{F_a;i}(X) \leq \varepsilon_{\boldsymbol{w}}(n, \delta)$.

Assumption 2(a) is a mild assumption which allows us to convert the $\hat{h}$ accuracy into $\hat{g}$ accuracy while Assumption 2(b) bounds the rates of decay of the weights in our linear smoothers.

Theorem 2. *Let $\hat{g}$ be estimated as using Algorithm 2 with $\{Y_i\}_{i \in I_1}$, sorted and then used as our evaluation points and linear smoothers used for regressions in Algorithm 1. Then provided Assumptions 1 & 2 hold we have that for $\delta \in (0, e^{-1})$ and sufficiently large n, w.p. at least $1 - \delta$,*

$$|\hat{g}(y|\boldsymbol{x}) - g^*(y|\boldsymbol{x})| \leq 2\left(\eta^{-1}\varepsilon_h(n, \delta/(2n)) + \xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/(2n))\right).$$

From this result we see that if our weights decay at rate faster than $\varepsilon_h(n, \delta)$ then this error will be dominated by the ε_h term. We believe this to hold in most cases and show that it does comfortably when using Nadaraya-Watson (NW) estimation [21, 31] with a box kernel in Appendix C.4. Furthermore, if the dependence on δ in both terms is of the form $\log^c(1/\delta)$ for some $c > 0$ then we obtain the same rate of estimation as for h^* up to polylog factors. This means we translate our desirable double robustness $\hat{h}$ over to $\hat{g}$. We also obtain finite sample bounds on $\mathbb{E}[|\hat{g}(Y|\boldsymbol{x}) - g^*(Y|\boldsymbol{x})| \mid A = 0, \hat{g}]$ with high probability and present this in Appendix C.3.

4 Numerical experiments

We now apply our approach to a series of simulated and real data scenarios in order to demonstrate the utility of our estimand and the effectiveness of our estimation procedure. For these, we use NW estimation as our regression procedure throughout. See appendix A.2 for details on NW estimation.

4.1 Simulated experiment

In this section, we test our method's performance in terms of our estimator's mean absolute error under simulated scenarios.[1] In each scenario we test against a separate estimator which estimates the two CCDFs separately and simply takes their difference, an IPW pseudo-outcome estimator detailed in Appendix A.3, the CQTE estimator of Kallus and Oprescu [14], and the oracle DR estimator where $\hat{\varphi}$ is replaced with φ (i.e. exact π, F_a are used). In this experiment, we return back to the set-up of example 1. We now change the frequency of the sine term by taking

$$Y|X=x, A=0 \sim \mathrm{N}(\sin(\gamma\pi x),\ 1^2), \qquad Y|X=x, A=1 \sim \mathrm{N}(2\sin(\gamma\pi x),\ 2^2),$$
$$\pi(x) = 0.4\sin(\gamma\pi x) + 0.5.$$

for $\gamma \in [0, 10]$ so that increasing γ imitates decreasing smoothness of our nuisance parameters. In our experiments half the samples are used to estimate the propensity score and CCDFs and the other half are used to regress against the pseudo-outcome. Our estimate $\hat{g}$ is then compared against g^* using a hold-out testing set. This process is repeated 500 times with new training data on each run. From this, a Monte-Carlo estimate of $\mathbb{E}_{\hat{g}}[\mathbb{E}_X[\mathbb{E}_{Y|X,A=0}[|\hat{g}(Y|X) - g^*(Y|X)|]]]$ is produced alongside 95% confidence intervals (CIs). In our first experiment, we let $2n = 1000$ and vary γ in $[0, 10]$. In our second experiment, we let $\gamma = 6$ and $2n$ vary in $[200, 5000]$. The results of this are shown in Figure 3.

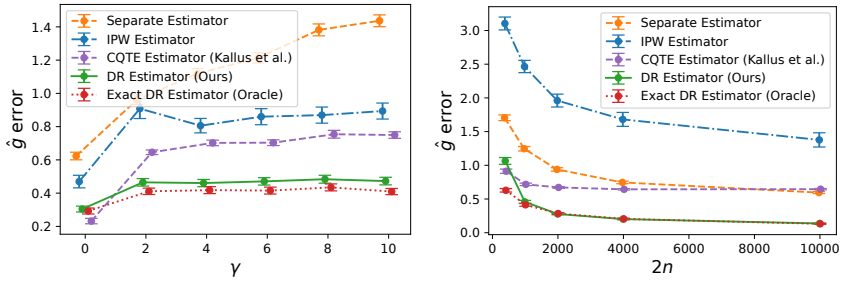


Figure 3: Mean absolute error with 95% CIs for various estimators. The left plot has fixed sample size ($2n = 1000$) and increasing γ. The right plot has $\gamma = 6$ and increasing sample size.

We can see that the average error decreases as the sample size $2n$ grows and mildly increases as γ grows. This is expected as increasing γ makes estimating the nuisance parameters more challenging. The result shows that the proposed DR method achieves the best performance compared with the Separate and IPW estimators and is only marginally outperformed by the oracle estimator. Additionally we see that the method of Kallus and Oprescu [14] is much more affected by the increase in γ. This is because unlike the CQC, the CQTE has a complexity that depends on the frequency term (γ). We also observe much better performance as the sample size increases. We hypothesise that the plateau in the CQTE approach is due to the difficulty of estimating the reciprocal of the PDF, which causes the estimation to be unstable irrespective of sample size. Further simulated experiments including 10-dimensional $\boldsymbol{x}$ and linear CQC are given in Appendix B.1.

4.2 Real world employment example

To show the performance in the real-world scenarios, we use a dataset on an employment programme which has been studied in various prior works [4, 5, 26]. Within the programme, some participants were given job placements or temporary help jobs while others received no intervention. Participants' earnings were then monitored over the next 8 quarters following their enrolments. We take their net earnings as our response (Y) and the employment intervention as the treatment ($A = 1$). We use each participant's age at their entry to the study as our covariate (X). We fit our quantile comparator function on 2,000 participants. Figure 4 shows our estimate of $\Delta^*(y|\boldsymbol{x}) = g^*(y|\boldsymbol{x}) - y$ for various values of $(y, \boldsymbol{x})$.

We see that participants around age 23 and between ages 32-37 benefit most from this scheme, as indicated by the darker colour on the heat map. Additionally, the lower quantile of the income distribution (wage $\leq$ \$7500) shows the least change, indicating that wage improvements primarily occur

[1] Code implementation can be found at: github.com/joshgivens/ConditionalOutcomeEquivalence

https://doi.org/10.52202/079017-3260

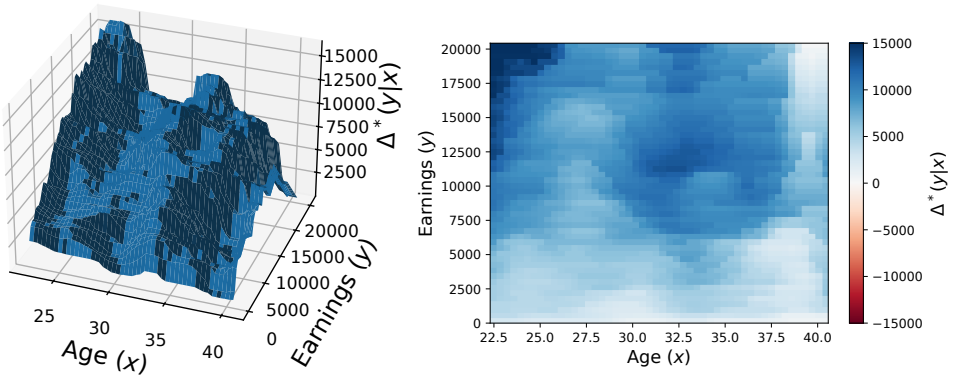


Figure 4: Surface and heat plot of $\Delta^*(y|\boldsymbol{x})$ for our employment data with X =Age, Y=Income.

for the higher income group. For participants aged 40, there appears to be little change in outcomes overall. To demonstrate our approach in a medical setting, we apply our method to evaluate the effectiveness of a colon cancer treatment, as detailed in Appendix B.3. For comparative purposes, we also provide an estimate of the CQTE for this data in Appendix B.2.

5 Limitations

The theory provided here gives a strong foundation and motivation for framing our problem in this particular manner. There is however a great deal more to be explored in this area from a theoretical perspective. For example, one immediate improvement would be to give a more general case where our weight decay (in Assumption 2 (b) for Theorem 2) is sufficiently fast. In addition more work needs to be done exploring the relationship between the smoothness of g^* and the smoothness h^*. Smoothness in h^* appears to be a stronger condition so ideally we would like to make theoretical statements directly on the smoothness of g^*. Interestingly our experiments on synthetic data do seem to suggest that it is the complexity of g^* which drives the estimation rate as in these experiments h^* increases in complexity while g^* remains constant. Additionally, changing our experiment in Section 4.1 to have uniform response so that both h^*, g^* are constant rather than just g^*, seems to give no material improvement to the performance of our estimator (see Appendix B.1.3).

Another limitation of our current estimation procedure is that it requires learning an estimand and then inverting it to obtain our final estimator. While this process is relatively simple, it could be streamlined and made more computationally efficient if we could produce a more direct estimator similar to the DR-Learner for CATE [15] or the CQTE estimator in Kallus and Oprescu [14].

Finally, while we have been able to provide more concrete examples of smoothness for the CQC these are still limited to the case of a deterministic treatment effect which we would like to expand upon. This is closely related to a more general limitation with quantile-based estimands, the CQC included, in that they lack meaningful interpretability for the individual. While the CATE can be viewed as the expected difference in an individual's outcome on and off the treatment, no such individual-level interpretation exists for the CQC or indeed any other estimand trying to learn higher level distributional information than the mean. As a result the CATE is still a more naturally interpretable estimand. To facilitate this interpretation however, one still needs to make assumptions about a lack of confounding between the treatment assignment and the potential outcomes, which are only verifiable in certain restrictive scenarios [30].

6 Conclusion

In this paper we have introduced a new treatment effect estimand, the conditional quantile comparator and demonstrated its efficacy both in terms of its doubly robust estimation, and its ability to provide valuable data insights. This is a promising direction as it allows quantile-based treatment effect exploration to "keep up" with the CATE in terms of estimation quality offering more flexibility as to which estimand can be used to best describe the data. For these reasons, we see the CQC as an exciting and worthwhile new direction within the HTE framework.

Acknowledgments and Disclosure of Funding

Josh Givens was supported by a PhD studentship from the EPSRC Centre for Doctoral Training in Computational Statistics and Data Science (COMPASS).

References

[1] Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117.

[2] Abadie, A. and Imbens, G. W. (2002). Simple and bias-corrected matching estimators for average treatment effects. Working Paper 283, National Bureau of Economic Research. Series: Technical working paper series.

[3] Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178. Publisher: Institute of Mathematical Statistics.

[4] Autor, D. H. and Houseman, S. N. (2010). Do temporary-help jobs improve labor market outcomes for low-skilled workers? Evidence from "Work First". *American Economic Journal: Applied Economics*, 2(3):96–128.

[5] Autor, D. H., Houseman, S. N., and Kerr, S. P. (2017). The effect of work first job placements on the distribution of earnings: An instrumental variable quantile regression approach. *Journal of Labor Economics*, 35(1):149–190.

[6] Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions (The Theory and Application of Isotonic Regression)*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd.

[7] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press.

[8] Chen, G. (2019). Nearest neighbor and kernel survival analysis: Nonasymptotic error bounds and strong consistency rates. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th international conference on machine learning*, volume 97 of *Proceedings of machine learning research*, pages 1001–1010. PMLR.

[9] Chen, P., Dong, W., Lu, X., Kaymak, U., He, K., and Huang, Z. (2019). Deep representation learning for individualized treatment effect estimation using electronic health records. *Journal of Biomedical Informatics*, 100:103303.

[10] Collins, F. S. and Varmus, H. (2015). A new initiative on precision medicine. *The New England journal of medicine*, 372(9):793–795. Place: United States.

[11] Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701. Publisher: [Wiley, The Econometric Society].

[12] Huynh, P., Villaluz, J., Bhandal, H., Alem, N., and Dayal, R. (2021). Long-Term Opioid Therapy: The Burden of Adverse Effects. *Pain Medicine*, 22(9):2128–2130.

[13] Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics*, 86(1):4–29.

[14] Kallus, N. and Oprescu, M. (2023). Robust and agnostic learning of conditional distributional treatment effects. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of the 26th international conference on artificial intelligence and statistics*, volume 206 of *Proceedings of machine learning research*, pages 6037–6060. PMLR.

[15] Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008 – 3049. Institute of Mathematical Statistics and Bernoulli Society.

[16] Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in neural information processing systems*, volume 24. Curran Associates, Inc.

 https://doi.org/10.52202/079017-3260

[17] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Bin Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

[18] Lakshminarayanan, B., Roy, D. M., and Teh, Y. W. (2014). Mondrian forests: Efficient online random forests. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in neural information processing systems*, volume 27. Curran Associates, Inc.

[19] Laurie, J. A., Moertel, C. G., Fleming, T. R., Wieand, H. S., Leigh, J. E., Rubin, J., McCormack, G. W., Gerstner, J. B., Krook, J. E., and Malliard, J. (1989). Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic. *Journal of clinical oncology*, 7(10):1447–1456. Place: United States.

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

[21] Nadaraya, E. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, 10(1):186–190.

[22] Nadeau, S. E., Wu, J. K., and Lawhern, R. A. (2021). Opioids and chronic pain: An analytic review of the clinical evidence. *Frontiers in Pain Research*, 2.

[23] Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

[24] Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine*, 375(13):1216–1219. Place: United States.

[25] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[26] Powell, D. (2020). Quantile Treatment Effects in the Presence of Covariates. *The Review of Economics and Statistics*, 102(5):994–1005.

[27] Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

[28] Semenova, V. and Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289.

[29] Singh, R., Xu, L., and Gretton, A. (2023). Kernel methods for causal functions: dose, heterogeneous and incremental response curves. *Biometrika*, 111(2):497–516. tex.eprint: https://academic.oup.com/biomet/article-pdf/111/2/497/57467664/asad042.pdf.

[30] VanderWeele, T. J. (2008). The sign of the bias of unmeasured confounding. *Biometrics. Journal of the International Biometric Society*, 64(3):702–706. Publisher: [Wiley, International Biometric Society].

[31] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

[32] Yang, F. and Barber, R. F. (2019). Contraction and uniform convergence of isotonic regression. *Electronic Journal of Statistics*, 13(1):646 – 677. Publisher: Institute of Mathematical Statistics and Bernoulli Society.

[33] Ying Zhang, Lei Wang, M. Y. and Shao, J. (2020). Quantile treatment effect estimation with dimension reduction. *Statistical Theory and Related Fields*, 4(2):202–213. Publisher: Taylor & Francis tex.eprint: https://doi.org/10.1080/24754269.2019.1696645.

[34] Zhou, T., Carson, W. E. t., and Carlson, D. (2022). Estimating Potential Outcome Distributions with Collaborating Causal Networks. *Transactions on machine learning research*, 2022. Place: United States.

A Additional details

A.1 Notation table

Table 1: Table of notation.

Notation	Definition	Description
Y	RV on $\mathcal{Y} \subseteq \mathbb{R}$	Response
X	RV on $\mathcal{X} \subseteq \mathbb{R}^d$	Covariates
A	RV on $\{0,1\}$	Treatment
Z	(Y, X, A)	
$\pi(\boldsymbol{x})$	$\mathbb{P}(A=1\|X=\boldsymbol{x})$	Propensity Score
$F_a(y\|\boldsymbol{x})$	$\mathbb{P}(Y \leq y\|X=\boldsymbol{x}, A=a)$	Conditional CDF (CCDF)
$F_a^{-1}(\alpha\|\boldsymbol{x})$	$\inf\{y \in \mathcal{Y} \| F_a(y\|\boldsymbol{x}) \geq \alpha\}$	Conditional quantile function
$\tau(\boldsymbol{x})$	$\mathbb{E}[Y\|X=\boldsymbol{x}, A=1] - \mathbb{E}[Y\|X=\boldsymbol{x}, A=0]$	Conditional average treatment effect (CATE)
$\tau_q(\alpha\|\boldsymbol{x})$	$F_1^{-1}(\alpha\|\boldsymbol{x}) - F_0^{-1}(\alpha\|\boldsymbol{x})$	Conditional quantile treatment effect (CQTE)
D	$\{Z_i\}_{i=1}^{2n}$	IID data sample
I_a	$\{i \in [2n] \| A_i = 1\}$	Treatment a index
D_a	$\{Z_i\}_{i \in I_a}$	
n_a	$\|I_a\|$	
$[n]$	$\{1, \ldots, n\}$	
$\|\boldsymbol{x}\|$	$\sqrt{\frac{1}{d}\sum_{i=1}^d x_i^2}$	Euclidean Norm
$\|\boldsymbol{x}\|_1$	$\frac{1}{d}\sum_{j=1}^d \|x_j\|$	1-norm
$\|\boldsymbol{x}\|_\infty$	$\max_{j \in [d]} \|w_j\|$	∞-norm
$g^*(y\|\boldsymbol{x})$	$F_1^{-1}(F_0(y\|\boldsymbol{x})\|\boldsymbol{x})$	Conditional quantile comparator CQC
$h^*(y_0, y_1\|\boldsymbol{x})$	$F_1(y_1\|\boldsymbol{x}) - F_{Y\|X,0(y_0\|\boldsymbol{x})}$	CCDF contrasting function
$\Delta^*(y\|\boldsymbol{x})$	$g^*(y\|\boldsymbol{x}) - y$	Quantile difference function
$\hat{f}$	Estimate of f/f^*	
$\varphi_{y_0,y_1}(y, \boldsymbol{x}, a)$	$\frac{a-\pi(\boldsymbol{x})}{\pi(\boldsymbol{x})(1-\pi(\boldsymbol{x}))}\{\mathbb{1}\{y \leq y_a\} - F_a(y_a\|\boldsymbol{x})\} + F_1(y_1\|\boldsymbol{x}) - F_0(y_0\|\boldsymbol{x})$	The Pseudo-Outcome
$\hat{\varphi}_{y_0,y_1}(y, \boldsymbol{x}, a)$	φ with π, F_a replaced by $\hat{\pi}, \hat{F}_a$	
$\mathcal{I}, \mathcal{J}$	$\{1, \ldots, n\}, \{n+1, \ldots, 2n\}$	Sample splitting indices
$D_\mathcal{I}, D_\mathcal{J}$	$\{Z_i\}_{i \in \mathcal{I}}, \{Z_j\}_{j \in \mathcal{J}}$	Data split
$\boldsymbol{w} \in \mathbb{R}^n$	$w_j \equiv w_j(\boldsymbol{x})$	$\hat{h}$ estimation weights
$\hat{h}(y_0, y_1\|\boldsymbol{x})$	$\sum_{j \in \mathcal{J}} w_j(\boldsymbol{x})\hat{\varphi}_{y_0,y_1}(Z_j)$	h^* estimate
$h'(y_0, y_1\|\boldsymbol{x})$	$\sum_{j \in \mathcal{J}} w_j(\boldsymbol{x})\varphi_{y_0,y_1}(Z_j)$	Oracle h^* estimate
$\|\phi\|_{\boldsymbol{w}^s}$	$\sqrt{\|\boldsymbol{w}^s\|_1^{-1}\sum_{i \in \mathcal{J}} w_j^s\, \mathbb{E}\left[\phi(Z)^2\|X=X_i\right]}$	
ξ	$\pi(\boldsymbol{x}') \in [\xi, 1-\xi]$ for all $\boldsymbol{x}' \in \mathcal{X}$	
$\varepsilon_{\hat{\varphi}}(n, \delta)$	$\|\hat{\varphi} - \varphi\|_{\boldsymbol{w}^2} \leq \varepsilon_{\hat{\varphi}}(n, \delta)$ w.p. $1-\delta$	$\hat{\varphi}$ error bound w.h.p.
$\varepsilon_\alpha(n, \delta)$	$\|\hat{\pi} - \pi\|_{\boldsymbol{w}^2} \leq \varepsilon_\alpha(n, \delta)$ w.p. $1-\delta$	$\hat{\pi}$ error bound w.h.p.
$\varepsilon_\beta(n, \delta)$	$\|\hat{F}_a(y\|.) - F_a(y\|.)\|_{\boldsymbol{w}^2} \leq \varepsilon_\beta(n, \delta)$ w.p. $1-\delta$	$\hat{F}_a$ error bound w.h.p.
γ	Smoothness of $h^*(y_0, y_1\|.)$	
$\varepsilon_\gamma(n, \delta)$	$\mathbb{E}[h'(y_0, y_1\|\boldsymbol{x}) - h^*(y_0, y_1\|\boldsymbol{x})\|X_\mathcal{J}, D_\mathcal{I}]\|$ $+ \sqrt{2\log(1/\delta)/n}\,\|\boldsymbol{w}\|\|\varphi - h(y_0, y_1\|.)\|_{\boldsymbol{w}^2}$ $+ 2\xi^{-1}\|\boldsymbol{w}\|_\infty \varepsilon_\delta(n)/3$	h' error bound w.h.p.
$\varepsilon_{T_n}(n, \delta)$	$\sqrt{2\log(1/\delta)/n}\,\varepsilon_{\boldsymbol{w}}(n, \delta)\varepsilon_{\hat{\varphi}}(n, \delta)$	$T_n(\boldsymbol{x})$ error bound w.h.p.
$\varepsilon_h(n, \delta)$	$\varepsilon_{T_n}(n, \delta/4) + (\varepsilon_\alpha \cdot \varepsilon_\beta)(n, \delta/4) + \varepsilon_\gamma(n, \delta/4)$	$\hat{h}$ error bound w.h.p.
$\mathrm{Iso}(p)$	$\{\boldsymbol{\alpha} \in \mathbb{R}^p \| \alpha_l \leq \alpha_{l+1} \forall l \in [p-1]\}$	Set of isotonic vectors
$P_{\mathrm{Iso}(p)})(\boldsymbol{\alpha}')$	$\mathrm{argmin}_{\boldsymbol{\alpha} \in \mathrm{Iso}(p)}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|$	Isotonic projection
$\boldsymbol{w}_{F_a} \in \mathbb{R}^n$	$w_{F_a;i}(\boldsymbol{x}) \equiv w_{F_a;i}(\boldsymbol{x}', X_\mathcal{I})$	$\hat{F}_a$ estimation weights

https://doi.org/10.52202/079017-3260

Notation	Definition	Description
$\hat{F}_a(y\|\boldsymbol{x}')$	$\sum_{i\in\mathcal{J}} w_{F_a}(\boldsymbol{x})\mathbb{1}\{Y_j \le y\}$	CCDF estimate
η	$F_1(y'\|\boldsymbol{x}) > \eta\ \forall, y' \in B_s(g^*(y\|\boldsymbol{x}))$	Density lower bound
$\varepsilon_{\boldsymbol{w}}(n,\delta)$	$\max\{\|\boldsymbol{w}(\boldsymbol{x})\|_\infty, \|\boldsymbol{w}_{F_a}(X)\|_\infty\}$ w.p. $1-\delta$	Variance lower bound w.h.p.
	Appendix Notation	
$k(\boldsymbol{x},\boldsymbol{x}')$		NW estimation kernel
$m_f(\boldsymbol{x})$	$\mathbb{E}[f(Z)\|X=\boldsymbol{x}]$	
$\hat{m}_f(\boldsymbol{x})$	$\sum_{j\in\mathcal{J}} w_j(\boldsymbol{x})f(Z_j)$	
$\hat{b}(\boldsymbol{x})$	$m_{\hat{\varphi}-\varphi}(\boldsymbol{x})$	
$\varepsilon_\delta(n)$	$\max\{\log(1/\delta)/n, 1/n\}$	
$[n]_0$	$\{0,1,\dots,n\}$	
$f(\Delta y)$	$\lim_{\varepsilon\downarrow 0} f(y) - f(y-\varepsilon)$	Step size of f at y

A.2 Nadaraya-Watson estimation

Throughout, we use NW estimation as our standard non-parametric regression technique. For a kernel $k : \mathcal{X}\times\mathcal{X} \to [0,\infty)$, IID data sample $D \coloneqq \{(Y_i, X_i)\}_{i=1}^n$, and $\boldsymbol{x}\in\mathcal{X}$ the NW estimate of $\mathbb{E}[Y|X=\boldsymbol{x}]$ is given by

$$\sum_{i=1}^n \frac{k(\boldsymbol{x},X_i)}{\sum_{j=1}^n k(\boldsymbol{x},X_j)} Y_i.$$

Applying this to our pseudo-outcome regression in Algorithm 1, we get that

$$\hat{h}(y_0,y_1|\boldsymbol{x}) \coloneqq \sum_{i\in\mathcal{J}} \frac{k(\boldsymbol{x},X_i)}{\sum_{j\in\mathcal{J}} k(\boldsymbol{x},X_j)} \hat{\varphi}_{y_0,y_1}(Y_i,X_i,A_i) \tag{7}$$

$$= \sum_{i\in\mathcal{J}} \frac{k(\boldsymbol{x},X_i)}{\sum_{j\in\mathcal{J}} k(\boldsymbol{x},X_j)} \Bigg(\frac{A_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)(1-\hat{\pi}(X_i))} \cdot \Big\{\mathbb{1}\{Y_i \le y_{A_i}\} - \hat{F}_{Y|X,A_i}(y_{A_i}|X_i)\Big\} + \hat{F}_a(y_1|X_i) - \hat{F}_0(y_0|X=X_i)\Bigg) \tag{8}$$

where for any $\boldsymbol{x}'\in\mathcal{X}$, $a\in\{0,1\}$

$$\hat{\pi}(\boldsymbol{x}') \coloneqq \sum_{i\in\mathcal{I}} \frac{k(\boldsymbol{x}',X_i)}{\sum_{j\in\mathcal{I}} k(\boldsymbol{x}',X_j)} \mathbb{1}\{A_i=1\} \tag{9}$$

$$\hat{F}_a(y_a|\boldsymbol{x}') \coloneqq \sum_{i\in\mathcal{I}} \frac{k(\boldsymbol{x}',X_i)}{\sum_{j\in\mathcal{I}} k(\boldsymbol{x}',X_j)\mathbb{1}\{A_j=a\}} \mathbb{1}\{Y_i\le y_a\}\mathbb{1}\{A_i=a\}. \tag{10}$$

Note that for our theoretical results we do not specify how our nuisance parameters are estimated and the results only depend on the accuracy of our estimation of the nuisance parameters.

Remark 6. *We can re-write* (8) *as:*

$$\hat{h}(y_0,y_1|\boldsymbol{x}) \coloneqq \sum_{i\in\mathcal{I}_1} \frac{k(\boldsymbol{x},X_i)}{\sum_{i=1}^n k(\boldsymbol{x},X_i)} \Bigg(\frac{1}{\hat{\pi}(X_i)} \Big\{\mathbb{1}\{Y_i\le y_1\} - \hat{F}_1(y_1|X_i)\Big\} + \hat{F}_1(y_1|X_i) - \hat{F}_0(y_0|X_i)\Bigg)$$
$$- \sum_{i\in\mathcal{I}_0} \frac{k(\boldsymbol{x},X_i)}{\sum_{i=1}^n k(\boldsymbol{x},X_i)} \Bigg(\frac{1}{1-\hat{\pi}(X_i)} \Big\{\mathbb{1}\{Y_i\le y_0\} - \hat{F}_0(y_0|X_i)\Big\} + \hat{F}_0(y_0|X_i) - \hat{F}_1(y_1|X_i)\Bigg)$$

Which can be helpful in terms of the practical implementation.

A.3 IPW pseudo-outcome estimator

Another pseudo-outcome one can use for estimating the treatment effect is the based on inverse propensity weighting. Specifically we can take our pseudo-outcome to be

$$\psi_{y_0,y_1}(Y', X', A') \coloneqq \frac{A' - \pi(X')}{\pi(X')(1-\pi(X'))} \mathbb{1}\{Y' \leq y_{A'}\}. \tag{11}$$

If we regress against it using NW estimation, and also use NW estimation for our nuisance parameter estimation as in (9) & (10), our estimate $\hat{h}$ will be increasing in y_1 and decreasing in y_0 meaning we do not need to perform any isotonic projection.

A.4 Additional experimental details

As the method of Kallus and Oprescu [14] is one for estimating the CQTE, we transform it into an estimator of the CQC (which we denote by $\hat{g}$) using the following formula where $\hat{\tau}_q(.|.)$ is our CQTE estimator

$$\hat{g}(y|\boldsymbol{x}) = \hat{\tau}_q\left(F_{Y|X,0}(y|\boldsymbol{x})|\boldsymbol{x}\right) + y.$$

using the true CCDF, $F_{Y|X,0}$.

This is the inverse of equation (5) which defines the CQTE in terms of the CQC. Conversely we also tested transforming all our CQC estimators into CQTE estimators (using exactly equation (5) with $\hat{g}$ replacing g^*) and testing the accuracy on this space and found similar results.

Each experiment took no longer than 1 hour to run on a single 4 core CPU with 8GB of RAM.

The bandwidths of the kernels for the NW estimation of the nuisance parameters were chosen by a limited grid search on additional simulated data by validating against the true value of the nuisance parameters. While this is unrealistic in practice it was done to make estimation of the nuisance parameters as strong as possible. This was to err on the side of caution as strong nuisance parameter estimation would naturally favour the baseline approaches.

Code to implement our approach alongside Jupyter notebooks running our numerical experiments can be found in the supplementary materials. The code to implement the kernels is adapted from `https://github.com/wittawatj/kernel-gof/` which is free to use under the MIT Licence.

B Additional results

B.1 Simulation results

We run additional simulated experiments in order to further test and explore our approach. Throughout, our overall experimental set-up is the same as in Section 4.1

B.1.1 10-dimensional example

To test our problem in higher dimensions we ran experiments where X was 10-dimensional with X uniform on $[-1,1]^{10}$. That is each component X_j was independent with $X_j \sim U(-1,1)$. We then took

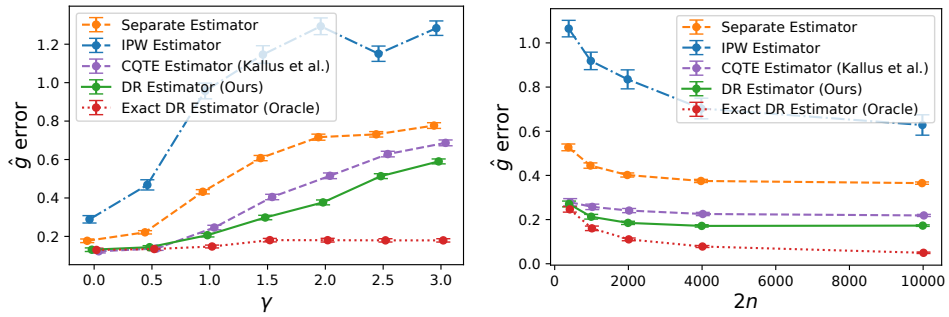
$$\begin{aligned} Y|X = \boldsymbol{x}, A = a &\sim N(\sin(\gamma\pi(\boldsymbol{\beta}^\top \boldsymbol{x}), 1) \\ \pi(\boldsymbol{x}) &= 0.4\sin(\gamma\pi(\boldsymbol{\beta}^\top \boldsymbol{x})) + 0.5 \end{aligned}$$

where $\boldsymbol{\beta}$ was randomly sampled from $N(0,(0.2)^2)$ and $\gamma \in [0,3]$. This gave the maximum gradient multiplied by $0.5\,\mathrm{diam}(\mathcal{X}) = \sqrt{d}$ close to 1 making the rate of change of the CDFs and propensities similar to our 1-dimensional examples.

In our first experiment we take $2n = 1000$ and $\gamma \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$. The results of this are shown in Figure 5a In our second experiment we take $\gamma = 1$ and $2n \in \{200, 500, 1000, 2000, 5000\}$. The results of this are shown in Figure 5b.

As we can see the DR approach performs the best with it performing close to oracle for low sample size and low frequency. As the frequency increases the DR estimator deviates from the oracle.

 https://doi.org/10.52202/079017-3260



(a) Average error as γ increases with 95% CIs for various approaches.

(b) Average error as sample size n increases with 95% CIs for various approaches.

Figure 5: Estimation accuracy of 10-dimensional CQC estimation procedures for highly varying CCDFs and constant CQC with respect to $\boldsymbol{x}$.

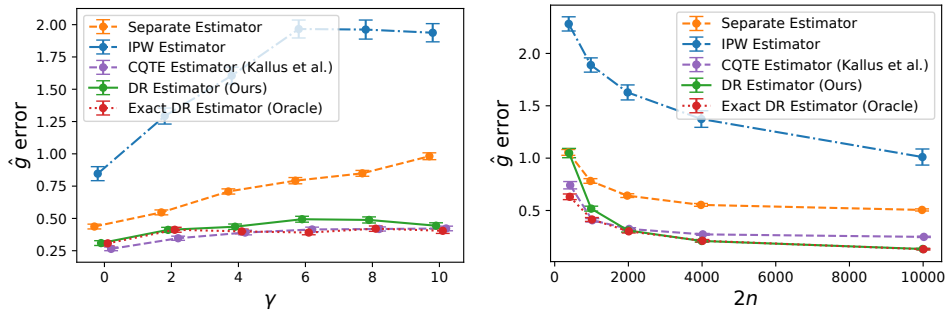
B.1.2 Varying CQC

In this experiment our set-up is

$$Y|X, A=0 \sim N\left(\frac{\sin(\gamma\pi x)}{0.5x+1.5}, 1\right)$$
$$Y|X, A=1 \sim N\left(\sin(\gamma\pi x)+0.25x+0.75, (0.5x+1.5)^2\right)$$
$$\pi(x) = 0.4\sin(\gamma\pi x)+0.5$$

for $\gamma \in [0, 10]$. This gives $g^* = (y+0.5)(0.5x+1.5)$ which is still simpler than the individual CCDFs but now does depend upon x.

In our first experiment we take $2n = 1000$ and $\gamma \in \{0, 2, 4, 6, 8, 10\}$. The results of this are shown in figure 5a In our second experiment we take $\gamma = 6$ and $2n \in \{200, 500, 1000, 2000, 5000\}$. The results of this are shown in Figure 5b.



(a) Average error as n increases with 95% CIs for various approaches.

(b) Average error as n increases with 95% CIs for various approaches.

Figure 6: Estimation accuracy of CQC estimation procedures for highly varying CCDFs and linear CQC with respect to x.

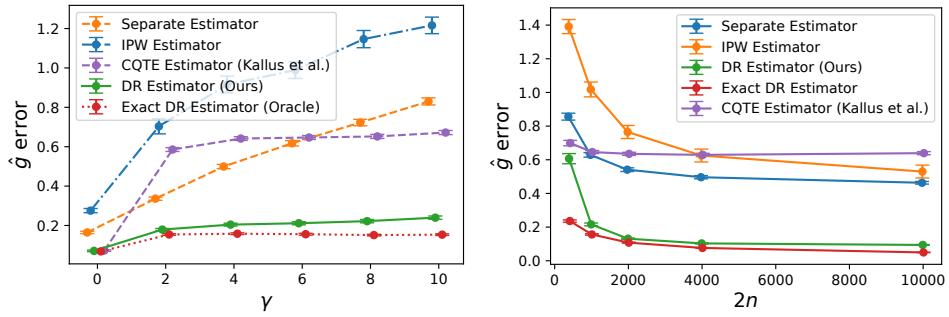
B.1.3 Constant h^*

In our previous examples, while g^* has been simple and or constant, h^* has actually included the high frequency sine term. In this experiment we adjust our original illustrative example to give a constant h^* as well. Specifically we take

$$Y|X=x, A=0 \sim \text{Unif}(\sin(\gamma\pi x), \sin(\gamma\pi x)+1),$$
$$Y|X=x, A=1 \sim \text{Unif}(2\sin(\gamma\pi x), 2\sin(\gamma\pi x)+2).$$

for $\gamma \in [0, 10]$. We also take the propensity to be $\pi(x) = 0.4\sin(\gamma\pi x) + 0.5$. In this case $h^*(y_0, y_1|\boldsymbol{x}) = \frac{1}{2}y_1 - y_0$ which does not depend on $\boldsymbol{x}$ for any $y_0, y_1 \in \mathcal{Y}$.

In our first experiment we take $2n = 1000$ and $\gamma \in \{0, 2, 4, 6, 8, 10\}$. The results of this are shown in figure 5a In our second experiment we take $\gamma = 6$ and $2n \in \{200, 500, 1000, 2000, 5000\}$. The results of this are shown in Figure 5b. As we can see we obtain very similar results to original



(a) Average error as n increases with 95% CIs for various approaches.

(b) Average error as n increases with 95% CIs for various approaches.

Figure 7: Constant h^*

Figure 8: Estimation accuracy of CQC estimation procedures for highly varying CCDFs and constant h^* and CQC.

illustrative example suggesting that our method is effectively exploiting simplicity in g^* rather than in h^*.

B.2 Employment example: Comparing to the CQTE

To further explore the interpretability of our estimator we provide an estimate of the CQTE for comparison. This can be seen in Figure 9.

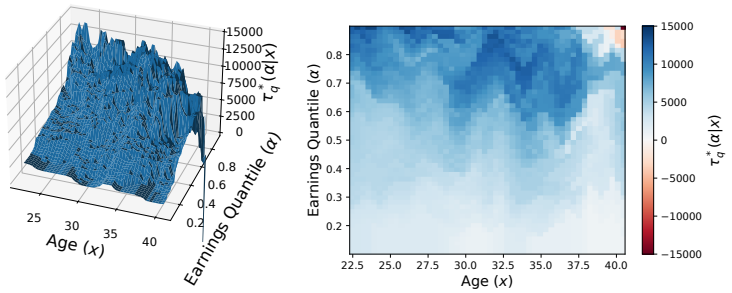


Figure 9: Surface and heat plot of the CQTE, $\tau_q^*(\alpha|\boldsymbol{x})$, for our employment data with X =Age, α=Income Quantile.

In these plots we can see that the interpretation of the CQTE is less immediate than for the case of the CQC. This is because, for each covariate value, the quantile value corresponds to a different untreated response making comparisons between various values of the covariates less direct. From this plot we do still see that higher income quantiles are associated with a greater increase in wages. Interestingly the value of the CQTE plummets 90% at the highest quantile for individuals of age 40. This is likely an estimation error due to the limited data at higher quantiles as and higher ages.

B.3 Colon cancer treatment

To demonstrate the effectiveness of our approach in medical settings, we apply it to a trial on the effect of colon cancer treatment on survival time/time to remission. This dataset was originally introduced in Laurie et al. [19] and can be found in the “survival" package in R and loaded with the line `data(colon, package="survival")`. In this dataset 929 are randomised to either receive Placebo or Levamisole. They are then followed up for a period of up to 3329 days and the time till

https://doi.org/10.52202/079017-3260

either death or recurrence of their cancer. For our analysis we take our response (Y) as the time to first of death or recurrence. For simplicity, when an individual makes it to the end of a trial without an event and is censored we take that to be the time of their event. Looking at the data censoring times are mostly at around 2,000-3,000 days while over 90% of death or recurrence events that do occur, occur within 1,500 days. Therefore censored individuals will still have better responses than those who had events, as we would want. As our covariate, we took the patients' ages when joining the trial. We show a 3D plot and heat plot of the estimate of $\Delta(y|\boldsymbol{x}) = g^*(y|\boldsymbol{x}) - y$ in Figure 10.

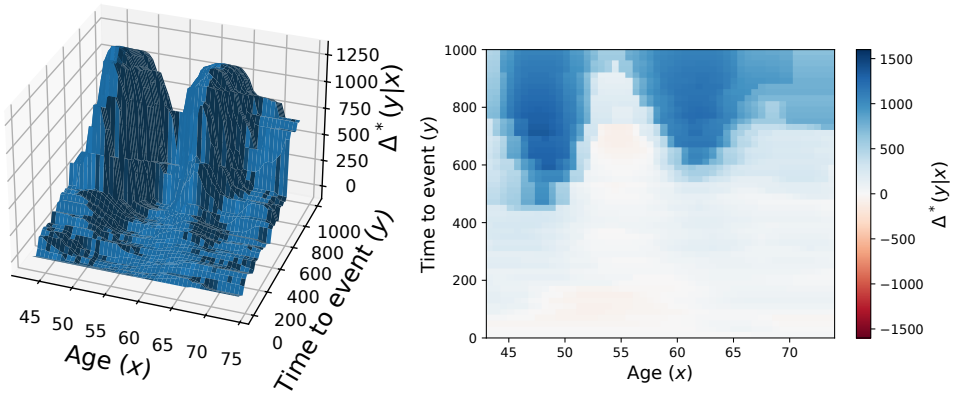


Figure 10: Surface plot and heat plot of $\Delta(y|\boldsymbol{x})$ over $y, \boldsymbol{x}$ for colon cancer trial data with X =Age, Y=Time to Event.

Interestingly from Figure 10 we see that the treatment appears to have little effect on patients aged 53-57. For the reaming patients, we can see that there is relatively little effect on the time to event for events of 400 days or less while at around 500 days the difference in time to event jumps to 1000 days. This jump takes the time to event (TTE) to the time when censoring begins. This gives evidence that rather than the treatment delaying the time to event, it increases the proportion of the patients who have no event during the trial essentially fully treating those patients for the duration of the trial. This shows the value of the CQC it was able to reveal meaningful information about the treatment beyond simply its positive effect. On top of this, we are able to identify the censoring within our results without explicitly controlling for it in any way.

C Additional theory

C.1 Proof of $\hat{h}$ accuracy

We first define some additional notation. For an output f which depends upon $z = (y, x, a)$ define $m_f(\boldsymbol{x}) = \mathbb{E}[f(Z)|X = \boldsymbol{x}]$ and

$$\hat{m}_f(\boldsymbol{x}) = \sum_{j \in \mathcal{J}} w_j f(Z_j),$$

with w_j defined as before so that $\hat{m}_{\hat{\varphi}}(\boldsymbol{x})$ is our estimator and $\hat{m}_{\varphi}(\boldsymbol{x})$ is the oracle estimator.

Additionally define

$$\hat{b}(\boldsymbol{x}) \coloneqq m_{\hat{\varphi}-\varphi}(\boldsymbol{x}),$$

in other words the bias in our estimate of the pseudo-outcome for a given $\boldsymbol{x}$. We will also work with $\hat{m}_b(\boldsymbol{x})$ where we view b as being a function of z.

Finally for $n \in \mathbb{N}$, $\delta \in (0, 1)$, define

$$\varepsilon_\delta(n) \coloneqq \begin{cases} \log(2/\delta)/n & \text{if } \delta \in (0, 2/e] \\ 1/n & \text{otherwise.} \end{cases}.$$

Theorem 3 (Stability Result)**.** *For $\delta \in (0, 1)$ we have that with probability at least $1 - \delta$*

$$|\hat{m}_\varphi(\boldsymbol{x}) - m_\varphi(\boldsymbol{x})| \leq \mathbb{E}[\hat{m}_\varphi(\boldsymbol{x}) - m_\varphi(\boldsymbol{x})|X_{\mathcal{J}}, D_{\mathcal{I}}]| \tag{12}$$
$$+ \sqrt{2\varepsilon_\delta(n)}\|\boldsymbol{w}\|\|\varphi - h(y_0, y_1|.)\|_{\boldsymbol{w}^2} + \frac{2\|\boldsymbol{w}\|_\infty \varepsilon_\delta(n)}{3\xi} =: \mathbb{B}(\varphi).$$

Moreover, with probability at least $1-\delta$ *we have*

$$|\hat{m}_{\hat{\varphi}}(\boldsymbol{x}) - m_{\varphi}(\boldsymbol{x})| \leq \mathbb{B}(\varphi) + |\hat{m}_{\hat{b}}(\boldsymbol{x})| + \sqrt{2\varepsilon_{\delta}(n)}\|\boldsymbol{w}\|_2 \|\hat{\varphi} - \varphi\|_{\boldsymbol{w}^2} =: \mathbb{B}^{+}(\hat{\varphi}). \tag{13}$$

Proof. To prove (12) we first observe that since $\boldsymbol{w}$ is $D_{\mathcal{I}}, X_{\mathcal{J}}$-measurable we have

$$\mathbb{E}[\hat{m}_{\varphi}(\boldsymbol{x})^2 | D_{\mathcal{I}}, X_{\mathcal{J}}] - \mathbb{E}[\hat{m}_{\varphi}(\boldsymbol{x}) | D_{\mathcal{I}}, X_{\mathcal{J}}]^2 = \sum_{j \in \mathcal{J}} w_j^2 \mathbb{E}[(\varphi(Z_j) - \mathbb{E}[\varphi(Z)|X = X_j])^2 | D_{\mathcal{I}}, X_{\mathcal{J}}]$$

Note also that we have $\max_{j\in\mathcal{J}} |w_j\{\varphi(Z_j) - \mathbb{E}[\varphi(Z_j)|D_{\mathcal{I}}, X_{\mathcal{J}}]\}| \leq \|\boldsymbol{w}\|_{\infty} 2\xi^{-1}$ almost surely. Note also that $\{\varphi(Z_j)\}_j \in \mathcal{J}$ are conditionally independent given $D_{\mathcal{I}}, X_{\mathcal{J}}$. Hence, the bound (12) follows from two applications Bernstein's inequality Boucheron et al. [7, Theorem 2.10], applied conditionally on $D_{\mathcal{I}}, X_{\mathcal{J}}$.

Next we note that by the triangle inequality we have

$$\begin{aligned}
&|\mathbb{E}\{\hat{m}_{\hat{\varphi}}(\boldsymbol{x}) - m_{\varphi}(\boldsymbol{x}), | \, D_{\mathcal{I}}, X_{\mathcal{J}}\}| \\
&\leq |\mathbb{E}\{\hat{m}_{\hat{\varphi}}(\boldsymbol{x}) - \hat{m}_{\varphi}(\boldsymbol{x}) \,|\, D_{\mathcal{I}}, X_{\mathcal{J}}\}| + |\mathbb{E}\{\hat{m}_{\varphi}(\boldsymbol{x}) - m_{\varphi}(\boldsymbol{x}) \,|\, D_{\mathcal{I}}, X_{\mathcal{J}}\}| \\
&\leq \left| \sum_{j \in \mathcal{J}} w_j \mathbb{E}\left[\hat{\varphi}(Z_j) - \varphi(Z_j) | D_{\mathcal{I}}, X_{\mathcal{J}}\right] \right| + |\mathbb{E}\{\hat{m}_{\varphi}(\boldsymbol{x}) - m_{\varphi}(\boldsymbol{x}) \,|\, D_{\mathcal{I}}, X_{\mathcal{J}}\}| \\
&= \|\hat{m}_{\hat{b}}\|_{\boldsymbol{w},1} + |\mathbb{E}\{\hat{m}_{\varphi}(\boldsymbol{x}) - m_{\varphi}(\boldsymbol{x}) \,|\, D_{\mathcal{I}}, X_{\mathcal{J}}\}|
\end{aligned}$$

Moreover, since $\mathbb{E}[\hat{\varphi}(Z_j)|D_{\mathcal{I}}, X_{\mathcal{J}}]$ is the projection of $\hat{\varphi}(Z_j)$ onto the subspace of $D_{\mathcal{I}}, X_{\mathcal{J}}$-measureable functions we have

$$\begin{aligned}
&\|\boldsymbol{w}\|_2^2 \, \|\hat{m}_{\hat{\varphi}} - m_{\hat{\varphi}}(\boldsymbol{x})\|_{\boldsymbol{w}^2}^2 \\
&= \sum_{j \in \mathcal{J}} w_j^2 \, \mathbb{E}\{(\hat{\varphi}(Z) - \mathbb{E}\{\hat{\varphi}(Z)|X = X_j\})^2 | D_{\mathcal{I}}, X_{\mathcal{J}}\} \\
&\leq \sum_{i \in [n]} w_j^2 \, \mathbb{E}\{(\hat{\varphi}(Z_j) - \mathbb{E}\{\varphi(Z_j)|D_{\mathcal{I}}, X_{\mathcal{J}}\})^2 | D_{\mathcal{I}}, X_{\mathcal{J}}\} \\
&= \sum_{i \in [n]} w_j^2 \, \mathbb{E}(\{(\hat{\varphi}(Z_j) - \varphi(Z_j)) + (\varphi(Z_j) - \mathbb{E}\{\varphi(Z_j)|D_{\mathcal{I}}, X_{\mathcal{J}}\})\}^2 | D_{\mathcal{I}}, X_{\mathcal{J}}) \\
&= \sum_{i \in [n]} w_j^2 \, (\mathbb{E}\{(\hat{\varphi}(Z_j) - \varphi(Z_j))^2 | D_{\mathcal{I}}, X_{\mathcal{J}}\} + \mathbb{E}\{(\varphi(Z_j) - \mathbb{E}\{\varphi(Z_j)|D_{\mathcal{I}}, X_{\mathcal{J}}\})^2 | D_{\mathcal{I}}, X_{\mathcal{J}}\}) \\
&= \|\boldsymbol{w}\|_2^2 \, \{(\|\hat{m}_{\hat{\varphi}} - \hat{m}_{\varphi}\|_{\boldsymbol{w}^2}^2 + \|\hat{m}_{\varphi} - m_{\varphi}\|_{\boldsymbol{w}^2}^2\} \\
&\leq \|\boldsymbol{w}\|_2^2 \, \{(\|\hat{m}_{\hat{\varphi}} - \hat{m}_{\varphi}\|_{\boldsymbol{w}^2} + \|\hat{m}_{\varphi} - m_{\varphi}\|_{\boldsymbol{w}^2}\}^2.
\end{aligned}$$

Hence, to deduce (13) we apply the first bound with $\hat{\varphi}$ in place of φ to obtain the following bound with probability at least $1-\delta$,

$$\begin{aligned}
|\hat{m}_{\varphi}(\boldsymbol{x}) - m_{\varphi}(\boldsymbol{x})| \leq & |\mathbb{E}\{\hat{m}_{\hat{\varphi}}(\boldsymbol{x}) - m_{\varphi}(\boldsymbol{x}) \,|\, D_{\mathcal{I}}, X_{\mathcal{J}}\}| + \sqrt{2\varepsilon_{\delta}(n)} \, \|\boldsymbol{w}\|_2 \, \|\hat{m}_{\hat{\varphi}} - m_{\hat{\varphi}}\|_{\boldsymbol{w}^2} \\
& + 2\xi^{-1}\|\boldsymbol{w}\|_{\infty} \, \varepsilon_{\delta}(n)/3 \\
\leq & \|\hat{m}_{\hat{b}}\|_{\boldsymbol{w},1} + |\mathbb{E}\{\hat{m}_{\varphi}(\boldsymbol{x}) - m_{\varphi}(\boldsymbol{x}) \,|\, D_{\mathcal{I}}, X_{\mathcal{J}}\}| \\
& + \sqrt{2\varepsilon_n(\delta)}\|\boldsymbol{w}\|_2 \, \{(\|\hat{m}_{\hat{\varphi}} - \hat{m}_{\varphi}\|_{\boldsymbol{w}^2} + \|\hat{m}_{\varphi} - m_{\varphi}\|_{\boldsymbol{w}^2}\} \\
& + 2\xi^{-1}\|\boldsymbol{w}\|_{\infty} \, \varepsilon_{\delta}(n)/3 \\
= & \, \mathbb{B}(\varphi) + \|\hat{m}_{\hat{b}}\|_{\boldsymbol{w},1} + \sqrt{2\varepsilon_{\delta}(n)} \, \|\boldsymbol{w}\|_2 \, \|\hat{m}_{\hat{\varphi}} - \hat{m}_{\varphi}\|_{\boldsymbol{w}^2} \\
= & \, \mathbb{B}^{+}(\hat{\varphi}),
\end{aligned}$$

as required. □

We apply the following result from Kennedy [15].

 https://doi.org/10.52202/079017-3260

Proposition 4 (Proposition 2 from Kennedy [15]). *Suppose that* $\hat{b}(\boldsymbol{x}) = \hat{b}_1(\boldsymbol{x})\hat{b}_2(\boldsymbol{x})$ *then*

$$|\hat{m}_{\hat{b}}(\boldsymbol{x})| = \|\boldsymbol{w}\|_1 \|\hat{b}_1\|_{\boldsymbol{w}} \|\hat{b}_1\|_{\boldsymbol{w}}.$$

Proof. Follows from the Cauchy-Schwartz inequality. □

Proposition 5. *Our pseudo-outcome is conditionally unbiased,*

$$\mathbb{E}[\varphi_{y_0,y_1}(Z)|X = \boldsymbol{x}] = h^*(y_0, y_1|\boldsymbol{x}).$$

Proof. We have

$$\begin{aligned}\mathbb{E}[\varphi_{y_0,y_1}(Z)|X = \boldsymbol{x}, A = 1] &= \frac{1}{\pi(\boldsymbol{x})}\Big(\mathbb{E}[\mathbb{1}\{Y \le y_1\}|X = \boldsymbol{x}, A = 1] - F_1(y_1|\boldsymbol{x})\Big) + h^*(y_0, y_1|\boldsymbol{x}) \\ &= h^*(y_0, y_1|\boldsymbol{x}),\end{aligned}$$

and similarly $\mathbb{E}[\varphi_{y_0,y_1}(Z)|X = \boldsymbol{x}, A = 1] = h^*(y_0, y_1|\boldsymbol{x})$. The result now follows by the law of total expectation. □

C.1.1 Proof of Proposition 1

Proof of Proposition 1. Fix $y_0, y_1, \boldsymbol{x}$ Let $\hat{m}_{\hat{\varphi}}(\boldsymbol{x})$ be the regression of X against the estimated pseudo-outcome $\hat{\varphi}_{y_0,y_1}(Z)$ evaluated at $\boldsymbol{x}$ (i.e. $\hat{h}(y_0, y_1|\boldsymbol{x})$), let $\hat{m}_{\varphi}$ be the same but with the estimated pseudo-outcome replaced with the exact pseudo-outcome φ_{y_0,y_1}. Finally take

$$m_{\varphi}(\boldsymbol{x}) \coloneqq \mathbb{E}[\varphi_{y_0,y_1}(Z)|X = \boldsymbol{x}] = \mathbb{P}(Y \le y_1|X = \boldsymbol{x}, A = 1) - \mathbb{P}(Y \le y_1|X = \boldsymbol{x}, A = 0).$$

Theorem 3 gives us that with probability. at least $1 - \delta$

$$\begin{aligned}|\hat{m}_{\hat{\varphi}}(\boldsymbol{x}) - m_{\varphi}(\boldsymbol{x})| &\le \varepsilon_{\gamma}(n, \delta) + |\hat{m}_{\hat{b}}(\boldsymbol{x})| + \sqrt{2\varepsilon_{\delta}(n)}\|\boldsymbol{w}\|_2 \|\hat{\varphi} - \varphi\|_{\boldsymbol{w}^2} \\ &\le \varepsilon_{\gamma}(n, \delta) + |\hat{m}_{\hat{b}}(\boldsymbol{x})| + \sqrt{2\varepsilon_{\delta}(n)}\|\hat{\varphi} - \varphi\|_{\boldsymbol{w}^2}\end{aligned}$$

with $\hat{b}((y, \boldsymbol{x}, a)) \coloneqq \mathbb{E}[\hat{\varphi}(Z) - \varphi(Z)|X = \boldsymbol{x}]$.

To bound $\hat{b}$ we have that from Proposition 5,

$$\mathbb{E}[\varphi_{y_0,y_1}(Z)|X = \boldsymbol{x}] = \mathbb{P}(Y \le y_1|X = \boldsymbol{x}, A = 1) - \mathbb{P}(Y \le y_0|X = \boldsymbol{x}, A = 0).$$

Additionally, by splitting over the events $\{A = 1\}, \{A = 0\}$ and noting that $\mathbb{P}(A = 1|X = \boldsymbol{x}) = \pi(\boldsymbol{x})$ we have

$$\begin{aligned}\mathbb{E}[\hat{\varphi}_{y_0,y_1}(Z)|X = \boldsymbol{x}] = &\left(\frac{\pi(\boldsymbol{x})}{\hat{\pi}(\boldsymbol{x})}\right)\Big(\mathbb{P}(Y \le y_1|X = \boldsymbol{x}, A = 1) - \hat{\mathbb{P}}(Y \le y_1|X = \boldsymbol{x}, A = 1)\Big) \\ &- \frac{1 - \pi(\boldsymbol{x})}{1 - \hat{\pi}(\boldsymbol{x})}\Big(\mathbb{P}(Y \le y_0|X = \boldsymbol{x}, A = 0) - \hat{\mathbb{P}}(Y \le y_0|X = \boldsymbol{x}, A = 0)\Big) \\ &+ \hat{\mathbb{P}}(Y \le y_1|X = \boldsymbol{x}, A = 1) - \hat{\mathbb{P}}(Y \le y_0|X = \boldsymbol{x}, A = 0).\end{aligned}$$

Hence

$$\begin{aligned}\hat{b}(\boldsymbol{x}) = &\left(\frac{\pi(\boldsymbol{x})}{\hat{\pi}(\boldsymbol{x})} - 1\right)\Big(\mathbb{P}(Y \le y_1|X = \boldsymbol{x}, A = 1) - \hat{\mathbb{P}}(Y \le y_1|X = \boldsymbol{x}, A = 1)\Big) \\ &- \left(\frac{1 - \pi(\boldsymbol{x})}{1 - \hat{\pi}(\boldsymbol{x})} - 1\right)\Big(\mathbb{P}(Y \le y_0|X = \boldsymbol{x}, A = 0) - \hat{\mathbb{P}}(Y \le y_0|X = \boldsymbol{x}, A = 0)\Big).\end{aligned}$$

We have that using Proposition 4

$$\begin{aligned}|\hat{m}_{\hat{b}}(\boldsymbol{x})| &\leq \|\boldsymbol{w}\|_1 \left\|\frac{\pi}{\hat{\pi}} - 1\right\|_{\boldsymbol{w}} \left\|\mathbb{P}(Y \leq y_1|X, A=1) - \hat{\mathbb{P}}(Y \leq y_1|X, A=1)\right\|_{\boldsymbol{w}} \\ &+ \left\|\frac{1-\pi}{1-\hat{\pi}} - 1\right\|_{\boldsymbol{w}} \left\|\mathbb{P}(Y \leq y_0|X, A=0) - \hat{\mathbb{P}}(Y \leq y_0|X, A=0)\right\|_{\boldsymbol{w}} \\ &\leq \frac{1}{\xi}\sum_{a=0}^{1} \|\pi - \hat{\pi}\|_{\boldsymbol{w}} \left\|\mathbb{P}(Y \leq y_a|X, A=a) - \hat{\mathbb{P}}(Y \leq y_a|X, A=a)\right\|_{\boldsymbol{w}}.\end{aligned}$$

Now it is simply a matter of bounding all the relevant terms which we do through the following events

$$\begin{aligned}E_{\text{oracle}} &\coloneqq \left\{|\hat{m}_{\hat{\varphi}} - m_\varphi| \leq |\hat{m}_\varphi(\boldsymbol{x}) - m_\varphi(\boldsymbol{x})| + |\hat{m}_{\hat{b}}(\boldsymbol{x})| + \sqrt{2\varepsilon_\delta(n)}\|\boldsymbol{w}\|_2 \|\hat{\varphi} - \varphi\|_{\boldsymbol{w}^2}\right\} \\ E_{\hat{\varphi}} &\coloneqq \{\|\hat{\varphi} - \varphi\|_{\boldsymbol{w}^2} \leq \varepsilon_{\hat{\varphi}}(n, \delta/4)\} \\ E_\gamma &\coloneqq \{\hat{m}_\varphi - m_\varphi \leq \varepsilon_\gamma(n, \delta/4)\} \\ E_\alpha &\coloneqq \left\{\|\pi - \hat{\pi}\|_{\boldsymbol{w},2} \leq \varepsilon_\alpha(n, \delta/4)\right\} \\ E_{\beta,0} &\coloneqq \left\{\|\hat{\mathbb{P}}(Y \leq y_0|X, A=0) - \mathbb{P}(Y \leq y_0|X, A=0)\|_{\boldsymbol{w},2} \leq g_\beta(n, \delta/4)\right\} \\ E_{\beta,1} &\coloneqq \left\{\|\hat{\mathbb{P}}(Y \leq y_1|X, A=1) - \mathbb{P}(Y \leq y_1|X, A=1)\|_{\boldsymbol{w},2} \leq \varepsilon_\beta(n, \delta/4)\right\}\end{aligned}$$

According to our assumptions and previous results, $E_{\text{oracle}}, E_{\hat{\varphi}}, E_\gamma, E_\alpha \cap E_{\beta,0} \cap E_{\beta,1}$ separately occur w.p. at least $1 - \delta/4$. Then by the union bound, the intersection of all these events holds w.p. at least $1 - \delta$ and under these events

$$|\hat{m}_{\hat{\varphi}} - \hat{m}_\varphi + \hat{m}_\varphi - m_\varphi| \leq \varepsilon_{T_n}(n, \delta/4) + \frac{2}{\xi}\varepsilon_\alpha(n, \delta/4)\varepsilon_\beta(n, \delta/4) + \varepsilon_\gamma(n, \delta/4)\varepsilon_{\boldsymbol{w}}(n, \delta/4)$$

Where the first term comes from Proposition 3, the second from Proposition 4 and the final term from smoothness of g^*. □

C.2 Proof of $\hat{g}$ accuracy

Proposition 6 (Maximum step-size bound)**.** *Suppose that both the outer regression and estimation of CCDFs is fit using linear smoothers so that*

$$\hat{h}(y_0, y_1|\boldsymbol{x}) = \sum_{j \in \mathcal{J}} w_j(\boldsymbol{x}; X_{\mathcal{J}})\hat{\varphi}(Z_j) \qquad \hat{F}_a(y|\boldsymbol{x}') = \sum_{i \in \mathcal{I}} w_{F_a;i}(\boldsymbol{x}'; X_{\mathcal{J}})\mathbb{1}\{Y_i \leq y\}$$

with $w_j(\boldsymbol{x}), w_i(\boldsymbol{x}')$ Additionally suppose with probability at least $1 - \delta$

$$\max\left\{\max_{j \in \mathcal{J}} w_j(\boldsymbol{x}), \max_{i \in \mathcal{I}} w_{F_a;i}(X)\right\} \leq \varepsilon_{\boldsymbol{w}}(n, \delta).$$

Then with probability at least $1 - \delta$

$$\max_{j \in [n_1+1]_0} |\hat{h}(y_0, Y_1^{(j)}|\boldsymbol{x}) - \hat{h}(y_0, Y_1^{(j+1)}|\boldsymbol{x})| \geq \xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n)$$

where $[n_1 + 1]_0 \coloneqq [n_1 + 1] \cup \{0\}$, $\hat{h}(y_0, Y_1^{(0)}|\boldsymbol{x}) = -1$, $\hat{h}(y_0, Y_1^{(n_1+1)}|\boldsymbol{x})$, and $\{Y^{(i)}\}_{i \in I_1}$ is the sorted version of $\{Y_i\}_{i \in I_1}$.

Proof. We immediately have via union bounds that with probability at least $1 - \delta$

$$\max\left\{\max_{j \in \mathcal{J}} w_j(\boldsymbol{x}), \max_{j \in \mathcal{J}} w_{F_a;i}(X_j)\right\} \leq \varepsilon_{\boldsymbol{w}}(n, \delta/n).$$

Now we can try to bound the jump size of our function. To do so for a step function $f : \mathcal{Y} \to \mathbb{R}$ and a step points $y \in \mathcal{Y}$ we define

$$f(\Delta y) \coloneqq \lim_{\varepsilon \downarrow 0} f(y) - f(y - \varepsilon),$$

 https://doi.org/10.52202/079017-3260

i.e. the size of the step at y. From the definition of our pseudo-outcome, the jumps occur in different forms at $\{Y_j\}_{\{j\in\mathcal{J},A_j=1\}}$, $\{Y(i)\}_{\{i\in\mathcal{I}|A^{(i)}=1\}}$. Note that we are assuming Y continuous so these points are all distinct a.s. .

For Y_j with $j\in\mathcal{J}$ and $A_j=1$. We have

$$\begin{aligned}\hat{h}(y_0,\Delta Y_j|\boldsymbol{x}) &= w_{\varphi,j}\frac{1}{\hat{\pi}(X_j)}\mathbb{1}\{Y_j\le\Delta Y_j\}\\ &\le \frac{w_{\varphi,j}}{\xi}.\end{aligned}$$

For $i\in\mathcal{I}$ with $A^{(i)}=1$, we have that

$$\begin{aligned}|\hat{h}(y_0,\Delta Y_i|\boldsymbol{x})| &= \left|\sum_{j\in\mathcal{J}} w_j\left(-\mathbb{1}\{A_j=1\}\frac{1}{\hat{\pi}(X_j)}\hat{F}_1(\Delta Y_i|X_j)+\hat{F}_1(\Delta Y_i|X_j)\right)\right|\\ &\le \frac{1}{\xi}\max_{j\in\mathcal{J}}\hat{F}_1(\Delta Y_i|X_j)\\ &\le \frac{1}{\xi}\max_{j\in\mathcal{J}} w_{F_a;i}(X_j)\mathbb{1}\{Y_i\le\Delta Y_i\}\\ &= \frac{1}{\xi}\max_{j\in\mathcal{J}} w_{F_a;i}(X_j).\end{aligned}$$

Hence the maximum jump size is less than

$$\frac{1}{\xi}\max\left\{\max_{j\in\mathcal{J}} w_{\varphi,j}(\boldsymbol{x}),\max_{i\in\mathcal{I},j\in\mathcal{J}} w_{F_a;i}(X_j)\right\}.$$

Therefore combining this with our previous bounds gives that with probability at least $1-\delta$

$$\max_{i\in[n_1+1]_0}|\hat{h}(y_0,Y^{(i)}|\boldsymbol{x})-\hat{h}(y_0,Y^{(i+1)}|\boldsymbol{x})|\le\xi^{-1}\varepsilon_{\boldsymbol{w}}(n,\delta/n)$$

□

F_1 We also have a result ensuring the step size of the projected function is even smaller.

Proposition 7. *Let $\boldsymbol{\alpha}\in\mathbb{R}^p$ and $\tilde{\boldsymbol{\alpha}}\coloneqq P_{\mathrm{Iso}(p)}(\boldsymbol{\alpha})$. Then $|\alpha_l-\alpha_{l+1}|\ge|\tilde{\alpha}_l-\tilde{\alpha}_{l+1}|$ for all $l\in[p]$.*

Proof. We will prove this by first giving an algorithm to compute the projection. Define

$$\psi_l(\boldsymbol{\alpha})\coloneqq\begin{cases}\boldsymbol{\alpha} & \text{if } \alpha_l\le\alpha_{l+1}\\ \left(\alpha_1,\ldots,\frac{\alpha_l+\alpha_{l+1}}{2},\frac{\alpha_l+\alpha_{l+1}}{2},\alpha_m\right) & \text{if } \alpha_l>\alpha_l+1.\end{cases}$$

Now define $\boldsymbol{\alpha}^{(0,0)}=\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^{(n,i)}=\boldsymbol{\psi}_i(\alpha^{(n,i-1)})$ for $i\in[p-1]$ and $\boldsymbol{\alpha}^{(n,0)}=\boldsymbol{\alpha}^{(n-1,p-1)}$ for $n\in\mathbb{N}$. Finally define $\boldsymbol{\alpha}^{(t)}=\boldsymbol{\alpha}^{(\lfloor t/p\rfloor,\mathrm{mod}(t,p)}$. Then by Yang and Barber [32] we know that $\lim_{t\to\infty}\boldsymbol{\alpha}^{(t)}=P_{\mathrm{Iso}(p)}(\boldsymbol{\alpha})$.

Now if $|\alpha_{l+1}^{(n,l+1)}-\alpha_l^{(n,l+1)}|>|\alpha_{l+1}^{(n,l)}-\alpha_l^{(n,l)}|$. Then $\alpha_{l+1}^{(n,l+1)}$ has be moved by ψ_{l+1}. Hence $\alpha_{l+1}^{(n,l+1)}<\alpha_{l+1}^{(n,l)}$. For this to increase the distance then we must have

$$\alpha_{l+1}^{(n,l+1)}<\alpha_l^{(n,l+1)}.$$

Therefore as $\alpha_l^{(n+1,l-1)}\ge\alpha_l^{n,l+1}$, we have that

$$\alpha_{l+1}^{(n+1,l)}=\alpha_l^{(n+1,l)}.$$

By a similar (simpler) argument $|\alpha_l^{(n,l-1)}-\alpha_{l+1}^{(n,l-1)}|>|\alpha_l^{(n,l-2)}-\alpha_{l+1}^{(n,l-2)}|$ then $\alpha_l^{(n,l)}=\alpha_{l+1}^{(n,l)}$. Hence if any iteration moves adjacent points further apart, a later iteration will make them equal meaning that on convergence they will be equal. Hence we have proved our claim. □

We now have the result which justifies our choice to take the isotonic projection of the data which is take from Yang and Barber [32].

Proposition 8 (Theorem 1 from Yang and Barber [32]). *Let $\boldsymbol{z}^*, \hat{\boldsymbol{z}} \in \mathbb{R}^p$ with $\boldsymbol{z}^* \in \mathrm{Iso}(p)$. Then for $\tilde{\boldsymbol{z}} \coloneqq P_{\mathrm{Iso}(p)}(\hat{\boldsymbol{z}})$,*

$$\max_l |z_l^* - \tilde{z}_l| \leq \max_l |z_l^* - \hat{z}_l|.$$

We now use this isotonic projection to obtain supremum bounds on the accuracy of our estimator.

Proposition 9 (Supremum bound on $\hat{h}$ accuracy). *Fix $\boldsymbol{x} \in \mathcal{X}, y \in y_0$ and let $\hat{h}$ be our original estimate of h^*. For $m \in \mathbb{N}$, take $\{Y^{(l)}\}_{l=1}^m$ to be a potentially random set of points in $\mathcal{Y}$ in increasing order.*

Now define $\boldsymbol{\alpha} \in \mathbb{R}^m$ by $\alpha_l \coloneqq \hat{h}(y_0, Y^{(l)}|\boldsymbol{x})$ and $\tilde{\boldsymbol{\alpha}} \coloneqq P_{\mathrm{Iso}(m)}(\boldsymbol{\alpha})$. Finally, define $\tilde{h}$ to be the piecewise constant right continuous function with $\tilde{h}(y_0, Y^{(l)}|\boldsymbol{x}) \coloneqq \tilde{\alpha}_l$.

Suppose for any $y_1 \in \mathcal{Y}$, $n \in \mathbb{N}$, and $\delta, \delta' > 0$ that

$$\mathbb{P}\left(\left|\tilde{h}(y_0, y_1|\boldsymbol{x}) - h^*(y_0, y_1|\boldsymbol{x})\right| \geq \varepsilon_h(n, \delta)\right) \leq \delta$$

$$\mathbb{P}\left(\max_{l \in [m]_0} \left|\hat{h}(y_0, Y^{(l)}|\boldsymbol{x}) - \hat{h}(y_0, Y^{(l+1)}|\boldsymbol{x})\right| \geq \varepsilon_{step}(n, m, \delta)\right) \leq \delta$$

where we take $[m]_0 \coloneqq [m] \cup \{0\}$, $\hat{h}(y_0, Y^{(0)}|\boldsymbol{x}) \coloneqq -1$, and $\hat{h}(y_0, Y^{(m+1)}|\boldsymbol{x}) \coloneqq 1$.

Then for any $\delta, \delta' > 0, n \in \mathbb{N}$,

$$\mathbb{P}\left(\sup_{y_1 \in \mathcal{Y}} \left|\hat{h}(y_0, y_1|\boldsymbol{x}) - h^*(y_0, y_1|\boldsymbol{x})\right| \leq \varepsilon_h(n, \delta/m) + \varepsilon_{step}(n, m, \delta')\right) \geq 1 - \delta - \delta'.$$

Proof. For each $l \in [m]$ define the event

$$E_{h,l} \coloneqq \left\{\left|\hat{h}(y_0, Y_1^{(l)}|\boldsymbol{x}) - h^*(y_0, Y_1^{(l)}|\boldsymbol{x})\right| \leq \varepsilon_h(n, \delta/m)\right\}$$

and take $E_h \coloneqq \bigcap_{l=1}^n E_{h,l}$. Additionally define the event

$$E_{step} \coloneqq \left\{\max_{l \in [m]_0} \left|\hat{h}(y_0, Y^{(l)}|\boldsymbol{x}) - \hat{h}(y_0, Y^{(l+1)}|\boldsymbol{x})\right| < \varepsilon_{step}(n, m, \delta)\right\}$$

Then E_h and E_{step} hold w.p.s at least $1 - \delta$ and $1 - \delta'$ respectively. Also under E_h, E_{step}, by Propositions 7 & 8, we have that

$$\max_{l \in [m]} \left|\tilde{h}(y_0, Y_1^{(l)}|\boldsymbol{x}) - h^*(y_0, Y_1^{(l)}|\boldsymbol{x})\right| \leq \varepsilon_h(n, \delta/m)$$

$$\max_{l \in [m]_0} \left|\tilde{h}(y_0, Y^{(l)}|\boldsymbol{x}) - \tilde{h}(y_0, Y^{(l+1)}|\boldsymbol{x})\right| < \varepsilon_{step}(n, m, \delta)$$

We then have that for any $y_1 \in \mathcal{Y}$ there exists $i \in [m]$ such that $y_1^{(l-1)} \leq y_1 \leq y_1^{(l)}$.

Hence by monotonicity, we have the following 2 inequalities

$$\hat{h}(y_0, y_1|\boldsymbol{x}) - h^*(y_0, y_1|\boldsymbol{x}) \geq \underbrace{\tilde{h}(y_0, Y_1^{(l-1)}|\boldsymbol{x}) - h^*(y_0, Y_1^{(l)}|\boldsymbol{x})}_{\Lambda_1}$$

$$\hat{h}(y_0, y_1|\boldsymbol{x}) - h^*(y_0, y_1|\boldsymbol{x}) \leq \underbrace{\tilde{h}(y_0, Y_1^{(l)}|\boldsymbol{x}) - h^*(y_0, Y_1^{(l-1)}|\boldsymbol{x})}_{\Lambda_2}.$$

Now for the first inequality we have

$$\begin{aligned}\Lambda_1 &= \tilde{h}(y_0, Y_1^{(l-1)}|\boldsymbol{x}) - \tilde{h}^*(y_0, Y_1^{(l)}|\boldsymbol{x}) + \tilde{h}(y_0, Y_1^{(l)}|\boldsymbol{x}) - h^*(y_0, Y_1^{(l)}|\boldsymbol{x}) \\ &\geq -\varepsilon_h(n, \delta) - \varepsilon_{step}(n, m, \delta').\end{aligned}$$

 https://doi.org/10.52202/079017-3260

With the final inequality coming from $E_{h,l-1}$ and our definition of h. By a similar argument we get from $E_{h,l}$ that $\Lambda_2 \leq \varepsilon_h(n,\delta) + \varepsilon_{step}(n,m,\delta')$. Again we can use the same approach to get that under $E_{h,l-1}, E_{h,l}, E_{step}$

$$-\varepsilon_h(n,\delta) - \varepsilon_{step}(n,m,\delta') \leq h^*(y_0,y_1|\boldsymbol{x}) - \tilde{h}(y_0,y_1|\boldsymbol{x}) \leq \varepsilon_h(n,\delta) + \varepsilon_{step}(n,m,\delta').$$

Hence for our specific y_1 under $E_{h,l}, E_{h,l-1}, E_{step}$

$$\left|h^*(y_0,y_1|\boldsymbol{x}) - \tilde{h}(y_0,y_1|\boldsymbol{x})\right| \leq \varepsilon_h(n,\delta) + \varepsilon_{step}(n,m,\delta')$$

Now as this inequality holds for arbitrary y_1 under $E_h \cap E_{step}$ (an event which not depend on our choice of y_1) and the intersection of these two events holds w.p. at least $1-\delta-\delta'$ by the union bound we have our result. □

Our final key result before we can piece them all together will be to obtain accuracy in g from our accuracy in h^*

Theorem 10 (Single point accuracy bound)**.** *Fix $\boldsymbol{x}, y_0$, assume that $h, \tilde{h}$ are strictly monotonic in y_1 and suppose that*

$$\sup_{y_1}|h^*(y_0,y_1|\boldsymbol{x}) - \tilde{h}(y_0,y_1|\boldsymbol{x})| < \varepsilon.$$

Additionally let $\alpha \coloneqq F_0(y_0|\boldsymbol{x})$ and assume that $f_{Y|X,A=1}(y_1|\boldsymbol{x}) > \eta$ for all $y_1 \in F_1^{-1}(B_\varepsilon(\alpha)|\boldsymbol{x})$ where here we are taking $F_1^{-1}(A|\boldsymbol{x})$ to be the pre-image in $\mathcal{Y}$ of the set A for fixed $\boldsymbol{x}$.

Then

$$|g^*(y_0|\boldsymbol{x}) - \hat{g}(y_0|\boldsymbol{x})| \leq \frac{2\varepsilon}{\eta},$$

where $g^*(y_0|\boldsymbol{x})$ is the unique y_1 such that $h^*(y_0,y_1|\boldsymbol{x}) = 0$.

Proof. Let $y_1^* \coloneqq g^*(y_0|\boldsymbol{x})$ and $\hat{y}_1 \coloneqq \hat{g}(y_0|\boldsymbol{x})$ so that $F_1(y_1^*|\boldsymbol{x})$. From our accuracy assumption on $\hat{h}$ in the set-up of the Theorem we have

$$\begin{aligned}
|F_1(y_1^*|\boldsymbol{x}) - F_1(\hat{y}_1|\boldsymbol{x})| &= |h^*(y_0,y_1^*|\boldsymbol{x}) - h^*(y_0,\hat{y}_1|\boldsymbol{x})| \\
&= |h^*(y_0,\hat{y}_1|\boldsymbol{x})| \\
&\leq |h^*(y_0,\hat{y}_1|\boldsymbol{x}) - \hat{h}(y_0,\hat{y}_1|\boldsymbol{x})| + |\hat{h}(y_0,\hat{y}_1|\boldsymbol{x})| \\
&\leq |h^*(y_0,\hat{y}_1|\boldsymbol{x}) - \hat{h}(y_0,\hat{y}_1|\boldsymbol{x})| + |\hat{h}(y_0,y_1^*|\boldsymbol{x})| \\
&\leq |h^*(y_0,\hat{y}_1|\boldsymbol{x}) - \hat{h}(y_0,\hat{y}_1|\boldsymbol{x})| + |\hat{h}(y_0,y_1^*|\boldsymbol{x}) - h^*(y_0,y_1^*|\boldsymbol{x})| \\
&\leq 2\varepsilon,
\end{aligned}$$

where the second and fifth line come from the definition of y_1^* and the fourth from the definition of $\hat{y}_1$.

If we define $\partial_k f$ to be the derivative with respect to the k^{th} argument of f then we have

$$\begin{aligned}
|\hat{g}(y_0|\boldsymbol{x}) - g^*(y_0|\boldsymbol{x})| &= |F_1^{-1}(F_1(y_1^*|\boldsymbol{x})|\boldsymbol{x}) - F_1^{-1}(F_1(\hat{y}_1|\boldsymbol{x})|\boldsymbol{x})| \\
&= \left|\int_{F_1(y_1^*|\boldsymbol{x})}^{F_1(\hat{y}_1|\boldsymbol{x})} \partial_1 F_1^{-1}(\beta|\boldsymbol{x})\mathrm{d}\beta\right| \\
&= \left|\int_{F_1(y_1^*|\boldsymbol{x})}^{F_1(\hat{y}_1|\boldsymbol{x})} \frac{1}{f((F_1^{-1}(\beta|\boldsymbol{x})|\boldsymbol{x})}\mathrm{d}\beta\right| \\
&\leq 2\varepsilon \max_{y_1 \in (y_1^*,\hat{y}_1)} \frac{1}{f(y_1|\boldsymbol{x})} \\
&\leq \frac{2\varepsilon}{\eta}.
\end{aligned}$$

□

C.2.1 Proof of Theorem 2

We can now finally combine all these results to prove Theorem 2.

Proof of Theorem 2. From proposition 1 we have that

$$\mathbb{P}\left(\left|\hat{h}(y_0, y_1|\boldsymbol{x}) - h^*(y_0, y_1|\boldsymbol{x})\right| \leq \varepsilon_h(n, \delta)\right) \geq 1 - \delta$$

with

$$\varepsilon_h(n, \delta) \coloneqq \varepsilon_{T_n}(n, \delta/4) + \varepsilon_\alpha(n, \delta/4)\varepsilon_\beta(n, \delta/4) + \varepsilon_\gamma(n, \delta/4).$$

Now define $\{Y^{(i)}\}_{i=1}^n$ to be the sorted version of $\{Y_i\}_{i=1}^n$. Then by Proposition 6 we have

$$\mathbb{P}\left(\max_{j\in[n]_0} |\hat{h}(y_0, Y_1^{(i)}|\boldsymbol{x}) - h^*(y_0, Y_1^{(i)}|\boldsymbol{x})| \geq \varepsilon_{h-step}(n, \delta)\right) \leq \delta.$$

Plugging this into Proposition 9 with $\varepsilon_{step}(n, m, \delta)$ replaced by $\xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n))$ and $\varepsilon_h(n, \delta)$ replaced with $\varepsilon_h(n, \delta/2)$ gives

$$\mathbb{P}\left(\sup_{y_1\in\mathcal{Y}} \left|\tilde{h}(y_0, y_1|\boldsymbol{x}) - h^*(y_0, y_1|\boldsymbol{x})\right| \leq \varepsilon_h(n, \delta/(2n)) + \xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n))\right) \geq 1 - \delta.$$

Now let $y_1^* \coloneqq g^*(y_0|\boldsymbol{x})$. Then for any $y_1 \in \mathcal{Y}$ if $|y_1' - y_1^*| > s$ this implies that $|F_1(y_1'|\boldsymbol{x}) - F_1(y_1^*|\boldsymbol{x})| > s\eta$ by our lower bound on the density in $B_s(y_1^*)$. Hence by the contrapositive, if $|F_1(y_1'|\boldsymbol{x}) - F_1(y_1^*|\boldsymbol{x})| \leq s\eta$ then $|y_1' - y_1^*| < s$.

Now as

$$\begin{aligned} |F_1(y_1'|\boldsymbol{x}) - F_1(y_1^*|\boldsymbol{x})| \leq s\eta &\Leftrightarrow y_1' \in F_1^{-1}(B_{s\eta}(F_1(y_1^*))) \\ &\Leftrightarrow y_1' \in B_{s\eta}F_1^{-1}(B_{s\eta}(F_0(y_0))) \end{aligned}$$

Hence if n is sufficiently large so that $\varepsilon \coloneqq \varepsilon_h(n, \delta/(2n)) + \xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n)) \leq \eta s$ then we satisfy the bounded density condition of Theorem 10. Therefore we can plug our bound into theorem 10 gives

$$\mathbb{P}\left(|\hat{g}(y_0|\boldsymbol{x}) - g^*(y_0|\boldsymbol{x})| \leq 2\left(\eta^{-1}\varepsilon_h(n, \delta/(2n)) + \xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n))\right)\right) \geq 1 - \delta.$$

□

C.3 Extension to expectation

Proposition 11. *Let Y_0, D be RVs on $\mathcal{Y}, \mathcal{Z}^n$ respectively (with D representing data used to fit a model and Y_0 representing the point where the model is fit). Now take $l(Y_0, D)$ to be a non-negative bounded loss so that $l(Y_0, D) < l_{\max}$ a.s. . Suppose that for any $\delta > 0$, for all $y \in \mathcal{Y}$*

$$\mathbb{P}(l(y, D) > \varepsilon(n, \delta)) < \delta$$

Then for any $t, \delta_0 \in [0, 1]$ and $p, q \in [1, \infty]$ such that $1/p + 1/q = 1$

$$\mathbb{P}\left(\mathbb{E}[l(Y_0, D)|D] \leq t^{1/q}\mathbb{E}[l(Y_0, D)^p|D]^{1/p} + \varepsilon(n, \delta_0)\right) \geq 1 - \frac{\delta_0}{t}.$$

In particular if $l(y, D) < l_{\max}$ a.s. then taking $q = 1, p = \infty$ yields

$$\mathbb{P}\left(\mathbb{E}[l(Y_0, D)|D] \leq tl_{\max} + \varepsilon(n, \delta_0)\right) \geq 1 - \frac{\delta_0}{t}.$$

Proof. First fix $t, \delta_0 \in [0, 1]$ We will first bound the probability that the number of y which don't satisfy our bound isn't too large. We do this by defining the event

$$A \coloneqq \{l(Y_0, D) > \varepsilon(n, \delta_0)\},$$

and now aim to bound the probability that $\mathbb{P}(A|D) > t$ (this probability is just w.r.t Y_0 treating D as fixed.)

 https://doi.org/10.52202/079017-3260

By Markov's inequality,

$$\begin{aligned}\mathbb{P}(\mathbb{P}(A|D) > t) &\leq \frac{1}{t}\mathbb{E}[\mathbb{P}(A|D)] \\ &= \frac{1}{t}\mathbb{E}[\mathbb{P}(A|Y_0)] \quad \text{by Fubini's Theorem} \\ &< \frac{1}{t}\mathbb{E}[\delta_0] = \frac{\delta_0}{t}.\end{aligned}$$

Now define $B \coloneqq \{\mathbb{P}(A|D) \leq t\}$ so that $\mathbb{P}(B) = 1 - \frac{\delta_0}{t}$. Then under B

$$\begin{aligned}\mathbb{E}[l(Y_0, D)|D] &= \mathbb{E}[\mathbb{1}_A l(Y_0, D)|D] + \mathbb{E}[l(Y_0, D)|A^c, D]\mathbb{P}(A^c|D) \\ &\leq \mathbb{E}[\mathbb{1}_A l(Y_0, D)|D] + \varepsilon(n, \delta_0) \quad \text{by definitions of } A \\ &\leq \mathbb{E}[\mathbb{1}_A|D]^{1/q}\mathbb{E}[l(Y_0, D)^p|D]^{1/p} + \varepsilon(n, \delta_0) \quad \text{by Holder's inequality} \\ &\leq t^{1/q}\mathbb{E}[l(Y_0, D)^p|D]^{1/p}t + \varepsilon(n, \delta_0) \quad \text{As we are assuming } B.\end{aligned}$$

□

Corollary 12. *Assume that $F_1(y|\boldsymbol{x}) > \eta$ for all $y_1 \in \mathcal{Y}$. Then, for our fixed $\boldsymbol{x} \in \mathcal{X}$ and $\delta \in (0, e^{-1})$, w.p. at least,*

$$1 - \frac{\delta \operatorname{diam}(\mathcal{Y})}{2\left(\eta^{-1}\varepsilon_h(n, \delta/2n) + \xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n)\right))},$$

$$\mathbb{E}\Big[\, |\hat{g}(Y|\boldsymbol{x}) - g^*(Y|\boldsymbol{x})| \;\Big|A = 0, \hat{g}\Big] \leq 4\left(\eta^{-1}\varepsilon_h(n, \delta/n) + 2\xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n))\right)$$

where $\varepsilon_h(\delta, n) \coloneqq \varepsilon_{T_n}(n, \delta/4) + \varepsilon_\alpha(n, \delta/4)\varepsilon_\beta(n, \delta/4) + \varepsilon_\gamma(n, \delta/4)$.

In particular for $2\left(\eta^{-1}\varepsilon_h(n, \delta/n) + 2\xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n))\right) \leq \log(e_1(n)/\delta)^a/e_2(n)$. *For* $\delta < 1/e$ *we have that w.p. at least* $1 - \delta$

$$\mathbb{E}\Big[\, |\hat{g}(Y|\boldsymbol{x}) - g^*(Y|\boldsymbol{x})| \;\Big|A = 0, \hat{g}\Big] \leq \log(2\operatorname{diam}(\mathcal{Y})e_2(n)e_1(n)/\delta)^a/e_2(n)$$

Proof. Plugging the result of theorem 2 into proposition 11, noting that $|\hat{g}(y_0|\boldsymbol{x}) - g^*(y|\boldsymbol{x})| \leq \operatorname{diam}(\mathcal{Y})$ and taking

$$t = \frac{2}{\operatorname{diam}(\mathcal{Y})}\left(\eta^{-1}\varepsilon_h(n, \delta/(2n)) + \xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n))\right)$$

yields that w.p. at least

$$1 - \frac{\delta \operatorname{diam}(\mathcal{Y})}{2\left(\eta^{-1}\varepsilon_h(n, \delta/n) + \xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n))\right)},$$

$$\mathbb{E}\Big[\, |\hat{g}(Y|\boldsymbol{x}) - g^*(Y|\boldsymbol{x})| \;\Big|A = 0, \hat{g}\Big] \leq 4\left(\eta^{-1}\varepsilon_h(n, \delta/n) + 2\xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n))\right).$$

Following from this if $2\left(\eta^{-1}\varepsilon_h(n, \delta/n) + 2\xi^{-1}\varepsilon_{\boldsymbol{w}}(n, \delta/n))\right) \leq \log(e_1(n)/\delta)^\alpha/e_2(n)$. Then we have

$$\mathbb{P}\left(\mathbb{E}\Big[\, |\hat{g}(Y|\boldsymbol{x}) - g^*(Y|\boldsymbol{x})| \;\Big|A = 0, \hat{g}\Big] \geq 2\log(e_1(n)/\delta)^a/e_2(n)\right) \leq \frac{\delta \operatorname{diam}(\mathcal{Y})e_2(n)}{\log(e_1(n)/\delta)^a}$$

for $\delta < 1/e$ we have

$$\begin{aligned}\delta < 1/e &\Rightarrow \delta < \exp\left\{-(\frac{1}{2})^{1/a}\right\} \\ &\Leftrightarrow \log(1/\delta)^a > \frac{1}{2} \\ &\Rightarrow \log(e_1(n)/\delta)^a > \frac{1}{2} \\ &\Rightarrow 2\delta\log(e_1(n)/\delta)^a > \delta \\ &\Leftrightarrow 2\delta > \frac{\delta}{\log(e_1(n)/\delta)^a}.\end{aligned}$$

Hence

$$\mathbb{P}\left(\mathbb{E}\Big[\,|\hat{g}(Y|\boldsymbol{x}) - g^*(Y|\boldsymbol{x})|\;\Big|A=0,\hat{g}\Big] \geq 2\log(e_1(n)/\delta)^a/e_2(n)\right) \leq 2\delta\,\mathrm{diam}(\mathcal{Y})e_2(n).$$

Finally $\delta' = 2\delta\,\mathrm{diam}(\mathcal{Y})e_2(n)$ gives

$$\mathbb{P}\left(\mathbb{E}\Big[\,|\hat{g}(Y|\boldsymbol{x}) - g^*(Y|\boldsymbol{x})|\;\Big|A=0,\hat{g}\Big] \geq 2\log(2\,\mathrm{diam}(\mathcal{Y})e_1(n)e_2(n)/\delta')^a/e_2(n)\right) \leq \delta'.$$

□

C.4 Application and justification of NW estimation with box kernel

We aim to show that the box kernel satisfies some of our conditions. Specifically conditions 2 & 3 from Proposition 1. We first start by bounding the step size.

Proposition 13 (Effective sample size)**.** *Suppose that for our $\boldsymbol{x} \in \mathcal{X}$ there exists $C_0, r_0 > 0$ such that for any $r \in (0, r_0)$*

$$\mathbb{P}(X \in B_r(\boldsymbol{x})) \geq C_0 r^d.$$

Now for $r \in (0, r_0)$ take our kernel to be $k_r(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \mathbb{1}\{\|\boldsymbol{x} - \boldsymbol{x}'\| \leq r\}$. Then w.p. at least $1-\delta$

$$\sum_{j\in\mathcal{J}} \mathbb{1}\{X_j \in B_r(\boldsymbol{x})\} \geq \left(nC_0r^d - \sqrt{2nC_0r^d\log(1/\delta)} - \log(1/\delta)/3)\right)^{-1}.$$

Proof. We have $\mathbb{1}\{X_j \in B_r(\boldsymbol{x})\} \leq 1$ and $\mathbb{E}[\mathbb{1}\{X_j \in B_r(\boldsymbol{x})\}^2] = \mathbb{E}[\mathbb{1}\{X_j \in B_r(\boldsymbol{x})\}^2] = \mathbb{P}(X_j \in B_r(\boldsymbol{x})) = C_0 r^d$. Therefore by one sided Bernstein's inequality with $\varepsilon = \log(1/\delta)$ we get

$$\mathbb{P}\left(\sum_{j\in\mathcal{J}} \mathbb{1}\{X_j \in B_r(\boldsymbol{x})\} \leq nC_0r^d - \sqrt{2nC_0r^d\log(1/\delta)} - \frac{1}{3}\log(1/\delta)\right) \leq \delta.$$

This gives our desired result. □

Note that $\sum_{j\in\mathcal{J}} \mathbb{1}\{X_j \in B_r(\boldsymbol{x})\}$ is also the effective sample size of our estimation as it is the number of samples used in the average.

Now that we have the effective sample size result in terms of our kernel radius r, we need to obtain the optimal rate of decay of r for our estimation.

Proposition 14 (NW estimation with box kernel)**.** *let $\hat{m}_f(\boldsymbol{x})$ be the NW estimation of $m_f(\boldsymbol{x}) \coloneqq \mathbb{E}[f(Z)|X=\boldsymbol{x}]$ using IID copies $(Z_i)_{i=1}^n$ of Z. and assume $|f(Z)| \leq M$. For a fixed $r \in (0, r)$, use kernel k_r as defined above and suppose the same assumptions hold. Suppose that $m_f(\boldsymbol{x})$ is α smooth for $\alpha \leq 1$ (i.e. α-Holder continuous.) Then for sufficiently large n depending on C_0, α and $\delta \leq 2e^{-1}$ w.p. at least $1-\delta$,*

$$|\hat{m}_f(\boldsymbol{x}) - m_f(\boldsymbol{x})| \leq \frac{2M+1}{C_0}\log(2/\delta)n^{-\frac{1}{2+d/\alpha}}.$$

Proof. With our kernel define $\mathcal{I}_r \coloneqq \{i \in [n] | k_r(\boldsymbol{x}, X_i) = 1\}$ and $n_r = |I_r|$. Now define the event

$$E_n \coloneqq \{n_r \geq C_0 n r^d - \sqrt{2C_0r^d n\log(2/\delta)} - \log(2/\delta)/3\}.$$

 https://doi.org/10.52202/079017-3260

Then by Proposition 13 this event occur w.p. at least $1-\delta/2$. Now if we define $\varepsilon_i \coloneqq f(Z_i)-m_f(X_i)$ then we have $\mathbb{E}[\varepsilon_i|X_i]=0$. Also, we have

$$\begin{aligned}
|\hat{m}_f(\boldsymbol{x})-m_f(\boldsymbol{x})| &= \left|\frac{1}{n_r}\sum_{i\in\mathcal{I}_r} f(Z_i) \;-\; m_f(\boldsymbol{x})\right| \\
&= \left|\frac{1}{n_r}\sum_{i\in\mathcal{I}_r} m_f(X_i)+\varepsilon_i \;-\; m_f(\boldsymbol{x})\right| \\
&\leq \frac{1}{n_r}\sum_{i\in\mathcal{I}_r}|m_f(X_i)-m_f(\boldsymbol{x})| + \left|\frac{1}{n_r}\sum_{i\in\mathcal{I}_r}\varepsilon_i\right| \\
&\leq r^\alpha + \left|\frac{1}{n_r}\sum_{i\in\mathcal{I}_r}\varepsilon_i\right|.
\end{aligned}$$

With the final equality coming by our smoothness condition and definition of $\mathcal{I}_r$. Now by Hoeffding bounds we have that

$$\mathbb{P}\left(\left|\sum_{i\in\mathcal{I}_r}\varepsilon_i\right| \geq \sqrt{\frac{2\log(4/\delta)M^2}{n_r}}\right) \leq \frac{\delta}{2}.$$

Hence by combining this event and E_n then we have that with probability at least $1-\delta$

$$\begin{aligned}
|\hat{m}_f(\boldsymbol{x})-m_f(\boldsymbol{x})| &\leq r^\alpha + \sqrt{\frac{2\log(2/\delta)M^2}{n_r}} \\
&\leq r^\alpha + \sqrt{\frac{2\log(2/\delta)M^2}{C_0 n r^d - \sqrt{2C_0 n r^d \log(2/\delta)} - \log(2/\delta)/3}}.
\end{aligned}$$

Now for sufficiently large n

$$C_0 n r^d - \sqrt{2C_0 n r^d \log(2/\delta)} - \log(2/\delta)/3 \geq \frac{C_0 n r^d}{2\log(2/\delta)}.$$

This in turn gives

$$|\hat{m}_f(\boldsymbol{x})-m_f(\boldsymbol{x})| \leq r^\alpha + \frac{2\log(2/\delta)M}{\sqrt{C_0 n r^d}}.$$

Then the optimal choice of r is such that

$$\begin{aligned}
r^\alpha &= \frac{1}{\sqrt{nr^d}} \\
\Leftrightarrow r^{\alpha+\frac{d}{2}} &= n^{-\frac{1}{2}} \\
\Leftrightarrow r &= n^{-\frac{1}{2\alpha+d}} \\
\Leftrightarrow r^\alpha &= n^{-\frac{1}{2+d/\alpha}} = \frac{1}{\sqrt{nr^d}}.
\end{aligned}$$

Hence plugging this in we get that for sufficiently large n depending on C_0, α we have that for $\delta \leq 2e^{-1}$ w.p. at least $1-\delta$

$$|\hat{m}_f(\boldsymbol{x})-m_f(\boldsymbol{x})| \leq \frac{2M+1}{C_0}\log(2/\delta)n^{-\frac{1}{2+d/\alpha}}.$$

□

Proposition 15 (Final weight decay)**.** *Suppose that for our $\boldsymbol{x}\in\mathcal{X}$ there exists $C_0, r_0 > 0$ such that for any $r\in(0,r)$*

$$\mathbb{P}(X\in B_r(\boldsymbol{x})) \geq C_0 r^d.$$

Suppose that we are performing NW estimation of an α smooth function at points $\boldsymbol{x} \in \mathcal{X}$ using kernel $k_{r_n}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \mathbb{1}\{\|\boldsymbol{x} - \boldsymbol{x}'\| \leq r\}$ with r_n decaying optimally. Now define

$$w_j \coloneqq \frac{k_{r_n}(\boldsymbol{x}, X_j)}{\sum_{j' \in \mathcal{J}} k_{r_n}(\boldsymbol{x}, X_{j'})}$$

the weight of the j^{th} component. Then with probability at least $1 - \delta$

$$\max_{j \in \mathcal{J}} w_j \leq = \frac{\log(2/\delta)}{C_0} n^{\frac{-2}{2+d/\alpha}}.$$

Proof. We have

$$\begin{aligned}\max_{j \in \mathcal{J}} w_j &\leq \frac{1}{\sum_{j \in \mathcal{J}} k(\boldsymbol{x}, X_j)} \\ &= \frac{1}{\sum_{j \in \mathcal{J}} \mathbb{1}\{X_j \in B_r(\boldsymbol{x})\}}.\end{aligned}$$

Hence by proposition 13 we have w.p. at least $1 - \delta$

$$\max_{j \in \mathcal{J}} w_j \leq \left(C_0 n r^d - \sqrt{2 C_0 n r^d \log(1/\delta)} - \log(1/\delta)/3)\right)^{-1}.$$

We know from Proposition 14 that the optimal radius decay gives $\frac{1}{nr^d} = n^{-\frac{2}{2+d/\gamma}}$. Plugging this into our result gives that

$$\begin{aligned}\max_{j \in \mathcal{J}} w_j &\leq \left(C_0 n^{\frac{2}{2+d/\gamma}} - \sqrt{2 C_0 \log(1/\delta)} n^{\frac{1}{2+d/\gamma}} - \log(1/\delta)/3)\right)^{-1} \\ &\leq \frac{\log(1/\delta)}{C_0 n^{\frac{2}{2+d/\gamma}}} \quad \text{For sufficiently large } n \text{ depending on } \gamma, C_0.\end{aligned}$$

□

Note that for a γ smooth function the MSE is $C^{-1} n^{-\frac{1}{2+d/\gamma}}$. Hence the box kernel comfortably satisfies our weight decay condition 2 in Assumptions 2.

Additionally now if we have an α smooth function and assume that for any $\boldsymbol{x}' \in \mathcal{X}$ there exists $C_0(\boldsymbol{x}')$ such that for all $r \in (0, r_0)$ $\mathbb{P}(X \in B_r(\boldsymbol{x}') \geq C_0 r^d$. Then we have that w.p. at least $1 - \delta$,

$$w_{F_a;i}(X, X^{\mathcal{I}}) \leq \frac{\log(2/\delta)}{C_0(X)} n^{-\frac{2}{2+d/\alpha}}.$$

Thus if we assume that there exists $C'' > 0$ such that w.p. at least $1 - \delta$, $\frac{1}{C_0(X)} \leq C'' \sqrt{\log(1/\delta)}$ then we get that w.p. at least $1 - \delta$

$$w_{F_a;i}(X, X^{\mathcal{I}}) \leq C'' \log^{3/2}(3/\delta) n^{-\frac{2}{2+d/\alpha}}.$$

This then gives our condition 2 in Assumptions 2.

We now try to bound $\frac{\sum_{j \in \mathcal{J}} w_j^2 \sigma(X_j)}{\mathbb{E}\left[\sum_{j \in \mathcal{J}} w_j^2 \sigma(X_j)\right]}$.

Corollary 16 (Accuracy under NW estimation with box kernel)**.** *Suppose that our linear smoother is NW estimation with the box kernel and optimally decaying radius additionally assume that:*

- $\varepsilon_{\hat{\varphi}}(n, \delta) = e_{\hat{\varphi}}(n) \sqrt{\log(1/\delta)}$ *with* $e_1 = o(1)$.
- $\varepsilon_{\alpha}(n, \delta) = C_{\alpha} \sqrt{\log(1/\delta)} n^{-\frac{1}{2+d/\alpha}}$.
- $\varepsilon_{\beta}(n, \delta) = C_{\beta} \sqrt{\log(1/\delta)} n^{-\frac{1}{2+d/\beta}}$.

 https://doi.org/10.52202/079017-3260

- $\beta > \frac{d}{2(1+d/\gamma)}$. *Then for any δ, for sufficiently large n the following events each separately hold w.p. at least $1-\delta$,*

$$\left|\hat{h}(y_0, y_1|\boldsymbol{x}) - h^*(y_0, y_1|\boldsymbol{x})\right| \leq C_h \frac{\log^{3/2}(1/\delta)}{e_h(n)}$$
$$|\hat{g}(y|\boldsymbol{x}) - g^*(y|\boldsymbol{x})| \leq \frac{C_g}{\eta\xi} \frac{\log^{3/2}(n/\delta)}{e_h(n)}$$
$$\mathbb{E}\Big[\, |\hat{g}(Y|\boldsymbol{x}) - g^*(Y|\boldsymbol{x})| \;\Big| A = 0, \hat{g}\Big] \leq \frac{C_{g,2}\operatorname{diam}(\mathcal{Y})}{\xi\eta} \frac{\log^{3/2}(e_2(n)n/\delta)}{e_2(n)}$$

with $C_h, C_g, C_{g,2}$ depending on $C_\alpha, C_\beta, C', C''$ (where C', C'' are the constants define in Proposition 15.

Proof. For any NW estimator we have that $\|\boldsymbol{w}\|_1, \|\boldsymbol{w}\|_2 \leq 1$ a.s. meaning we can take $\varepsilon_{\boldsymbol{w}}(n, \delta) \equiv 1$. We also have that $\varepsilon_\gamma(n, \delta) = C_\gamma \log(1/\delta) n^{-\frac{1}{1+d/\gamma}}$. Hence $\varepsilon_{T_n} = \sqrt{2\varepsilon_\delta(n)} e_{\hat{\varphi}}(n) \sqrt{\log(1/\delta)}$. This then gives

$$\begin{aligned}
\left|\hat{h}(y_0, y_1|\boldsymbol{x}) - h^*(y_0, y_1|\boldsymbol{x})\right| &\leq C_\gamma \log^{1/2}(7/\delta) n^{-\frac{1}{2+d/\gamma}} + C_\alpha C_\beta \log(7/\delta) n^{-\left(\frac{1}{2+d/\alpha} + \frac{1}{2+d/\beta}\right)} \\
&\quad + \sqrt{2\varepsilon_{\delta/4}(n)} e_{\hat{\varphi}}(n) \sqrt{\log(7/\delta)} \\
&\leq C_h \log(1/\delta) n^{-\frac{1}{2+d/\gamma}} + \log(1/\delta) n^{-\left(\frac{1}{2+d/\alpha} + \frac{1}{2+d/\beta}\right)} \\
&\leq C_h \frac{\log(1/\delta)}{e_h(n)}
\end{aligned}$$

for $\delta < e^{-1}$ and $C_h = 7(C_\gamma + C_\alpha C_\beta + \sqrt{2})$ giving our first result.

Using our weight decay results for NW estimation and plugging into Proposition 6 we get that

$$\varepsilon_{h-step} = C'' \xi^{-1} \frac{\log^{3/2}}{e_h(n)}(n/\delta).$$

Hence plugging into Theorem 2 gives

$$\begin{aligned}
|\hat{g}(y|\boldsymbol{x}) - g^*(y|\boldsymbol{x})| &\leq C_h \sqrt{2} \eta^{-1} \frac{\log(n/\delta)}{e_h(n)} + 4C'' \xi^{-1} \log^{3/2}(n/\delta) n^{-\frac{1}{2+d/\gamma}} \\
&\leq \frac{C_g}{\xi\eta} \frac{\log^{3/2}(n/\delta)}{e_h(n)}
\end{aligned}$$

with $C_g = C_h\sqrt{2} + 4C''$ giving our second result. Finally by Corollary 12 for $\delta < e^{-1}$ w.p. at least $1-\delta$,

$$\mathbb{E}\left[|\hat{g}(Y|\boldsymbol{x}) - g^*(Y|\boldsymbol{x})| \;\middle|\; A = 0, \hat{g}\right] \leq \frac{C_g \operatorname{diam}(\mathcal{Y})}{\xi\eta} \frac{\log^{3/2}(ne_h(n)/\delta)}{e_h(n)}.$$

□

Note that both ε_{T_n} and ε_{h-step} are actually both $o(\varepsilon_\gamma(n, c))$, thus if we were allowed to take n sufficiently large depending upon δ then we would have all $\log^{3/2}$ terms replaced with $\log$ terms and the dependence on C', C'' removed.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We introduce a a new estimand, the CQC. We claim this to maintain the distributional information of the CQTE which we show through the definition and also on real world scenarios in Section 4. We claim that our method is able to obtain double robustness which we show in proposition 1 and Theorem 2. We also demonstrate its strong performance empirically on simulated data scenarios in Section 4 and Appendix B.1

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a dedicated limitations section where we discuss limitations in our theoretical results regarding both assumptions of sufficiently quickly decaying weights in our linear smoother and requiring smoothness in h^* rather than in g^*. We also discuss the limitations in the interpretability of quantile based approaches in general when exploring treatment effect as well as limitations in the interpretability of our assumptions within our theory. Finally we briefly address the methodological limitations of our approach acknowledging this as a potential area for improvement.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

https://doi.org/10.52202/079017-3260

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The two main results of the paper, Proposition 1 and Theorem 2 are proved in sections C.1 and C.2. The assumptions for these results are given in Assumptions 1 & 2 with the implications of these assumptions discussed.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper gives detailed information about the experimental set-up and explicit algorithms for methods introduced. The specific use of algorithm 1 with NW estimation is also described Appendix A.2. In addition code is provided in the supplementary materials with 'SimulatedExperiments.ipynb' reproducing all of the experiments. We also provide a link to a public github repository containig this in a footnote on page 9.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with

the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is provided in the supplementary materials with each function documented as well as in a linked public github repository. Jupyter Notebooks are provided to reproduce all of the experiments and real world data settings used within the paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

 https://doi.org/10.52202/079017-3260

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Most of the experimental details are provided alongside the experimental results in Section 4 with additional details provided in Appendix A.4. In addition code to replicate all of the experiments is provided in the supplmentary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For our experimental results we explain that 500 replications of the entire fitting procedure (including newly generated data making each run entirely independent) was run and that 95% confidence intervals are provided for the mean absolute error of each estimator.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Information on runtime and compute resources is provided in Appendix A.4. Overall computational resources were very low with all experiments run on a personal laptop and each one taking less than an hour even with a large number of replications.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: No outsourcing was done for this project/paper and all datasets used are publicly available and completely anonymised. Additionally we feel there is essentially no scope for our work to cause negative societal impact.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the potential positive impact of our work in terms of further empowering HTE analysis. We do not for see any potential negative impacts of our work as it is simply furthering the field of HTE and quantile based treatment effect analysis. To our knowledge, there is no suggestion that these fields could be used maliciously or unfairly beyond the ways that any regression technique or analysis could.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

 https://doi.org/10.52202/079017-3260

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There are no new datasets released in our work. Our models use standard non-parametric regression techniques and adapt these to best estimate our estimand, the CQC meaning there is no real risk of misuse beyond the misuse of any regression technique.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets are publicly available and properly referenced. Proprietary code used is also open source and used in accordance with its licence. The licence for this code can also be found in code itself. The model we use is our own and so no licensing is required.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only asset is the code which is implemented in our method which is provided in the supplementary materials alongside a licence.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing or research with human subjects is used.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

https://doi.org/10.52202/079017-3260

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.