Robust and Agnostic Learning of Conditional Distributional Treatment Effects

Nathan Kallus **Miruna Oprescu**
Cornell University and Cornell Tech

Abstract

The conditional average treatment effect (CATE) is the best measure of individual causal effects given baseline covariates. However, the CATE only captures the (conditional) average, and can overlook risks and tail events, which are important to treatment choice. In aggregate analyses, this is usually addressed by measuring the distributional treatment effect (DTE), such as differences in quantiles or tail expectations between treatment groups. Hypothetically, one can similarly fit conditional quantile regressions in each treatment group and take their difference, but this would not be robust to misspecification or provide agnostic best-in-class predictions. We provide a new robust and model-agnostic methodology for learning the conditional DTE (CDTE) for a class of problems that includes conditional quantile treatment effects, conditional super-quantile treatment effects, and conditional treatment effects on coherent risk measures given by f-divergences. Our method is based on constructing a special pseudo-outcome and regressing it on covariates using any regression learner. Our method is model-agnostic in that it can provide the best projection of CDTE onto the regression model class. Our method is robust in that even if we learn these nuisances nonparametrically at very slow rates, we can still learn CDTEs at rates that depend on the class complexity and even conduct inferences on linear projections of CDTEs. We investigate the behavior of our proposal in simulations, as well as in a case study of 401(k) eligibility effects on wealth.

1 INTRODUCTION

Measuring treatment-effect heterogeneity along observed covariates is an important tool for interpreting the results of A/B tests on online platforms, program evaluations in social science whether experimental or observational, and clinical trials in medicine. These analyses can help diagnose how the treatment works, understand for whom it does and does not work, and assess fairness (Heckman et al., 1997; Crump et al., 2008; Kent et al., 2010; Kallus, 2022). On the other hand, measuring distributional treatment effects (DTEs) such as quantile treatment effects (QTEs) is an important tool for understanding the impact of interventions beyond the mean, especially when outcomes are naturally very skewed, like income or platform usage (Bitler et al., 2006; Firpo, 2007; Belloni et al., 2017; Kallus et al., 2019).

Motivated by the need to assess *both* heterogeneity *and* distributional impact, in this paper we study flexible, agnostic, and robust machine learning tools to estimate *conditional* DTE (CDTE) functions. Recent advances in causal machine learning have offered new methods for assessing effect heterogeneity by learning conditional *average* treatment effects (CATEs) (Imai and Ratkovic, 2013; Athey and Imbens, 2016; Wager and Athey, 2018; Künzel et al., 2019; Kennedy, 2020; Nie and Wager, 2021). These works have highlighted the importance of learning CATEs *directly*, rather than learning conditional-average outcomes by treatment arm and taking their difference (also known as the plug-in approach). One issue with the plug-in approach is that it can wash out the effect signal (not robust): *e.g.*, many variables strongly predict baselines but only a few modulate the effect. Another issue is that it fails to give best-in-class predictions (not agnostic): *e.g.*, taking the difference of the best linear predictions of outcome by arm does not yield the best linear prediction of treatment effect.

We tackle the same challenges for CDTEs. We consider CDTEs for a very rich class of distributional metrics that includes quantiles, super-quantiles, and other coherent risk measures. Given any distributional metric in our class, we construct a pseudo-outcome that combines an initial guess for the CDTE along with a debiasing term. Our algorithm is then to regress this pseudo-outcome on covariates, using any given blackbox learner. In the case of the CATE, our method recovers the DR-Learner (Kennedy, 2020). We show that this procedure is robust in the sense that the blackbox regression mimics having used the pseudo-outcome with the *true* CDTE as the "initial guess", thus removing any bias or noise

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

from fitting the baselines. We further show that our method is model-agnostic in the sense that, if the blackbox is not a universal approximator, we still get the best approximation to the CDTE function offered by the blackbox. Lastly, we show that our estimating procedure allows for valid statistical inference on the best linear projection of CDTE, thus enabling interpretable analyses of distributional effect heterogeneity. We demonstrate in a comprehensive simulation study that we obtain uniformly better performance than the plug-in approach for several types of CDTEs. Finally, we apply our method on a real world study of 401(k) eligibility and its impact on financial wealth.

2 PROBLEM SETUP

We consider either an experimental or observational dataset with two treatments, denoted by 0 and 1. Each unit in the dataset is a draw from a population of baseline covariates $X \in \mathcal{X}$, treatment indicator $A \in \{0,1\}$, and observed outcome $Y \in \mathbb{R}$. The dataset consists of n such independent draws, $Z_i = (X_i, A_i, Y_i) \sim Z = (X, A, Y)$, $i = 1, \ldots, n$. We define the propensity score as $e^*(X) = \mathbb{P}(A = 1 \mid X)$. We assume throughout that $e^*(X) \in (0,1)$ almost surely, known as overlap.

Each unit is additionally associated with two unobserved potential outcomes, $Y(0)$, $Y(1) \in \mathbb{R}$, representing the potential outcome we would observe if (possibly counter to fact) each treatment were applied. We assume we observe the potential outcome corresponding to the treatment indicator, $Y = Y(A)$, which also encapsulates an assumption of no interference between unit treatments. We assume unconfoundedness (ignorability) throughout: $Y(a) \perp\!\!\!\perp A \mid X$. For experimental data this is ensured by design via random assignment of A (often with covariate-agnostic assignment, $A \perp\!\!\!\perp X$). For observational data, this is an assumption that all potential sources of confounding are captured in X. For our purposes, the only difference between the two cases is whether the propensity score $e^*(X)$ is known.

We are interested in the differences between the conditional distributions of $Y(1)$ and $Y(0)$, given X. In the next section, we describe specific metrics for these differences.

Notation. Given a distribution F, we define $\mathbb{E}_F[f(Z)] = \int f(z)dF(z)$. We let $\widehat{\mathbb{E}}_n$ denote the empirical expectation $\widehat{\mathbb{E}}_n f(Z) = \frac{1}{n}\sum_{i=1}^n f(Z_i)$. For a parameter f, we reserve f^* to represent its true value and $\widehat{f}$ a value learned from the data. We let $\|f\| := \mathbb{E}_F[f(z)^2]^{1/2}$ be the L_2 norm of f. We use D to denote directional derivatives: $D_h F(h)|_{h=h'} = \frac{\partial}{\partial a} F(h' + a)|_{a=0}$, whenever this exists. If h is a vector of functions, $D_h F(h) = (D_{h_1} F(h), ..., D_{h_k} F(h))$ and $D_{h_i} F(h)|_{h=h'} = \frac{\partial}{\partial a} F((h'_1, ..., h'_i + a, ..., h'_k))|_{a=0}$. We let $\text{conv}(\mathcal{S})$ denote the convex hull of $\mathcal{S}$. For two numbers a, b, we take $a \lesssim b$ to mean $a \leq Cb$ for some universal constant C and $a \asymp b$ to mean $cb \leq a \leq Cb$ for some constants c and C. Finally, we let $\overline{1,n}$ denote the set of integers $\{1, ..., n\}$.

3 CONDITIONAL DISTRIBUTIONAL TREATMENT EFFECTS

CDTEs are functions mapping x to a difference in some statistic of the conditional distributions of $Y(1)$ and $Y(0)$, given $X = x$. Examples of such statistics are the mean, yielding the CATE, and the τ-quantile, yielding the CQTE. In this paper, we handle a very wide range of CDTEs given by statistics defined by *moment equations* (Chamberlain, 1992; Ai and Chen, 2003).

Definition 1 (Moment Statistics)**.** Given $\rho : \mathbb{R}^{m+2} \to \mathbb{R}^{m+1}$, we define a statistic of a distribution F on $\mathbb{R}$ as $\kappa^*(F)$ for $(\kappa^*(F), h^*(F)) \in \mathbb{R} \times \mathbb{R}^m$ any solution (if it exists) to the moment equation

$$\mathbb{E}_F[\rho(Y, \kappa, h)] = 0. \tag{1}$$

Definition 2 (CDTEs)**.** Let $F_{Y(1)|X}$ and $F_{Y(0)|X}$ denote the conditional distributions of $Y(1)$ and $Y(0)$ given X, respectively. Fix a statistic $\kappa^*(F)$ given by Definition 1. The corresponding CDTE is given by:

$$\text{CDTE}(X) = \kappa^*(F_{Y(1)|X}) - \kappa^*(F_{Y(0)|X}). \tag{2}$$

For brevity, we will define the following functions (scalar-valued and $\mathbb{R}^m$-valued, respectively)

$$\kappa_a^*(X) = \kappa^*(F_{Y(a)|X}), \;\; h_a^*(X) = h^*(F_{Y(a)|X}), \;\; a = 0, 1.$$

We also assume these functions exist in that at least one solution to Eq. (1) exists for $F_{Y(a)|X}$ (see Remark 3 regarding multiplicity). In the examples below, we give specific names for κ or h (*e.g.*, $q(F;\tau)$ for the τ-quantile of F), in which case we use analogous abbreviations (*e.g.*, $q_a(X;\tau)$).

Note that under unconfoundedness and overlap, $F_{Y(a)|X}$ is the same as $F_{Y|X,A=a}$, the conditional distribution of Y given X and $A = a$. Therefore, κ_a^*, h_a^*, and the CDTE are all identifiable from the data. That is, despite being defined in terms of potential outcomes, the CDTE depends only on the distribution of observed data (X, A, Y) and not on that of the unobserved data $(X, Y(0), Y(1))$, under our assumptions. The question we address in this paper is *how* to learn CDTEs from data.

We next review some important examples of CDTEs that fit into our framework. First note that CATE fits into this setup by setting $\rho(y, \kappa) = y - \kappa$ with $m = 0$ in Eq. (1) (no h's). The power of our framework is that it captures many CDTEs of interest beyond averages.

3.1 Example 1: Conditional Quantile Treatment Effects

Our first example is the **conditional quantile treatment effect** (CQTE). Given $\tau \in (0,1)$, the quantile at level τ of the distribution F is defined as $q(F;\tau) = \inf\{y : F(y) \geq$

$\tau\}$ (identifying F with its cumulative distribution function). Whenever F has a positive derivative at $q(F;\tau)$, it is given by Definition 1 with $m = 0$ (no h's) and

$$\rho(y, q) = \tau - \mathbb{I}[y \le q]. \tag{3}$$

The corresponding CDTE is called the CQTE at level τ.

Non-conditional QTEs are an important tool for quantifying the effects of treatments throughout the outcome distribution (Firpo, 2007; Belloni et al., 2017; Kallus et al., 2019). This is especially important when we suspect that the outcome distribution might be skewed or heavy-tailed (*e.g.*, income).

CQTEs offer an opportunity to assess such effects at the individual level. In particular, the CQTE can capture the potential increase in an individual's chance to have very poor outcomes due to treatment, even if the average effects are good or neutral. That is, if $A = 1$ denotes an intervention, then CQTEs answer the prediction question: given an individual's covariates, how would intervening affect the 10%-worst possible outcomes.

A related quantity is the difference between the τ and $1-\tau$ conditional quantiles across treatments: $\omega(X;\tau) = q(F_{Y(1)|X};\tau) - q(F_{Y(0)|X};1-\tau)$. This quantity can bound the (unobservable) individual treatment effect:

$$\mathbb{P}\left(\omega(X;\tau) \le Y(1) - Y(0) \le \omega(X;1-\tau) \mid X\right) \ge 1 - 4\tau.$$

For the sake of brevity and uniform treatment we focus on CDTEs (*i.e.*, using the same statistic for both potential outcome distributions), but our results readily extend to quantities like this as well.

3.2 Example 2: Conditional Super-Quantile Treatment Effects

Our second example is the **conditional super-quantile treatment effect** (CSQTE). Given $\tau \in (0,1)$, the super-quantile (also known as conditional value-at-risk or tail expectation) at level τ of a distribution F is defined as

$$\begin{aligned}\mu(F;\tau) &= \inf_\beta \beta + \tfrac{1}{1-\tau}\textstyle\int_\beta^\infty (1 - F(y))\, dy \\ &= \inf_\beta \mathbb{E}_F[\beta + (1-\tau)^{-1}\max\{y-\beta, 0\}].\end{aligned} \tag{4}$$

The corresponding CDTE is called the CSQTE at level τ.

Note that a β realizing the above infimization is the quantile at level τ, $q(F;\tau)$. Therefore, given positive density at the quantile, the super-quantile is given in Definition 1 by setting $m = 1$ and

$$\rho(y, \mu, q) = ((1-\tau)^{-1} y \mathbb{I}[y \ge q] - \mu,\ \tau - \mathbb{I}[y \le q]). \tag{5}$$

The super-quantile $\mu(F;\tau)$ is the largest-possible subpopulation average among all subpopulations comprising a $1-\tau$ fraction of the population of values described by F. When $F(q(F;\tau)) = \tau$, this can be phrased as the average of values above the τ quantile. If we are interested in the left tail (average below a quantile), we can simply consider the the negative CSQTE in the negative outcomes. Unlike the quantile, the super-quantile is a coherent risk measure (Artzner et al., 1999).

In particular, while the CQTE captures the impact on the outcomes at a single probability level, they can fail to provide a full picture of the risk profile as they are indifferent to anything beyond the threshold of the quantile. In contrast, the CSQTE captures effects beyond quantile breakpoint and quantifies the impact on the *average*, say, 10%-worst (or, best) outcomes. Crucially, since super-quantiles are a coherent risk measure, making individual decisions based on CSQTEs (*e.g.*, intervene when the CSQTE is positive) is rational with respect to the coherent-risk axioms.

3.3 Example 3: Conditional f-Risk Treatment Effects

Our third example is a whole class of coherent risk measures. Coherent risk measures quantify how bad/good a random loss/reward is while satisfying certain axioms (monotonicity, sub-additivity, homogeneity, and translational invariance). A key result is that coherent risk measures are equivalent to distributionally robust optimization (Ruszczyński and Shapiro, 2006). In this section, we focus on the **conditional f-risk treatment effect** (CfRTE), a family of coherent risk measures generated by f-divergences.

Given a convex $f : \mathbb{R} \to \mathbb{R}$ with $f(1) = 0$, we define the conditional f-risk at level $\delta \ge 0$ as:

$$R^f(F;\delta) = \sup_{G \ll F,\, D^f(G\|F) \le \delta} \mathbb{E}_G[Y],$$

where $D^f(G\|F) = \mathbb{E}_F[f(dG/dF)]$. For example, the f-divergence for $f(x) = x\log x$ is the Kullback–Leibler divergence, and the f-risk is known as entropic value-at-risk (EVaR) which upper bounds the super-quantile at level $1 - e^{-\delta}$ (Ahmadi-Javid, 2012). The f-risk admits a dual formulation given by the following convex optimization problem (Rockafellar, 1974):

$$\begin{aligned}R^f(F;\delta) &= \inf_{\beta\ge 0, \lambda\in\mathbb{R}} \mathbb{E}_F\left[m(Y,\beta,\lambda;\delta)\right], \\ m(y,\beta,\lambda;\delta) &= \delta\beta + \lambda + \beta f^*\left(\beta^{-1}(y-\lambda)\right),\end{aligned}$$

where $f^*(x^*) = \sup_{x\in\mathbb{R}}\ xx^* - f(x)$ is the convex conjugate of f.

Therefore, under appropriate regularity, $R^f(F;\delta)$ is given by Definition 1 with

$$\begin{aligned}\rho(y, R, \beta, \lambda) = \big(&m(y,\beta,\lambda;\delta) - R^f,\ \tfrac{\partial}{\partial\beta} m(Z,\beta,\lambda;\delta), \\ &\tfrac{\partial}{\partial\lambda} m(Z,\beta,\lambda;\delta)\big).\end{aligned} \tag{6}$$

The corresponding CDTE is called the CfRTE at level δ. Frequently employed in actuarial science and finance, coherent risk measures are a fundamental tool for assessing

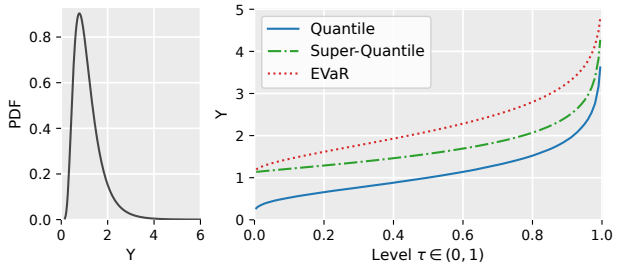


Figure 1: Comparison of quantiles, super-quantiles, and EVaRs for a right-truncated ($Y \leq 6$) Lognormal($\mu = 0$, $\sigma = 0.5$) at different risk levels $\tau \in (0, 1)$. *Note:* Level τ corresponds to $\delta = -\log(1 - \tau)$ for EVaR.

risk. CfRTE can therefore be used to perform risk-benefit analyses on the treatment effect profiles of affected groups.

Remark 1. A natural question is which risk measure to use in practice. Ultimately, the appropriate risk measure (and level) will be application-dependent and it is up to the practitioner to select a measure that best reflects the desired risk profile. For illustration purposes, in Figure 1 we show how our three examples (quantiles, super-quantiles and EVaRs) compare for a heavy-tailed distribution.

4 RELATED LITERATURE

Learning CATEs. CATE estimation is a central problem in causal learning and finds uses in both the analysis of causal interventions and decision support for personalization. Going beyond fully-parametric models, early advances relied on semiparametric models that imposed structure on the CATE function (Robins et al., 1992; Van der Laan and Robins, 2003; Vansteelandt and Joffe, 2014). Recently there has been a surge of interest in leveraging machine learning for CATE estimation. These flexible methods either employ specific machine learning models such as Bayesian regression trees (Hill, 2011; Hahn et al., 2020), random forests (RFs) (Wager and Athey, 2018; Oprescu et al., 2019), neural networks (Johansson et al., 2016; Atan et al., 2018; Shi et al., 2019), or allow for arbitrary blackbox meta-learners (Künzel et al., 2019; Nie and Wager, 2021) by leveraging efficient influence functions (Robins et al., 2017; Kennedy, 2020; Curth et al., 2020) and Neyman orthogonality (Chernozhukov et al., 2018; Foster and Syrgkanis, 2019). This paper is closest to the works on efficient influence functions and blackbox meta-learners, and we add to this literature by considering distributional statistics beyond averages.

Double machine learning and orthogonal statistical learning. Another vein of related literature is learning with nuisances. Both our work and the above methods based on regressing efficient influence functions may be phrased within the wider framework of orthogonal statistical learning (Foster and Syrgkanis, 2019), which extends double machine learning (DML) (Chernozhukov et al., 2018) from estimation to minimizing loss functions involving unknown nuisances subject to Neyman orthogonality (Neyman, 1959). Neyman orthogonal losses arise naturally from efficient influence functions (Ichimura and Newey, 2022), a fact that has been leveraged by Foster and Syrgkanis (2019); Kennedy (2020); Curth et al. (2020); Athey and Wager (2021) to tackle both CATE and policy learning. These methods have learning rates that adapt to the complexity of the target rather than that of the nuisances, but they largely focus on conditional averages. Our work builds on this line of research and is most similar to (Kennedy, 2020) in that we propose nuisance agnostic CDTE estimators with guarantees beyond those given by Neyman orthogonality.

Distributional treatment effects. The literature on DTEs can be split into two categories: (i) estimating cumulative distribution functions of potential outcomes and (ii) directly estimating distributional parameters of interest. In the first category, the main approach is to model conditional counterfactual distributions using distribution regression (Chernozhukov et al., 2013, 2020) or fully flexible approaches such as neural networks (Ge et al., 2020; Zhou and Carlson, 2021) and mean kernel embeddings (Park et al., 2021). Since these methods rely on plug-in estimation, they can be slow and biased when concerned with a particular DTE. The second category focuses on estimating a particular DTE. Existing orthogonal/efficient methods focus on unconditional DTEs (Firpo, 2007; Belloni et al., 2017; Kallus et al., 2019). Existing methods that tackle CDTEs rely on plug-in estimation (Park et al., 2021) or parametric methods (Hohberg et al., 2020). Our work bridges the gap by proposing robust, agnostic, and flexible CDTE learning.

5 PSEUDO-OUTCOME REGRESSION FOR CDTEs

In this section, we propose an algorithm for learning CDTEs from data. As a first step, consider the following naive, but straightforward estimation procedure:

$$\text{CDTE}^{\text{Plugin}}(X) = \widehat{\kappa}_1(X) - \widehat{\kappa}_0(X). \tag{7}$$

where $\widehat{\kappa}_a(\cdot)$ are estimates for $\kappa_a(\cdot)$ (see Section 5.2 regarding how these may be constructed). This estimator is known as a "plug-in" estimator or, in the context of CATE learning, a "T-learner" (Künzel et al., 2019). Unfortunately, this approach has several drawbacks. One concern is that the treatment effect signal can be easily masked by noise in the baseline predictors. In particular, while many variables may strongly predict baseline response, only a few strongly modulate effect. The effect function may often be simpler, sparser, and/or smoother than each baseline function. Therefore, the plug-in estimator can suffer from excessive bias inherent in fitting baseline estimators in high-dimensions or using flexible models. If, on the other hand, we seek to use a simple model such as a linear fit, we will find that differencing the best linear predictors of baseline outcomes does not yield the best linear predictor of effect. For these reasons,

it is imperative to learn CDTEs in a direct, model-agnostic, and robust way.

To address the short-comings of the plug-in estimator, we consider the plug-in prediction on each data point $\mathrm{CDTE}^{\text{Plugin}}(X_i)$, debias it using the observed action and outcome A_i, Y_i, and finally regress the debiased prediction on X again. Our first task is to propose an appropriate debiased pseudo-outcome for CDTE learning.

Definition 3 (CDTE Pseudo-Outcome)**.** Fix a statistic in Definition 1. Let $\nu_a^* = (\kappa_a^*, h_a^*)$ and

$$\alpha_a^*(X) = (J_a^*(X))_1^{-1},$$
$$\text{where } J_a^*(X) = D_{\nu_a}\{\mathbb{E}[\rho(Y, \nu_a) \mid X, A = a]\}\big|_{\nu_a = \nu_a^*}$$

provided that $(J_a^*(X))^{-1}$ exists. Here, $(J_a^*(X))_1^{-1}$ denotes the first row of the inverse Jacobian.

Given some e, α, ν serving as stand-ins for e^*, α^*, ν^*, we define the CDTE pseudo-outcome by

$$\begin{aligned}\psi(Z, e, \alpha, \nu) =& \kappa_1(X) - \kappa_0(X) \\ &- \tfrac{A - e(X)}{e(X)(1-e(X))}\alpha_A(X)^T \rho(Y, \nu_A(X)).\end{aligned} \tag{8}$$

We refer to e^*, α^*, ν^* as *nuisance functions*, as they are unknown functions needed to construct our pseudo-outcome. One initial motivation for Eq. (8) is that, by iterated expectations, we have $\mathbb{E}[\psi(Z, e^*, \alpha^*, \nu^*) \mid X] = \mathrm{CDTE}(X)$. That is, if we had plugged in the true nuisances, then the regression of our pseudo-outcome on X yields exactly what we want. This is, however, not so surprising because if we plug in $\nu = \nu^*$, the last term in Eq. (8) just has zero conditional expectation, given X, so we are left with $\kappa_1^*(X) - \kappa_0^*(X)$. The reason why Eq. (8) is special is that, as we will show, if we make small errors in the nuisances, the impact on the conditional expectation of our pseudo-outcome is even smaller, leading to robustness guarantees. This is in contrast to the plug-in approach, where errors in $\widehat{\kappa}_a(X)$ propagate directly to $\mathrm{CDTE}^{\text{Plugin}}$, which is just their difference.

The origin of Eq. (8) is that $\psi(Z, e^*, \alpha^*, \nu^*)$ is in fact the efficient influence function for the estimand $\mathbb{E}[\mathrm{CDTE}(X)]$ (for a review of influence functions see Kennedy (2022); Ichimura and Newey (2022)). In particular, if our statistic is simply the mean ($\rho(y, \kappa) = y - \kappa$) then $\psi(Z, e^*, \alpha^*, \nu^*)$ reduces to the familiar doubly-robust influence function, $\mathbb{E}[Y \mid X, A = 1] - \mathbb{E}[Y \mid X, A = 0] + \frac{A - e^*(X)}{e^*(X)(1-e^*(X))}(Y - \mathbb{E}[Y \mid X, A])$ that produces the CATE when regressed on X (as studied by Kennedy (2020)). As we never use the fact that $\psi(Z, e^*, \alpha^*, \nu^*)$ is the efficient influence function, we do not prove this or impose the necessary regularity conditions for this to actually hold precisely. Instead, we simply use a perturbation argument to essentially guess the form of Eq. (8), given which we directly prove our robustness and inference guarantees.

Algorithm 1 CDTE Learner

Input: Data $\{(X_i, A_i, Y_i) : i \in \overline{1, n}\}$, folds $K \geq 2$, nuisance estimators, regression learner

for $k \in \overline{1, K}$ **do**
  Use data $\{(X_i, A_i, Y_i) : i \neq k - 1 \pmod K\}$ to
  construct nuisance estimates $\hat{e}^{(k)}, \hat{\alpha}^{(k)}, \hat{\nu}^{(k)}$
  for $i = k - 1 \pmod K$ **do**
    Set $\widehat{\psi}_i = \psi(Z_i, \hat{e}^{(k)}, \hat{\alpha}^{(k)}, \hat{\nu}^{(k)})$
  end for
end for
return $\widehat{\mathrm{CDTE}}(x) = \widehat{\mathbb{E}}_n[\widehat{\psi} \mid X = x]$

5.1 The CDTE Learning Algorithm

We now describe our algorithm, which is summarized in Algorithm 1. We first split the data into K even folds. We then construct a pseudo-outcome for each data point by plugging in estimates of the nuisances into Eq. (8). The nuisances at a point are fit on data excluding the fold that the data point belongs to. This ensures that the data point and the nuisance estimates are independent without splitting the data into two and instead only using parts of it for each task. Finally, we regress the pseudo-outcome on X using a given regressor. We use $\widehat{\mathbb{E}}_n[W \mid X = x]$ to denote the function learned by the given regression method when regressing W on X given n data points (X_i, W_i), $i \in \overline{1, n}$. This notation affords us significant generality. For example, the regression method may be to minimize the sum of squared errors over some function class (*e.g.*, linear or neural nets) or it may be given by local polynomial regression or RFs.

5.2 Nuisance Estimation

Algorithm 1 requires nuisance estimators as inputs. Exactly how these nuisances are estimated may depend on the particular scenario. Let us first discuss the propensity $e^*(x)$. If it is known, as in experimental settings, we may simply set it as our estimate. Otherwise, we can estimate it using probabilistic classification (*e.g.*, logistic regression or neural nets with softmax output).

Next, we discuss $\nu_a^*(x)$. Most generally, since it is defined by solving the conditional moment restriction $\mathbb{E}[\rho(Y, \nu_a^*(X)) \mid X] = 0$, this nuisance may be learned by employing methods made for solving such models (Ai and Chen, 2003; Chen and Pouzo, 2009; Bennett et al., 2019; Bennett and Kallus, 2020; Athey et al., 2019; Khosravi et al., 2019; Dikkala et al., 2020). However, in some specific examples, more direct methods may be applicable. For both CQTE and CSQTE, $\nu_a^*(x)$ includes a conditional quantile function, which can be learned using any quantile regression method, whether minimizing the check loss (Koenker and Bassett Jr, 1978) or using forests (Meinshausen and Ridgeway, 2006). For CSQTE, we additionally need to fit the conditional super-quantile. Per Eq. (5), that nuisance

is given by the regression $\mathbb{E}[(1-\tau)^{-1}Y\mathbb{I}[Y \geq q_a(X;\tau)] \mid X, A=a]$. Therefore, one possibility is to split the training data (being all data excluding the k^{th} fold) into two halves, fit a conditional quantile estimate $\hat{q}_a^{(k)}(x;\tau)$ on one, set $\omega_i = (1-\tau)^{-1}Y_i\mathbb{I}[Y_i \geq \hat{q}_a^{(k)}(X_i;\tau)]$ on the other, and return $\hat{\mu}_a^{(k)}(x;\tau) = \widehat{\mathbb{E}}_{(1-1/k)n/2}[\omega \mid X=x, A=a]$. In particular, a given X-regression method can simply be applied once to $A=0$ and once to $A=1$. Appendix A of Dorn et al. (2021) provides guarantees for this procedure when the regression learner minimizes squared error over a class with bracketing entropy and Olma (2021) when using local linear regression.

Lastly, we discuss $\alpha_a^*(x)$. In some cases, it is a known function that need not be estimated. For CfRTE, it is equal to $(-1,0,0)$. In other cases, it is given directly by other nuisances. For CSQTE, it is equal to $(-1,(1-\tau)^{-1}q_a^*(x;\tau))$ so we can simply re-use the estimate we constructed for ν_a^*. In yet other cases, it is another nuisance that must be estimated. For CQTE, it is equal to $1/f_{Y|X=x,A=a}(q_a^*(x;\tau))$, the reciprocal of the density of $Y \mid X=x, A=a$ at the conditional quantile. One way to fit this suggested by Leqi and Kennedy (2021) is to split the training data into two halves, fit a conditional quantile estimate $\hat{q}_a^{(k)}(x;\tau)$ on one, set $\omega_i = K((Y_i - q_a^{(k)}(X_i;\tau))/b_n)/b_n$ on the other, where $K(u)$ is a kernel function such as the standard normal density, and return $\hat{\alpha}_a^{(k)}(x) = \widehat{\mathbb{E}}_{(1-1/k)n/2}[\omega \mid X=x, A=a]$.

6 GUARANTEES FOR LEARNING

In this section, we study the finite sample error rates for Algorithm 1 with arbitrary first- and second-stage estimators. For this and the next section, let us fix some CDTE with pseudo-outcome as in Definition 3 and estimation procedures input to Algorithm 1. We require the following boundedness conditions.

Assumption 1 (Boundedness)**.** For a nuisance realization set Ξ, there exist $c_1 > 0, c_2 \geq 0, c_3 \geq 0, c_4 > 0, c_5 \geq 0$ and matrices $G, H \in \{0,1\}^{(m+1)\times(m+1)}$ such that $\forall (e,\alpha,\nu_a) \in \Xi,\ i,j,l \in \overline{1,m+1},\ \bar{\nu}_a \in \text{conv}\{(\nu_a^*, \nu_a)\}$,

- $e^*(X), e(X) \in [c_1,\ 1-c_1]$
- $\left|D_{\nu_{a,j}}\mathbb{E}[\rho_i(Y,\nu_a) \mid X, A=a|_{\nu_a=\bar{\nu}_a}\right| \leq c_2 G_{ij}$
- $\left|D_{\nu_{a,l}}D_{\nu_{a,j}}\mathbb{E}[\rho_i(Y,\nu_a)|X=x, A=a]|_{\nu_a=\bar{\nu}_a}\right| \leq c_3 H_{jl}$
- $\det(D_{\nu_a}\{\mathbb{E}[\rho(Y,\nu_a) \mid X=x, A=a]\}\big|_{\nu_a=\bar{\nu}_a}) > c_4$
- $|\rho_i(Y,\bar{\nu}_a)| \leq c_5$

The first condition in Assumption 1 ensures that both treatments and controls can be observed for any X with some fixed probability. This is guaranteed in a randomized trial if $e^*(X)$ is a constant and otherwise is a standard assumption in observational studies. The other conditions encode the cross-term structure of the derivatives of our moments. In most examples, the requirement of ρ being bounded amounts to requiring Y to be bounded. Similar boundedness assumptions are often made in debiased machine learning for ATE and CATE to control remainder terms.

We can then prove the following *conditional* Neyman orthogonality for our pseudo-outcomes, which is the key step for learning guarantees.

Theorem 1 (Conditional Neyman Orthogonality)**.** *Suppose Assumption 1 holds and let* $(e,\alpha,\nu_a) \in \Xi$*. Then,*

$$\begin{aligned}
&\|\mathbb{E}\left[\psi(Z,e,\alpha,\nu) - \psi(Z,e^*,\alpha^*,\nu^*) \mid X\right]\| \lesssim \mathcal{E}(e,\alpha,\nu),\\
&\mathcal{E}(e,\alpha,\nu) = \textstyle\sum_{a=0}^{1}\big(\|\kappa_a - \kappa_a^*\|\,\|e-e^*\| \\
&\quad + \textstyle\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} G_{ij}\|\alpha_{a,i} - \alpha_{a,i}^*\|\,\|\nu_{a,j} - \nu_{a,j}^*\| \\
&\quad + \textstyle\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} H_{ij}\|\nu_{a,i} - \nu_{a,i}^*\|\,\|\nu_{a,j} - \nu_{a,j}^*\|\big),
\end{aligned} \tag{9}$$

where the $\lesssim$ *hides dependence on* c_1, c_2, c_3, c_4.

The result shows that whether we use the pseudo-outcome with oracle nuisances or estimated nuisances as a regression target, the difference is bounded by a *quadratic* form in the nuisance errors, wherein G, H govern which pairwise error products appear. Thus, even if nuisances are estimated slowly, the impact is marginal, as the error is squared.

We will show that, up to $\mathcal{E} = \sum_{k=1}^{K}\mathcal{E}(\hat{e}^{(k)}, \hat{\alpha}^{(k)}, \hat{\nu}^{(k)})$, our estimate $\widehat{\text{CDTE}}$ produced by Algorithm 1 behaves the same as if we applied our regression method to the pseudo-outcome with oracle nuisances, $\widehat{\text{CDTE}}(x) = \widehat{E}_n[\psi(Z,e^*,\alpha^*,\nu^*) \mid X=x]$. In particular, if $\mathcal{E}$ is generally smaller than the error $\|\widehat{\text{CDTE}} - \text{CDTE}\,\|$, then the leading behavior of $\widehat{\text{CDTE}}$ will be equivalent to that of $\widetilde{\text{CDTE}}$. The particular statement of this will depend on what last-stage regression method we are using.

The significance is that Algorithm 1 will behave like regressing an unbiased observation of CDTE on X, even though we used estimated nuisances. First, this means we can learn CDTE at rates that match the complexity of that function. Second, this implies that we can directly approximate CDTE and get model-agnostic best-in-class guarantees. For example, if we use linear regression, we would get the *best* linear approximation for CDTE at a rate of $n^{-1/2}$. This is especially important if we seek an interpretable model. In the next section, we show this also enables *inference*.

Remark 2. As shown in Appendix B, in many examples, Eq. (9) will contain only a few of the product terms, because many G, H entries are 0 and/or $\widehat{\alpha}(x,\widehat{\nu}) = \alpha^*(x,\widehat{\nu})$. This enables us to trade off slower rates in one nuisance for faster rates in another while maintaining the same rate for $\mathcal{E}$.

Remark 3. We never explicitly assume that the conditional moment restrictions identify $\nu_a^*(x)$ uniquely, only that they exist. While uniqueness is not necessary, the right-hand side of Eq. (9) will not vanish unless $\widehat{\nu}_a^{(k)}(x)$ converges to a *single*, *non-random* limit point $\nu_a^*(x)$. This is certainly

possible even under solution multiplicity (Imbens et al., 2021; Kallus and Mao, 2022). At the same time, usually $\nu_a^*(x)$ will be unique under very mild assumptions such as having continuous distributions.

First, we consider an empirical risk minimization algorithm (nonparametric least squares): given a class $\mathcal{F} \subset [\mathcal{X} \to \mathbb{R}]$,

$$\widehat{\mathbb{E}}_n[W \mid X = \cdot] \in \operatorname{argmin}_{f \in \mathcal{F}} \widehat{\mathbb{E}}_n (W - f(X))^2. \quad (10)$$

Theorem 2. *Suppose Assumption 1 holds and almost surely* $(\hat{e}^{(k)}, \hat{\alpha}^{(k)}, \hat{\nu}^{(k)}) \in \Xi$ *for* $k \in \overline{1, K}$. *Let* $\widehat{\mathbb{E}}_n[\cdot \mid X = x]$ *be as in Eq.* (10) *and suppose* $\mathcal{F}$ *is convex, closed and has bracketing entropy* $\log N_{[]}(\mathcal{F}, \epsilon) \lesssim \epsilon^{-r}$ *with* $0 < r < 2$ *and that* $|f(x)| \leq c_5 \, \forall f \in \mathcal{F}, x \in \mathcal{X}$. *Then,*

$$\|\widehat{\mathrm{CDTE}} - \mathrm{CDTE}\| \lesssim O_p(n^{-1/(2+r)}) + \mathcal{E}. \quad (11)$$

The rate $O_p(n^{-1/(2+r)})$ is generally the rate for regressing a *known* target using nonparametric least squares over $\mathcal{F}$ with such bracketing entropy. For example, if $\mathcal{F}$ is Hölder functions of smoothness β in d-dimensional inputs then it satisfies the entropy condition with $r = d/\beta$ (van der Vaart and Wellner, 1996, cor. 2.7.2) and $n^{-\beta/(2\beta+d)}$ is the optimal rate for regression in such a class (Stone, 1982). Thus, if $\mathcal{E} = O_p(n^{-1/(2+r)})$, our regression behaves as though we supplied it with the oracle nuisances.

Obtaining such a rate for $\mathcal{E}$ is generally lax because it is *quadratic* in nuisance-estimation errors. For example, if nuisances are estimated at the much slower rate $O_p(n^{-1/(4+2r)})$, the condition is ensured. The special structure in Theorem 1 further permits some trade-off between the rates of different nuisances. In Appendix B, we give the pseudo-outcome for each of the examples in Section 3 and instantiate Theorem 1 to exactly characterize the trade-offs.

Going beyond empirical risk minimizers, leveraging Theorem 1 and invoking a result of Kennedy (2020) [1] we can characterize the behavior when using a last-stage regression method satisfying certain stability properties.

Theorem 3. *Suppose Assumption 1 holds and almost surely for* $(\hat{e}^{(k)}, \hat{\alpha}^{(k)}, \hat{\nu}^{(k)}) \in \Xi$ *for* $k \in \overline{1, K}$, *and that, for any targets* Y *and* W, $\widehat{\mathbb{E}}_n[Y \mid X = x] + c = \widehat{\mathbb{E}}_n[Y + c \mid X = x]$ *and* $\|\widehat{\mathbb{E}}_n[W \mid X] - \mathbb{E}[W \mid X]\| \asymp \|\widehat{\mathbb{E}}_n[Y \mid X] - \mathbb{E}[Y \mid X]\|$ *whenever* $\mathbb{E}[Y \mid X = x] = \mathbb{E}[W \mid X = x]$. *Then,*

$$\|\widehat{\mathrm{CDTE}} - \mathrm{CDTE}\| \lesssim \|\widetilde{\mathrm{CDTE}} - \mathrm{CDTE}\| + \mathcal{E}. \quad (12)$$

7 GUARANTEES FOR INFERENCE

We next study how Algorithm 1 can be used for inference on the best linear projection of CDTE:

$$\gamma^* \in \operatorname{argmin}_{\gamma \in \mathbb{R}^p} \| \mathrm{CDTE} - \gamma^\top \phi(\cdot) \|,$$

for a given $\phi : \mathcal{X} \to \mathbb{R}^p$. In particular, ϕ may be subset just some of the features. Linear projections offer an interpretable view into distributional-effect heterogeneity. In the previous section, we showed Algorithm 1 can perform well at learning linear projections. Next, we show that we can further conduct inference on the coefficients, which can facilitate interpretation and credible conclusions.

Theorem 4 (Asymptotic Normality of Linear Projections)**.** *Suppose Assumption 1 holds and almost surely* $(\hat{e}^{(k)}, \hat{\alpha}^{(k)}, \hat{\nu}^{(k)}) \in \Xi$ *for* $k \in \overline{1, K}$. *Let* $\widehat{\gamma}$ *be the coefficient vector returned by Algorithm 1 when using ordinary least squares (OLS) on* $\phi(X)$ *as the regression blackbox for the final stage. Furthermore, assume* $\|\widehat{e}^{(k)} - e^*\| = o_p(1)$, $\|\widehat{\kappa}_a^{(k)} - \kappa_a^*\| \|\widehat{e}^{(k)} - e^*\| = o_p(n^{-1/2})$, $\|\widehat{\alpha}_{a,i}^{(k)} - \alpha_{a,i}^*\| = o_p(n^{-1/4})$, $\|\widehat{\nu}_{a,i}^{(k)} - \nu_{a,i}^*\| = o_p(n^{-1/4})$, $\forall i \in \overline{1, m+1}$, $k \in \overline{1, K}$. *Then,* $\widehat{\gamma}$ *satisfies*

$$\sqrt{n}(\widehat{\gamma} - \gamma^*) \rightsquigarrow \mathcal{N}(0, \Sigma^*), \quad (13)$$

where Σ^* *is the asymptotic covariance matrix for the linear regression of* $\psi(Z, e^*, \alpha^*, \nu^*)$ *on* $\phi(X)$.

Theorem 4 implies that we can just use out-of-the-box OLS with the built-in OLS inference as the final stage regression in Algorithm 1. The inference results would remain valid, provided nuisances are estimated slowly but not too slowly. In particular, we should generally use robust (aka sandwich or Huber–White) standard errors, as we do not expect the projection to have homoskedastic errors.

8 EMPIRICAL RESULTS

In this section, we demonstrate our method and theoretical results by applying Algorithm 1 to learn CSQTEs (Section 3.2). We first benchmark its performance in simulated data and then illustrate its use in a study of 401(k) eligibility and its effect on wealth accumulation. We provide additional results for CQTEs and CfRTEs in Appendix C. Replication code is available at `https://github.com/CausalML/CDTE`.

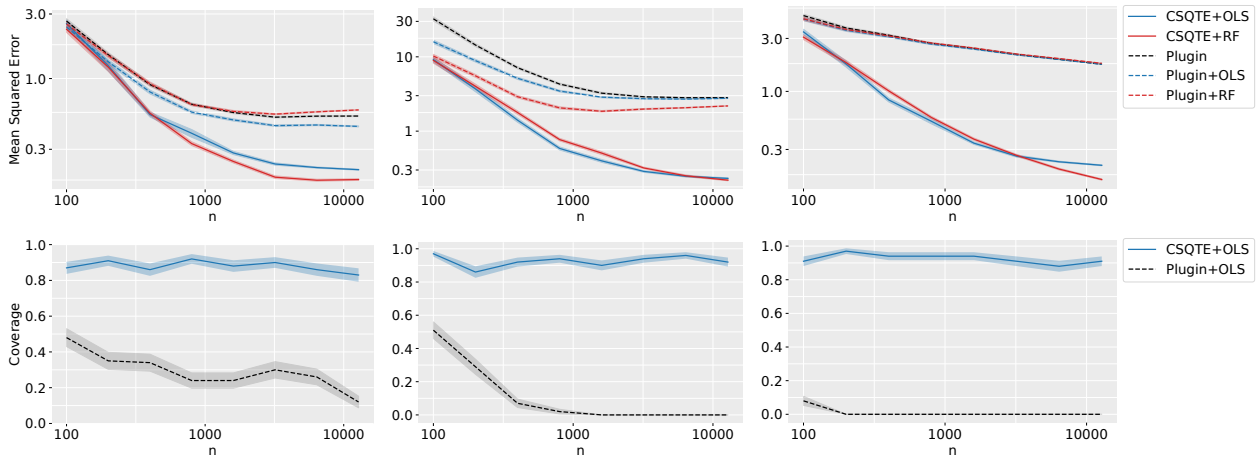
8.1 Simulation Study

We sample from the following data generating process:

$$X \sim \mathrm{Unif}([0,1]^{10}), \quad A \sim \mathrm{Bernoulli}(\sigma(6X_0 - 3)),$$
$$Y \mid X, A \sim \mathrm{Lognormal}(X_0 + AX_1, 0.2),$$

where σ is the logistic sigmoid. The coefficients in the expression for A are chosen such that the true propensity lies in the range $[0.05, 0.95]$. We seek to measure CSQTE at level $\tau = 0.75$, *i.e.*, the conditional average of values above the 0.75 conditional quantile. For this DGP, the true CSQTE at $\tau = 0.75$ is given by $\mu_1(X; \tau) - \mu_0(X; \tau) = 1.29(e^{X_0 + X_1} - e^{X_0})$, which is heterogeneous in X.

[1] The result appears as Theorem 1 in v2 of the arxiv preprint.



(a) Flexible $\hat{\mu}$ (b) Misspecified $\hat{\mu}$ (c) Slow $\hat{\mu}$

Figure 2: Mean squared error (MSE) and 95% confidence interval coverage for different CSQTE learners. Shaded regions depict plus/minus one standard error over 100 simulations.

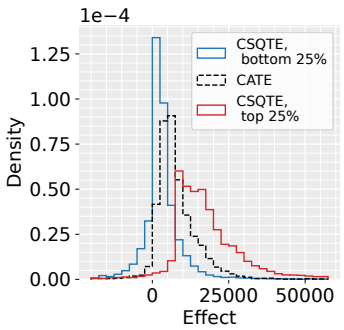
We estimate $e(X)$ using logistic regression and $q_a(X;\tau)$ using a quantile random forest (QRF) (Meinshausen and Ridgeway, 2006). We consider three options for estimating $\mu_a(X;\tau)$. First, we consider a *flexible* learner we term the superquantile RF (SQRF). SQRF uses a RF to calculate weights (like QRF) and then computes Eq. (4), replacing the expectation by the weighted average over the data. Second, we consider doing the same using a Gaussian kernel for calculating weights, choosing the bandwidth by Silverman's rule (Silverman, 2018). We refer to this as the *slow* learner because it suffers badly from the curse of dimension. Third, we consider estimating $\mu_a(X;\tau)$ using ordinary least squares for the regression $\frac{1}{1-\tau}\mathbb{E}[Y\mathbb{I}[Y \geq \widehat{q}_a(x;\tau)] \mid X = x]$ using the estimated quantile $\widehat{q}_a$. We term this estimator a *misspecified* learner since it is unlikely to fully capture the complexity of the superquantile (which is an exponential function of the features). For the final stage of Algorithm 1, we use either an ordinary least squares model (CSQTE+OLS) or a RF (CSQTE+RF). We set $K = 5$ as the number of folds required by Algorithm 1. All forests use `scikit-learn` (Pedregosa et al., 2011) defaults except for the minimum leaf size which is set to $n/20$ to control overfitting.

We compare the out-of-sample mean squared error (MSE) of the CSQTE estimator with that of the naive plug-in estimator from Eq. 7. To account for possible smoothing by the last stage regressor, we also construct Plugin+OLS and Plugin+RF given by running an additional OLS/RF model on the cross-fitted plug-in predictions. And, when the second stage algorithm is OLS, we check whether the 95%-confidence interval OLS returns for the X_1 coefficient contains the coefficient from the true projection. The results are shown in Figure 2. We run 100 simulations for each $n = 100, 200, \ldots, 12800$ and evaluate MSE over a fixed set of 500 random X values. Our CSQTE learner provides uniformly strong MSE performance, and the results show this is not just a consequence of the second-stage regression. For inference, we achieve good coverage whereas plug-in approaches yield little to no coverage. These findings confirm our theoretical results: the quadratic dependence on nuisance errors enables oracle rates when the nuisances are estimated slowly or are misspecified (Theorem 1, Theorem 2), and we can obtain valid inference when projecting onto linear spaces (Theorem 4).

8.2 Impact of 401(k) Eligibility on Financial Wealth

We apply the CSQTE estimator to study the impact of 401(k) eligibility on net financial assets. We use the dataset from Chernozhukov and Hansen (2004), which is based on the 1991 Survey of Income and Program Participation. The data contains 9,915 observations with 9 covariates such as age, income, education, family size, marital status, IRA participation, *etc.* While the eligibility for 401(k) (the treatment A) is not assigned at random, (Poterba et al., 1994; Chernozhukov and Hansen, 2004) argue that unconfoundedness can be assumed conditional on the observed covariates. The outcome of interest (Y) is the net financial assets of an individual, defined as the sum of 401(k) balance, bank accounts and interest-earning assets minus non-mortgage debt.

We apply the CSQTE estimator to understand effect heterogeneity beyond the conditional mean. Specifically, we estimate the average treatment effects on the bottom and top 25% of financial asset holders, conditional on covariates. We analyze the results alongside CATE estimates given by a DR-Learner (Kennedy, 2020) as implemented by Battocchi et al. (2019). For nuisance estimation, we use RF models with hyperparameters as in Chernozhukov et al. (2018). The superquanile model is SQRF described above, the quantile model is QRF, the outcome learner for CATE is a RF regression and the propensity model is a RF classifier.



Coefficient	CSQTE Bottom 25%	CATE	CSQTE Top 25%
Intercept	−0.021	−0.95	−2.07
($10,000)	(−1.06, 1.02)	(−2.42, 0.51)	(−7.04, 2.90)
Income	0.25**	0.21	−0.05
	(0.08, 0.43)	(−0.08, 0.50)	(−1.12, 1.01)
Age	105	232**	513
	(−75, 286)	(24, 441)	(−182, 1210)
Education	−801**	16	1340
	(−1440, −164)	(−1050, 1090)	(−2490, 5180)

Figure 3 & Table 1: Effect of 401(k) eligibility in dollars on the bottom 25%, top 25%, and average. The figure shows the distribution of CSQTEs and CATEs when using a random forest last stage. The table shows OLS inference on the projection of CSQTE and CATE on income, age, and education. "**" indicates statistical significance at level 0.05 (p-value < 0.05).

We consider two options for second-stage regressions. First, we consider using an RF model using all 9 covariates. We train the three estimators (CATE and upper/lower CSQTEs), and plot the distribution of predicted conditional effects on the 9,915 observations in Figure 3. We observe that 401(k) eligibility has a much higher financial impact on the high end of the conditional net worth distributions as compared to those on the lower end. While the CATE distribution reassuringly lies in between the bottom 25% and top 25% distributions, it fails to capture this disparity in effects.

To understand the heterogeneity in CSQTEs and CATEs, we use an OLS projection in the final stage. We choose the top three features that the RF last stage picked up as most important (see Figure 6 in Appendix C): income, age and education. The resulting coefficients and 95% confidence intervals are depicted in Table 1. For the bottom 25%, income and education are the main (statistically significant) drivers of the average effects. The positive income coefficient suggests that 401(k) eligibility provides more gains to those who potentially have the funds to invest. Likewise, the negative education coefficient means that the effects are higher among less educated earners. We hypothesize that higher educated earners might have a more comprehensive financial education and would save and invest regardless of 401(k) eligibility, whereas 401(k) availability provides lower educated earners a default investment option. For the top 25% asset holders, higher age and education rather than income have the largest impact on 401(k) eligibility returns. This group also displays more variability as none of the coefficients are statistically significant. The CATEs provide middle-of-the-road estimates that miss out on trends at the two ends of the spectrum. Thus, CSQTEs are a powerful tool for uncovering different behaviours across the conditional distribution that a simple average would obscure.

9 CONCLUDING REMARKS

We provide new tools to assess distributional effect heterogeneity through agnostic and robust learning of CDTEs in a generic framework that includes quantiles, super-quantiles, and f-risk measures. These tools can be used to analyze treatments as well as to support personalized decisions that take risk into account. We provide strong guarantees for both learning and inference, and demonstrate our methods in both synthetic and real data. We further discuss some limitations of our work in Appendix D.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1939704 and the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Award Number DE-SC0023112.

References

A. Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

C. Ai and X. Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

O. Atan, J. Jordon, and M. Van der Schaar. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

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

K. Battocchi, E. Dillon, M. Hei, G. Lewis, P. Oka, M. Oprescu, and V. Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Es-

timation. https://github.com/microsoft/EconML, 2019. Version 0.13.

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

A. Bennett and N. Kallus. The variational method of moments. *arXiv preprint arXiv:2012.09422*, 2020.

A. Bennett, N. Kallus, and T. Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019.

M. P. Bitler, J. B. Gelbach, and H. W. Hoynes. What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012, 2006.

G. Chamberlain. Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 567–596, 1992.

X. Chen and D. Pouzo. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1):46–60, 2009.

V. Chernozhukov and C. Hansen. The effects of 401 (k) participation on the wealth distribution: an instrumental quantile regression analysis. *Review of Economics and statistics*, 86(3):735–751, 2004.

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

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

V. Chernozhukov, I. Fernandez-Val, and M. Weidner. Network and panel quantile effects via distribution regression. *Journal of Econometrics*, 2020.

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

A. Curth, A. M. Alaa, and M. van der Schaar. Estimating structural target functions using machine learning and influence functions. *arXiv preprint arXiv:2008.06461*, 2020.

N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

J. Dorn, K. Guo, and N. Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*, 2021.

S. Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276, 2007.

D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Q. Ge, X. Huang, S. Fang, S. Guo, Y. Liu, W. Lin, and M. Xiong. Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection. *Frontiers in genetics*, page 1578, 2020.

P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

J. J. Heckman, J. Smith, and N. Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 1997.

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

M. Hohberg, P. Pütz, and T. Kneib. Treatment effects beyond the mean using distributional regression: Methods and guidance. *PloS one*, 15(2):e0226514, 2020.

H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13 (1):29–61, 2022.

K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

G. Imbens, N. Kallus, and X. Mao. Controlling for unmeasured confounding in panel data using minimal bridge functions: From two-way fixed effects to factor models. *arXiv preprint arXiv:2108.03849*, 2021.

F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.

N. Kallus. Treatment effect risk: Bounds and inference. *arXiv preprint arXiv:2201.05893*, 2022.

N. Kallus and X. Mao. Debiased inference on identified linear functionals of underidentified nuisances via penalized minimax estimation. *arXiv preprint arXiv:2208.08291*, 2022.

N. Kallus, X. Mao, and M. Uehara. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *arXiv preprint arXiv:1912.12945*, 2019.

E. H. Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

D. M. Kent, P. M. Rothwell, J. P. Ioannidis, D. G. Altman, and R. A. Hayward. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials*, 11(1):1–11, 2010.

K. Khosravi, G. Lewis, and V. Syrgkanis. Non-parametric inference adaptive to intrinsic dimension. *arXiv preprint arXiv:1901.03719*, 2019.

R. Koenker. *Quantile regression*, volume 38. Cambridge university press, 2005.

R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

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

L. Leqi and E. H. Kennedy. Median optimal treatment regimes. *arXiv preprint arXiv:2103.01802*, 2021.

N. Meinshausen and G. Ridgeway. Quantile regression forests. *Journal of Machine Learning Research*, 7(6), 2006.

J. Neyman. Optimal asymptotic tests of composite hypotheses. *Probability and statsitics*, pages 213–234, 1959.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

T. Olma. Nonparametric estimation of truncated conditional expectation functions. *arXiv preprint arXiv:2109.06150*, 2021.

M. Oprescu, V. Syrgkanis, and Z. S. Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941. PMLR, 2019.

J. Park, U. Shalit, B. Schölkopf, and K. Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. In *International Conference on Machine Learning*, pages 8401–8412. PMLR, 2021.

S. Passi and S. Barocas. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 39–48, 2019.

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

J. M. Poterba, S. F. Venti, and D. A. Wise. 401 (k) plans and tax-deferred saving. *Studies in the Economics of Aging*, pages 105–142, 1994.

J. M. Robins, S. D. Mark, and W. K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, pages 479–495, 1992.

J. M. Robins, L. Li, R. Mukherjee, E. T. Tchetgen, and A. van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

R. T. Rockafellar. *Conjugate duality and optimization*. SIAM, 1974.

A. Ruszczyński and A. Shapiro. Optimization of convex risk functions. *Mathematics of operations research*, 31 (3):433–452, 2006.

C. Shi, D. Blei, and V. Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

M. J. Van der Laan and J. M. Robins. *Unified methods for censored longitudinal data and causality*, volume 5. Springer, 2003.

A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

S. Vansteelandt and M. Joffe. Structural nested models and g-estimation: the partially realized promise. *Statistical Science*, 29(4):707–731, 2014.

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

T. Zhou and D. Carlson. Estimating potential outcome distributions with collaborating causal networks. *arXiv preprint arXiv:2110.01664*, 2021.

A PROOFS OF MAIN THEOREMS

A.1 Proof of Theorem 1

We first note the following useful identities:

1. (Functional analog of Taylor's expansion - 2nd order). Let $F : \mathcal{F}^d \to \mathbb{R}$ where $\mathcal{F}$ is a vector space of functions. For any $h, h' \in \mathcal{F}^d$, assume $t \mapsto F(t \cdot h + (1-t) \cdot h')$ has second order derivatives in an open interval containing $[0, 1]$, then $\exists\ \overline{h} \in \text{conv}(\{h, h'\})$ such that

$$F(h') = F(h) + \sum_{i=1}^{d} D_{h_i} F(h)(h'_i - h_i) + \sum_{i=1}^{d}\sum_{j=1}^{d} D_{h_i} D_{h_j} F(\overline{h})(h'_i - h_i)(h'_j - h_j)$$

2. For all $i \in \overline{1, m}, a \in \{0, 1\}, x \in \mathcal{X}$, the following hold:

$$\mathbb{E}[\rho_i(Y, \nu_a^*) \mid X = x, A = a] = 0 \qquad \text{(Moment equations)}$$

$$\sum_{i=1}^{m+1} \alpha_{a,i}(x, \nu_a^*) D_{\nu_{a,j}} \mathbb{E}[\rho_i(Y, \nu_a^*) \mid X = x, A = a] = \mathbb{I}[j = 1] \qquad \text{(Definition of } \alpha)$$

Proof. We wish to bound the term:

$$\mathcal{E}(e, \alpha, \nu) = \|\mathbb{E}[\psi(Z, e, \alpha, \nu) \mid X = x] - \mathbb{E}[\psi(Z, e^*, \alpha^*, \nu^*) \mid X = x]\|$$

To simplify our calculations, we write ψ as a difference of two terms $\psi^1 - \psi^0$ given by:

$$\psi^1(Z, e, \alpha, \nu) = \kappa_1(X) - \frac{A}{e(X)} \sum_{i=1}^{m+1} \alpha_{1,i}(X, \nu_{1,i}) \rho(Y, \nu_{1,i})$$

$$\psi^0(Z, e, \alpha, \nu) = \kappa_0(X) - \frac{1-A}{1-e(X)} \sum_{i=1}^{m+1} \alpha_{1,i}(X, \nu_{0,i}) \rho(Y, \nu_{0,i})$$

With this notation, we have:

$$\mathcal{E}(e, \alpha, \nu) \lesssim \sum_{a=0}^{1} \|\mathbb{E}\left[\psi^a(Z, e, \alpha, \nu) - \psi^a(Z, e^*, \alpha^*, \nu^*) \mid X = x\right]\|$$

The inequality above follows from the inequalities $(a+b)^2 \leq 2(a^2 + b^2)$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ (for non-negative a, b). Without loss of generality, we consider the bound for ψ^1. The proof for ψ^0 is very similar. We have:

$$\begin{aligned}
&\mathbb{E}\left[\psi^1(Z, e, \alpha, \nu) - \psi^1(Z, e^*, \alpha^*, \nu^*) \mid X = x\right] \\
&\quad = \sum_{a=0}^{1} \mathbb{E}\left[\psi^1(Z, e, \alpha, \nu) - \psi^1(Z, e^*, \alpha^*, \nu^*) \mid X = x, A = a\right] P(A = a \mid X = x) \\
&\quad = \kappa_1(x) - \kappa_1^*(x) - \frac{e^*(x)}{e(x)} \sum_{i=1}^{m+1} \alpha_{1,i}(x, \nu_1) \mathbb{E}[\rho_i(Y, \nu_1) \mid X = x, A = 1] \\
&\quad = \kappa_1(x) - \kappa_1^*(x) - \underbrace{\frac{e^*(x)}{e(x)} \sum_{i=1}^{m+1} (\alpha_{1,i}(x, \nu_1) - \alpha_{1,i}^*(x, \nu_1^*)) \mathbb{E}[\rho_i(Y, \nu_1) \mid X = x, A = 1]}_{\Lambda_1} \\
&\qquad\qquad - \underbrace{\frac{e^*(x)}{e(x)} \sum_{i=1}^{m+1} \alpha_{1,i}^*(x, \nu_1^*) \mathbb{E}[\rho_i(Y, \nu_1) \mid X = x, A = 1]}_{\Lambda_2}
\end{aligned}$$

where in the last equality we subtracted and then added back a term. We now study Λ_1 and Λ_2. For both these quantities, we do a Taylor expansion of $\mathbb{E}[\rho_i(Y,\nu_1) \mid X=x, A=1]$ around ν_1^*. For Λ_1, it suffices to use the Taylor expansion to first order only. For Λ_2, we perform a second order expansion:

$$\begin{aligned}
\mathbb{E}[\rho_i(Y,\nu_1) \mid X=x, A=1] &= \mathbb{E}[\rho_i(Y,\nu_1^*) \mid X=x, A=1] \\
&\quad + \sum_{j=1}^{m+1} D_{\nu_j}\mathbb{E}[\rho_i(Y,\overline{\nu}) \mid X=x, A=1](\nu_{1,j}(x) - \nu_{1,j}^*(x)) && \text{(first order expansion)} \\
\mathbb{E}[\rho_i(Y,\nu_1) \mid X=x, A=1] &= \mathbb{E}[\rho_i(Y,\nu_1^*) \mid X=x, A=1] \\
&\quad + \sum_{j=1}^{m+1} D_{\nu_j}\mathbb{E}[\rho_i(Y,\nu_1^*) \mid X=x, A=1](\nu_{1,j}(x) - \nu_{1,j}^*(x)) \\
&\quad + \frac{1}{2}\sum_{j=1}^{m+1}\sum_{l=1}^{m+1} D_{\nu_j}D_{\nu_l}\mathbb{E}[\rho_i(Y,\overline{\nu}) \mid X=x, A=1](\nu_{1,j}(x) - \nu_{1,j}^*(x))(\nu_{1,l}(x) - \nu_{1,l}^*(x)) \\
& && \text{(second order expansion)}
\end{aligned}$$

Then Λ_1 is given by:

$$\begin{aligned}
\Lambda_1 &= \frac{e^*(x)}{e(x)}\sum_{i=1}^{m+1}(\alpha_{1,i}(x,\nu_1) - \alpha_{1,i}^*(x,\nu_1^*))\mathbb{E}[\rho_i(Y,\nu_1) \mid X=x, A=1] \\
&= \frac{e^*(x)}{e(x)}\sum_{i=1}^{m+1}(\alpha_{1,i}(x,\nu_1) - \alpha_{1,i}^*(x,\nu_1^*))\Bigg(\cancelto{0}{\mathbb{E}[\rho_i(Y,\nu_1^*) \mid X=x, A=1]} \\
&\qquad\qquad + \sum_{j=1}^{m+1} D_{\nu_j}\mathbb{E}[\rho_i(Y,\overline{\nu}) \mid X=x, A=1](\nu_{1,j}(x) - \nu_{1,j}^*(x))\Bigg) \\
& && \text{(first order expansion)} \\
&= \frac{e^*(x)}{e(x)}\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} D_{\nu_j}\mathbb{E}[\rho_i(Y,\overline{\nu}) \mid X=x, A=1](\alpha_{1,i}(x,\nu_1) - \alpha_{1,i}^*(x,\nu_1^*))(\nu_{1,j}(x) - \nu_{1,j}^*(x)) \\
\|\Lambda_1\| &\le \frac{c_2}{c_1}\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} G_{ij}\|\alpha_{1,i} - \alpha_{1,i}^*\|\,\|\nu_{1,j} - \nu_{1,j}^*\| \\
&\lesssim \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} G_{ij}\|\alpha_{1,i} - \alpha_{1,i}^*\|\,\|\nu_{1,j} - \nu_{1,j}^*\|
\end{aligned}$$

We proceed in a similar way for Λ_2, but this time using the second order expansion:

$$\begin{aligned}
\Lambda_2 &= \frac{e^*(x)}{e(x)}\sum_{i=1}^{m+1}\alpha_{1,i}^*(x,\nu_1^*)\mathbb{E}[\rho_i(Y,\nu_1) \mid X=x, A=1] \\
&= \frac{e^*(x)}{e(x)}\sum_{i=1}^{m+1}\alpha_{1,i}^*(x,\nu_1^*)\Bigg(\cancelto{0}{\mathbb{E}[\rho_i(Y,\nu_1^*) \mid X=x, A=1]} \\
&\qquad + \sum_{j=1}^{m+1} D_{\nu_j}\mathbb{E}[\rho_i(Y,\nu_1^*) \mid X=x, A=1](\nu_{1,j}(x) - \nu_{1,j}^*(x)) \\
&\qquad + \frac{1}{2}\sum_{j=1}^{m+1}\sum_{l=1}^{m+1} D_{\nu_j}D_{\nu_l}\mathbb{E}[\rho_i(Y,\overline{\nu}_1) \mid X=x, A=1](\nu_{1,j}(x) - \nu_{1,j}^*(x))(\nu_{1,l}(x) - \nu_{1,l}^*(x))\Bigg) \\
&= \frac{e^*(x)}{e(x)}\sum_{j=1}^{m+1}\underbrace{\sum_{i=1}^{m+1}\alpha_{1,i}^*(x,\nu_1^*)D_{\nu_j}\mathbb{E}[\rho_i(Y,\nu_1^*) \mid X=x, A=1]}_{=\mathbb{I}[j=1]}(\nu_{1,j}(x) - \nu_{1,j}^*(x)) && \text{(see useful identities)}
\end{aligned}$$

$$+\frac{1}{2}\frac{e^*(x)}{e(x)}\sum_{i=1}^{m+1}\sum_{j=1}^{m+1}\sum_{l=1}^{m+1}\Bigg(\alpha^*_{1,i}(x,\nu^*_1)D_{\nu_j}D_{\nu_l}\mathbb{E}[\rho_i(Y,\overline{\nu}_1)\mid X=x,A=1]$$
$$\cdot(\nu_{1,j}(x)-\nu^*_{1,j}(x))(\nu_{1,l}(x)-\nu^*_{1,l}(x))\Bigg)$$
$$=\frac{e^*(x)}{e(x)}(\kappa_1(x)-\kappa^*_1(x))$$
$$+\frac{1}{2}\frac{e^*(x)}{e(x)}\sum_{i=1}^{m+1}\sum_{j=1}^{m+1}\sum_{l=1}^{m+1}\Bigg(\alpha^*_{1,i}(x,\nu^*_1)D_{\nu_j}D_{\nu_l}\mathbb{E}[\rho_i(Y,\overline{\nu}_1)\mid X=x,A=1]$$
$$\cdot(\nu_{1,j}(x)-\nu^*_{1,j}(x))(\nu_{1,l}(x)-\nu^*_{1,l}(x))\Bigg)$$

Adding back the κ_1 terms to Λ_2, we have:

$$\kappa_1(x)-\kappa_1(x)-\Lambda_2=\frac{1}{e(x)}(e(x)-e^*(x))(\kappa_1(x)-\kappa^*_1(x))$$
$$+\frac{1}{2}\frac{e^*(x)}{e(x)}\sum_{i=1}^{m+1}\sum_{j=1}^{m+1}\sum_{l=1}^{m+1}\Bigg(\alpha^*_{1,i}(x,\nu^*_1)D_{\nu_j}D_{\nu_l}\mathbb{E}[\rho_i(Y,\overline{\nu}_1)\mid X=x,A=1]$$
$$\cdot(\nu_{1,j}(x)-\nu^*_{1,j}(x))(\nu_{1,l}(x)-\nu^*_{1,l}(x))\Bigg)$$
$$\|\kappa_1(x)-\kappa_1(x)-\Lambda_2\|\le\frac{1}{c_1}\|\kappa_1-\kappa^*_1\|\|e-e^*\|+\frac{c_3c_4}{2c_1}\sum_{j=1}^{m+1}\sum_{l=1}^{m+1}H_{jl}\|\nu_{1,j}-\nu^*_{1,j}\|\,\|\nu_{1,l}-\nu^*_{1,l}\|$$
$$\lesssim\|\kappa_1-\kappa^*_1\|\|e-e^*\|+\sum_{j=1}^{m+1}\sum_{l=1}^{m+1}H_{jl}\|\nu_{1,j}-\nu^*_{1,j}\|\,\|\nu_{1,l}-\nu^*_{1,l}\|$$

Constants c_1,c_2,c_3,c_4 and binary matrices H,G are defined in Assumption 1. Putting everything together, we have:

$$\left\|\mathbb{E}\left[\psi^1(Z,e,\alpha,\nu)-\psi^1(Z,e^*,\alpha^*,\nu^*)\mid X=x\right]\right\|=\|\kappa_1(x)-\kappa_1(x)-\Lambda_2-\Lambda_1\|$$
$$\lesssim\|\kappa_1(x)-\kappa_1(x)-\Lambda_2\|+\|\Lambda_1\|$$
$$\lesssim\|\kappa_1-\kappa^*_1\|\|e-e^*\|+\sum_{j=1}^{m+1}\sum_{l=1}^{m+1}H_{jl}\|\nu_{1,j}-\nu^*_{1,j}\|\,\|\nu_{1,l}-\nu^*_{1,l}\|$$
$$+\sum_{i=1}^{m+1}\sum_{j=1}^{m+1}G_{ij}\|\alpha_{1,i}-\alpha^*_{1,i}\|\,\|\nu_{1,j}-\nu^*_{1,j}\|$$

Putting the ψ^0 and ψ^1 bounds together, we obtain the desired result:

$$\mathcal{E}(e,\alpha,\nu)\lesssim\sum_{a=1}^{1}\Bigg(\|\kappa_a-\kappa^*_a\|\,\|e-e^*\|+\sum_{i=1}^{m+1}\sum_{j=1}^{m+1}G_{ij}\|\alpha_{a,i}-\alpha^*_{a,i}\|\,\|\nu_{a,j}-\nu^*_{a,j}\|$$
$$+\sum_{i=1}^{m+1}\sum_{j=1}^{m+1}H_{ij}\|\nu_{a,i}-\nu^*_{a,i}\|\,\|\nu_{a,j}-\nu^*_{a,j}\|\Bigg)$$

□

A.2 Proof of Theorem 2

Proof. Let $I_k=\{i:i=k-1\pmod K\}$ and $I_k^C=\{i:i\neq k-1\pmod K\}$ be the k^{th} data fold and its complement, let $n_k=|I_k|\in\{\lfloor n/K\rfloor,\lceil n/K\rceil\}$, and let $\lambda_k=n_k/n$ (note none of I_k,n_k,λ_k are random). Let $\mathbb{P}g(Z)$ denote expectation

over Z alone (that is, conditioning on the data, should g depend on it) and $\hat{\mathbb{P}}_k g(Z)$ denote empirical expectation over I_k. Further, let $\mathbb{P}(g(Z) \mid X)$ denote the conditional expectation, $\|g\|_2 = \mathbb{P}g^2$, and $\Pi_{\mathcal{F}}(g) \in \operatorname{argmin}_{f \in \mathcal{F}} \|f - g\|_2$.

Define the pseudo-outcome random variables:

$$\begin{aligned}
\psi &= \psi(Z, e^*, \alpha^*, \nu^*), \\
\hat{\psi}_k &= \psi(Z, \hat{e}^{(k)}, \hat{\alpha}^{(k)}, \hat{\nu}^{(k)}), \\
\hat{\psi} &= \sum_{k=1}^{K} \lambda_k \hat{\psi}_k.
\end{aligned}$$

Further, define the squared-error objectives:

$$\begin{aligned}
\hat{R}_k(f) &= \hat{\mathbb{P}}_k(f(X) - \hat{\psi}_k)^2, \\
\hat{R}(f) &= \sum_{k=1}^{K} \lambda \hat{R}_k(f), \\
\tilde{R}_k(f) &= \hat{\mathbb{P}}_k(f(X) - \hat{\psi}_k)^2, \\
\tilde{R}(f) &= \sum_{k=1}^{K} \lambda \tilde{R}_k(f), \\
\bar{R}(f) &= \mathbb{P}(f(X) - \hat{\psi})^2, \\
R(f) &= \mathbb{P}(f(X) - \psi)^2.
\end{aligned}$$

Note that $\tilde{R}(f) - \bar{R}(f) = \tilde{R}(f') - \bar{R}(f')$ for any f, f', that is, $\tilde{R}(f)$ and $\bar{R}(f)$ are equal up to additive constants. Finally, we have the predictors:

$$\begin{aligned}
\hat{f} &\in \operatorname*{argmin}_{f \in \mathcal{F}} \hat{R}(f), \\
\bar{f} = \Pi_{\mathcal{F}}(\mathbb{P}(\hat{\psi} \mid X)) &\in \operatorname*{argmin}_{f \in \mathcal{F}} \bar{R}(f) = \operatorname*{argmin}_{f \in \mathcal{F}} \tilde{R}(f), \\
f^* = \Pi_{\mathcal{F}}(\mathbb{P}(\psi \mid X)) &\in \operatorname*{argmin}_{f \in \mathcal{F}} R(f).
\end{aligned}$$

Note $\hat{f} = \widehat{\mathrm{CDTE}}$ and $f^* = \mathrm{CDTE}$; we rename them for brevity.

We then have:

$$\begin{aligned}
\left\|\hat{f} - f^*\right\| &\leq \left\|\hat{f} - \bar{f}\right\| + \|\bar{f} - f^*\| \\
&= \left\|\hat{f} - \bar{f}\right\| + \left\|\Pi_{\mathcal{F}}(\mathbb{P}(\hat{\psi} \mid X)) - \Pi_{\mathcal{F}}(\mathbb{P}(\psi \mid X))\right\| \\
&\leq \underbrace{\left\|\hat{f} - \bar{f}\right\|}_{\mathrm{Error}_A} + \underbrace{\left\|\mathbb{P}(\hat{\psi} \mid X) - \mathbb{P}(\psi \mid X)\right\|}_{\mathrm{Error}_B},
\end{aligned}$$

where the inequality is because $\mathcal{F}$ is closed convex

We first tackle Error_B:

$$\begin{aligned}
\left\|\mathbb{P}(\hat{\psi} \mid X) - \mathbb{P}(\psi \mid X)\right\| &= \left\|\sum_{k=1}^{K} \lambda_k (\mathbb{P}(\hat{\psi}_k \mid X) - \mathbb{P}(\psi \mid X))\right\| \\
&\leq \sum_{k=1}^{K} \lambda_k \left\|\mathbb{P}(\hat{\psi}_k \mid X) - \mathbb{P}(\psi \mid X)\right\|,
\end{aligned}$$

which we bound as $\lesssim \mathcal{E}$ by Theorem 1.

Next, we address Error_A. Note that $\left\|\hat{f} - \bar{f}\right\| \geq t$ means that $\tilde{R}(\hat{f}) - \tilde{R}(\bar{f}) \geq t^2$ and $\hat{R}(\hat{f}) - \hat{R}(\bar{f}) \leq 0$. By intermediate value theorem and convexity of $\tilde{R}, \hat{R}, \mathcal{F}$, we have for some $f \in [\hat{f}, \bar{f}]$ that $\tilde{R}(f) - \tilde{R}(\bar{f}) = t^2$ and $\hat{R}(f) - \hat{R}(\bar{f}) \leq 0$.

This in turn implies that $\exists f \in \mathcal{F}$ with $\left\|f - \bar{f}\right\| \leq t$ and $(\tilde{R}(f) - \tilde{R}(\bar{f})) - (\hat{R}(f) - \hat{R}(\bar{f})) \geq t^2$. Because $\tilde{R} - \hat{R}$ is average of $\tilde{R}_k - \hat{R}_k$, this in turn implies that for at least one $k \in \overline{1, K}$ we have $\exists f \in \mathcal{F}$ with $\left\|f - \bar{f}\right\| \leq t$ and $(\tilde{R}_k(f) - \tilde{R}_k(\bar{f})) - (\hat{R}_k(f) - \hat{R}_k(\bar{f})) \geq t^2$. Since $|\psi| \leq c_5, \left|\hat{\psi}_k\right| \leq c_5, |f(x)| \leq c_5$, we have $(f(X) - \hat{\psi}_k)^2 - (\bar{f}(X) - \hat{\psi}_k)^2 \leq 4c_5 \left|f(X) - \bar{f}(X)\right|$.

Therefore, by union bound and iterated expectations,

$$\mathbb{P}\left(\left\|\hat{f} - \bar{f}\right\| \geq t\right) \leq \sum_{k=1}^{K} \mathbb{E}\, \mathbb{P}\left(\exists g \in \{\left|f - \bar{f}\right| : f \in \mathcal{F}\} \,:\, \mathbb{P}g^2 \leq t^2,\ (\mathbb{P} - \hat{\mathbb{P}}_k)g \geq t^2/(4c_5) \mid \{Z_i : i \in I_k^C\}\right).$$

By lemma 3.4.2 of van der Vaart and Wellner (1996) and Markov's inequality, we have that

$$\mathbb{P}\left(\exists g \in \{\left|f - \bar{f}\right| : f \in \mathcal{F}\} \,:\, \mathbb{P}g^2 \leq t^2,\ (\mathbb{P} - \hat{\mathbb{P}}_k)g \geq t^2/(4c_5) \mid \{Z_i : i \in I_k^C\}\right) \leq 4c_5 J(1 + c_5 J/(\sqrt{n_k}t^2))/t^2,$$

where $J = t + \int_0^t \sqrt{\epsilon^{-r}} d\epsilon \leq t + \frac{2}{2-r} t^{1-r/2}$. Thus, $\text{Error}_A = O_p(n^{-1/(2+r)})$. □

A.3 Proof of Theorem 3

Proof. Using the stability conditions for the estimator $\widehat{\mathbb{E}}_n$ outlined in Theorem 3, we can immediately apply Theorem 1 from Kennedy (2020) (as appears in the v2 preprint on arxiv) with the cross-fitted nuisances to obtain:

$$\begin{aligned}
\left\|(\widehat{\text{CDTE}}) - (\text{CDTE})\right\| &\lesssim \left\|(\widetilde{\text{CDTE}}) - (\text{CDTE})\right\| \\
&\quad + \sum_{k=1}^{K} \left\|\mathbb{E}[\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) \mid X = x] - \mathbb{E}[\psi(Z, e^*, \alpha^*, \nu^*) \mid X = x]\right\| \\
&\lesssim \left\|(\widetilde{\text{CDTE}}) - (\text{CDTE})\right\| + \mathcal{E}
\end{aligned}$$

where the last inequality was obtained by applying Theorem 1 to nuisance sets $(\widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)})$ in each fold $k \in \overline{1, K}$. □

A.4 Proof of Theorem 4

Proof. Let $\mathcal{I}_k = \{i : i = k - 1 \pmod K\}$ be the data indices in the k-th data fold, and let $\widehat{\mathbb{E}}_k f(Z) = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} f(Z_i)$ be the empirical expectation of data with indices in $\mathcal{I}_k$. The linear regression parameters $\gamma^*, \widetilde{\gamma}$ and $\widehat{\gamma}$ are given by:

$$\begin{aligned}
\gamma^* &= \mathbb{E}[\phi(X)\phi(X)^T]^{-1}\mathbb{E}[\text{CDTE}(X)\phi(X)] \\
&= \mathbb{E}[\phi(X)\phi(X)^T]^{-1}\mathbb{E}[\psi(Z, e^*, \alpha^*, \nu^*)\phi(X)] && \text{(consistency and iterated expectations)} \\
\widetilde{\gamma} &= \widehat{\mathbb{E}}_n[\phi(X)\phi(X)^T]^{-1}\widehat{\mathbb{E}}_n[\psi(Z, e^*, \alpha^*, \nu^*)\phi(X)] \\
&= \widehat{\mathbb{E}}_n[\phi(X)\phi(X)^T]^{-1}\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_k[\psi(Z, e^*, \alpha^*, \nu^*)\phi(X)] \\
\widehat{\gamma} &= \widehat{\mathbb{E}}_n[\phi(X)\phi(X)^T]^{-1}\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathbb{E}}_k[\psi(Z, \widehat{e}^{(k)}, {}^{(k)}, \widehat{\nu}^{(k)})\phi(X)]
\end{aligned}$$

We note that we can rewrite $\sqrt{n}(\widehat{\gamma} - \gamma^*)$ as:

$$\sqrt{n}(\widehat{\gamma} - \gamma^*) = \sqrt{n}(\widehat{\gamma} - \widetilde{\gamma}) + \underbrace{\sqrt{n}(\widetilde{\gamma} - \gamma^*)}_{\rightsquigarrow \mathcal{N}(0, \Sigma^*)} \tag{14}$$

The second term converges in distribution to the desired $\mathcal{N}(0, \Sigma^*)$. Thus, it suffices to show that the first term is $o_p(1)$. We decompose the first term into two components and study them separately:

$$\begin{aligned}
&\sqrt{n}(\widehat{\gamma} - \widetilde{\gamma}) \qquad (15)\\
&= \sqrt{n}\widehat{\mathbb{E}}_n[\phi(X)\phi(X)^T]^{-1}\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\mathbb{E}}_k[\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)})\phi(X)] - \widehat{\mathbb{E}}_k[\psi(Z, e^*, \alpha^*, \nu)\phi(X)]\right)\\
&= \sqrt{n}\widehat{\mathbb{E}}_n[\phi(X)\phi(X)^T]^{-1}\frac{1}{K}\sum_{k=1}^{K}\underbrace{\left(\mathbb{E}[\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)})\phi(X)] - \mathbb{E}[\psi(Z, e^*, \alpha^*, \nu^*)\phi(X)]\right)}_{\Lambda_{1,k}}\\
&\quad + \sqrt{n}\widehat{\mathbb{E}}_n[\phi(X)\phi(X)^T]^{-1}\frac{1}{K}\sum_{k=1}^{K}\underbrace{(\widehat{\mathbb{E}}_k - \mathbb{E})\left[(\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)})\phi(X)] - \psi(Z, e^*, \alpha^*, \nu^*))\phi(X)\right]}_{\Lambda_{2,k}}
\end{aligned}$$

We now show that $\Lambda_{1,k}$ and $\Lambda_{2,k}$ in the equation above are both $o_p(1/\sqrt{n})$. Let $\Sigma_\phi = \mathbb{E}[\phi(X)\phi(X)^T]$. Then, each element of $\Lambda_{1,k}$ is given by:

$$\begin{aligned}
(\Lambda_{1,k})_i &= \mathbb{E}[\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)})(\phi(X))_i] - \mathbb{E}[\psi(Z, e^*, \alpha^*, \nu^*)\phi(X)_i]\\
&= \mathbb{E}[\mathbb{E}[\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi(Z, e^*, \alpha^*, \nu^*) \mid X]\phi(X)_i]\\
\Rightarrow |(\Lambda_{1,k})_i| &\leq \|\mathbb{E}[\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi(Z, e^*, \alpha^*, \nu^*) \mid X]\| \, \|\phi(X)_i\| \qquad \text{(Cauchy-Schwartz)}\\
&\lesssim \left\{\sum_{a=1}^{1}\left(\|\widehat{\kappa}_a^{(k)} - \kappa_a^*\| \, \|\widehat{e}^{(k)} - e^*\| + \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} G_{ij}\|\widehat{\alpha}_{a,i}^{(k)} - \alpha_{a,i}^*\| \, \|\widehat{\nu}_{a,j}^{(k)} - \nu_{a,j}^*\|\right.\right.\\
&\qquad \left.\left. + \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} H_{ij}\|\widehat{\nu}_{a,i}^{(k)} - \nu_{a,i}^*\| \, \|\widehat{\nu}_{a,j}^{(k)} - \nu_{a,j}^*\|\right)\right\}\sqrt{(\Sigma_\phi)_{ii}} \qquad \text{(Thm. 1)}
\end{aligned}$$

By the theorem's assumptions, we have that $\Lambda_{1,k}$ is $o_p(1/\sqrt{n})$, as desired. We now see how we can control $\Lambda_{2,k}$. By Chebyshev's inequality, we have that $\Lambda_{2,k}$ is

$$o_p\left(n^{-1/2}\sum_{i=1}^{p}\mathbb{E}[(\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi(Z, e^*, \alpha^*, \nu^*))^2(\phi(X)_i)^2]^{1/2}\right) \qquad (16)$$

where we leveraged that $|\mathcal{I}_k| \simeq n/K$ and K is a fixed integer constant that doesn't depend on n. The sum in the expression above further reduces to:

$$\begin{aligned}
&\sum_{i=1}^{p}\mathbb{E}[(\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi(Z, e^*, \alpha^*, \nu^*))^2(\phi(X))_i^2]^{1/2}\\
&\qquad \leq \sum_{i=1}^{p}\|\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi(Z, e^*, \alpha^*, \nu^*)\|\|\phi(X)_i\|\\
&\qquad = \sum_{i=1}^{p}\sqrt{(\Sigma_\phi)_{ii}}\,\|\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi(Z, e^*, \alpha^*, \nu^*)\|
\end{aligned}$$

Thus, it remains to study the convergence of $\|\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi(Z, e^*, \alpha^*, \nu^*)\|$. We note that:

$$\begin{aligned}
\|\psi(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi(Z, e^*, \alpha^*, \nu^*)\| &= \sum_{a=0}^{1}\|\psi^a(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi^a(Z, e^*, \alpha^*, \nu^*)\|\\
&\leq \sum_{a=0}^{1}\left(\|\psi^a(Z, \widehat{e}^{(k)}, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi^a(Z, e^*, \widehat{\alpha}, \widehat{\nu})\|\right.\\
&\qquad \left. + \|\psi^a(Z, e^*, \widehat{\alpha}^{(k)}, \widehat{\nu}^{(k)}) - \psi^a(Z, e^*, \alpha^*, \nu^*)\|\right) \qquad \text{(Cauchy-Schwartz)}
\end{aligned}$$

where the ψ^a's are defined in the proof of Theorem 1. Without loss of generality, we study the convergence of ψ^1:

$$\begin{aligned}
&\|\psi^1(Z,\widehat{e}^{(k)},\widehat{\alpha}^{(k)},\widehat{\nu}) - \psi^1(Z,e^*,\widehat{\alpha}^{(k)},\widehat{\nu}^{(k)})\| \\
&\quad= \left\| \frac{A(\widehat{e}^{(k)}(X) - e^*(X))}{e^*(X)\widehat{e}^{(k)}(X)} \sum_{i=1}^{m+1} \widehat{\alpha}_{1,i}^{(k)}(X,\widehat{\nu}_1^{(k)})\rho_i(Y,\widehat{\nu}_1^{(k)}) \right\| \\
&\quad\leq \frac{(m+1)c_4c_5}{c_1^2}\|\widehat{e}^{(k)} - e^*\| \qquad \text{(from boundedness assumptions)} \\
&\quad\lesssim \|\widehat{e}^{(k)} - e^*\| \\
&\|\psi^1(Z,e^*,\widehat{\alpha},\widehat{\nu}) - \psi^1(Z,e^*,\alpha^*,\nu^*)\| \\
&\quad= \left\| \widehat{\kappa}_1^{(k)}(X) - \kappa_1(X) - \frac{A}{e^*(X)} \left(\sum_{i=1}^{m+1} \widehat{\alpha}_{1,i}^{(k)}(X,\widehat{\nu}_1^{(k)})\rho_i(Y,\widehat{\nu}_1^{(k)}) - \sum_{i=1}^{m+1} \alpha_{1,i}^*(X,\nu_1^*)\rho_i(Y,\nu_1^*) \right) \right\| \\
&\quad\leq \|\widehat{\kappa}_1^{(k)} - \kappa_1^*\| + \frac{c_5}{c_1}\sum_{i=1}^{m} \|\widehat{\alpha}_{1,i}^{(k)} - \alpha_{1,i}^*\| \\
&\quad\lesssim \|\widehat{\kappa}_1^{(k)} - \kappa_1^*\| + \sum_{i=1}^{m} \|\widehat{\alpha}_{1,i}^{(k)} - \alpha_{1,i}^*\|
\end{aligned}$$

Therefore:

$$\|\psi(Z,\widehat{e}^{(k)},\widehat{\alpha},\widehat{\nu}) - \psi(Z,e^*,\alpha^*,\nu^*)\| \lesssim \|\widehat{e}^{(k)} - e^*\| + \sum_{a=0}^{1} \left(\|\kappa_a^{(k)} - \kappa_a^*\| + \sum_{i=1}^{m} \|\widehat{\alpha}_{a,i}^{(k)} - \alpha_{a,i}^*\| \right)$$

Given our assumptions on the nuisance convergence rates, the term above is $o_p(1)$. Putting everything together, we obtain that $\Lambda_{1,k} = o_p(1/\sqrt{n})$ and $\Lambda_{1,k} + \Lambda_{2,k} = o_p(1/\sqrt{n})$. Going back to Eq. 15, we have that:

$$\sqrt{n}(\widehat{\gamma} - \widetilde{\gamma}) = \sqrt{n}\widehat{\mathbb{E}}_n[\phi(X)\phi(X)^T]^{-1}\frac{1}{K}\sum_{k=1}^{K}(\Lambda_{1,k} + \Lambda_{2,k})$$

Using the continuous mapping theorem, we have that $\widehat{\mathbb{E}}_n[\phi(X)\phi(X)^T]^{-1} \xrightarrow{P} \Sigma_\phi^{-1}$. Finally, by Slutsky's theorem and the fact that K does not grow with n, we obtain that $\sqrt{n}(\widehat{\gamma} - \widetilde{\gamma})$ is $o_p(1)$ and therefore:

$$\sqrt{n}(\widehat{\gamma} - \gamma^*) \rightsquigarrow \mathcal{N}(0,\Sigma^*)$$

as desired.

□

B APPLICATIONS OF THEOREM 1

B.1 Pseudo-outcomes and Rates for CQTE

Corollary 5 (Pseudo-outcomes and rates for CQTE)**.** *Assume the distribution F is continuous. Then, the pseudo-outcome for the conditional quantile treatment at level τ, CQTE$(x;\tau)$, is given by:*

$$\psi^{CQTE}(Z,e,q,f) = q_1(X;\tau) - q_0(X;\tau) + \frac{A - e(X)}{e(X)(1-e(X))}\frac{1}{f_A(x)}\left(\tau - \mathbb{I}[Y \leq q_A(x;\tau)]\right)$$

where $f_a(x) = f_{Y|X=x,A=a}(q_a(x;\tau))$ is the conditional density of Y evaluated at the conditional quantile. Furthermore, if the conditions of Assumption 1 are satisfied, the oracle rate deviation is given by:

$$\mathcal{E} \leq \sum_{k=1}^{K}\sum_{a=0}^{1} \|\widehat{q}_a^{(k)} - q_a^*\| \left(\|\widehat{e}^{(k)} - e^*\| + \left\| \frac{1}{\widehat{f}_a^{(k)}} - \frac{1}{f_a^*} \right\| + \|\widehat{q}_a^{(k)} - q_a^*\| \right)$$

Remark 4. We note that by Assumption 1, f_a and $\widehat{f}_a$ are bounded away from 0, and thus the rate deviation can be written as $\sum_{k=1}^{K}\sum_{a=0}^{1}\|\widehat{q}_a^{(k)}-q_a^*\|\left(\|\widehat{e}-e^*\|+\|\widehat{f}_a^{(k)}-f_a^*\|+\|\widehat{q}_a^{(k)}-q_a^*\|\right)$. When considering medians ($\alpha=0.5$), this result reduces to the result in Leqi and Kennedy (2021) for median effects.

Proof. For continuous CDFs, the conditional quantile at level τ is the solution to the moment:

$$\mathbb{E}[\tau-\mathbb{I}[Y\leq q_a(x;\tau)]\mid X=x,A=a]=0$$

The Jacobian is given by $J_a^*(X)=D_{q_a}\{\mathbb{E}[\tau-\mathbb{I}[Y\leq q_a(X;\tau)]\mid X=x,A=a]\}\big|_{q_a=q_a^*}=-f_a^*(x)$ where $f_a(x)=f_{Y|X=x,A=a}(q_a(x;\tau))$ is the conditional density evaluated at the conditional quantile. By Definition 3, $\alpha_a^*(X)=-1/f_a^*(X)$ and the pseudo-outcome is given by:

$$\psi^{\text{CQTE}}(Z,e,q,f)=q_1(X;\tau)-q_0(X;\tau)+\frac{A-e(X)}{e(X)(1-e(X))}\frac{1}{f_A(x)}\left(\tau-\mathbb{I}[Y\leq q_A(x;\tau)]\right)$$

We apply Theorem 1 with $\alpha_a=1/f_a$, $\nu_a=q_a$ and obtain that the oracle deviation is given by:

$$\mathcal{E}\leq\sum_{k=1}^{K}\sum_{a=0}^{1}\|\widehat{q}_a^{(k)}-q_a^*\|\left(\|\widehat{e}^{(k)}-e^*\|+\left\|\frac{1}{\widehat{f}_a^{(k)}}-\frac{1}{f_a^*}\right\|+\|\widehat{q}_a^{(k)}-q_a^*\|\right)$$

as desired.

□

B.2 Pseudo-outcomes and Rates for CSQTE

Corollary 6 (Pseudo-outcomes and rates for CSQTE)**.** *Assume the distribution F is continuous. Then, the pseudo-outcome for the conditional super-quantile treatment at level τ, CSQTE$(x;\tau)$, is given by:*

$$\begin{aligned}\psi^{\textit{CSQTE}}(Z,e,\mu,q)=\mu_1(X;\tau)-\mu_0(X;\tau)\\+\frac{A-e(X)}{e(X)(1-e(X))}\Big(q_A(X;\tau)+\frac{1}{1-\tau}(Y-q_A(X;\tau))\mathbb{I}[Y\geq q_A(X;\tau)]-\mu_A(X;\tau)\Big)\end{aligned}$$

Furthermore, if the conditions of Assumption 1 are satisfied, the oracle rate deviation is given by:

$$\mathcal{E}\leq\sum_{k=1}^{K}\sum_{a=0}^{1}\left(\|\widehat{\mu}_a^{(k)}-\mu_a^*\|\|\widehat{e}^{(k)}-e^*\|+\|\widehat{q}_a^{(k)}-q_a^*\|^2\right)$$

Proof. From Example 3.2, we have that for continuous CDFs, the conditional super-quantile at level τ is the solution to the conditional moments:

$$\begin{aligned}&\mathbb{E}[(1-\tau)^{-1}Y\mathbb{I}[Y\geq q_a(x;\tau)]-\mu_a(x;\tau)\mid X=x,A=a]=0\\&\mathbb{E}[\tau-\mathbb{I}[Y\leq q_a(x;\tau)]\mid X=x,A=a]=0\end{aligned}$$

Thus, the Jacobian $J_a^*(X)$ and its inverse if given by:

$$J_a^*(X)=\begin{pmatrix}-1 & -(1-\tau)^{-1}q_a(X;\tau)f_a(X)\\0 & -f_a(X)\end{pmatrix},\quad (J_a^*(X))^{-1}=\begin{pmatrix}-1 & (1-\tau)^{-1}q_a(X;\tau)\\0 & -1/f_a(X)\end{pmatrix}$$

where $f_a(x)=f_{Y|X=x,A=a}(q_a(x;\tau))$ is the conditional density evaluated at the conditional quantile. By Definition 3 with $\alpha_a=(-1,-(1-\tau)^{-1}q_a(X;\tau))$, $\nu_a=(\mu_a,q_a)$, the pseudo-outcome is given by:

$$\begin{aligned}\psi^{\text{CSQTE}}(Z,e,\mu,q)&=\mu_1(X;\tau)-\mu_0(X;\tau)+\frac{A-e(X)}{e(X)(1-e(X))}\frac{1}{1-\tau}\Big(\\&\qquad Y\mathbb{I}[Y\geq q_A(X;\tau)]-(1-\tau)\mu_A(X;\tau)-q_A(X;\tau)\left(\tau-\mathbb{I}[Y\leq q_A(X;\tau)]\right)\Big)\\&=\mu_1(X;\tau)-\mu_0(X;\tau)\\&\quad+\frac{A-e(X)}{e(X)(1-e(X))}\Big(q_A(X;\tau)+\frac{1}{1-\tau}(Y-q_A(X;\tau))\mathbb{I}[Y\geq q_A(X;\tau)]-\mu_A(X;\tau)\Big)\end{aligned}$$

We now apply Theorem 1. We first note that the binary matrices G and H are given by:

$$G = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad H = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Moreover, $\|\widehat{\alpha}_{a,1}^{(k)} - \alpha_{a,1}^*\| = 0$ since $\alpha_{a,1} = -1$ is a constant. Therefore, in the cross-products $\|\widehat{\alpha}_{a,i}^{(k)} - \alpha_{a,i}^*\|\|\widehat{\nu}_{a,j}^{(k)} - \nu_{a,j}^*\|$, the only surviving term is $\|\widehat{\alpha}_{a,2}^{(k)} - \alpha_{a,2}^*\|\|\widehat{\nu}_{a,2}^{(k)} - \nu_{a,2}^*\| = \|\widehat{q}_a^{(k)} - q_a^*\|^2$ due to $G_{21} = 0$. Furthermore, given H, the only surviving term in the cross-products $\|\widehat{\nu}_{a,i}^{(k)} - \nu_{a,i}^*\|\|\widehat{\nu}_{a,j}^{(k)} - \nu_{a,j}^*\|$ is $\|\widehat{\nu}_{a,2}^{(k)} - \nu_{a,2}^*\|^2 = \|\widehat{q}_a^{(k)} - q_a^*\|^2$. Thus, the oracle deviation is given by:

$$\mathcal{E} \leq \sum_{k=1}^{K} \sum_{a=0}^{1} \left(\|\widehat{\mu}_a^{(k)} - \mu_a^*\|\|\widehat{e}^{(k)} - e^*\| + \|\widehat{q}_a^{(k)} - q_a^*\|^2 \right)$$

as desired. □

B.3 Pseudo-outcomes and Rates for CfRTE

Corollary 7 (Pseudo-outcomes and rates for CfRTE)**.** *The conditional f-risk treatment effect at level δ, CfRTE$(x;\delta)$, is given by:*

$$\psi^{\mathrm{C}f\mathrm{RTE}}(Z, e, R, \beta, \lambda) = R_1^f(X;\delta) - R_0^f(X;\delta) + \frac{A - e(X)}{e(X)(1-e(X))}(m(Y, \beta_A, \lambda_A;\delta) - R_A^f(X;\delta))$$

Furthermore, if the conditions of Assumption 1 are satisfied, the oracle rate deviation is given by:

$$\mathcal{E} \leq \sum_{k=1}^{K} \sum_{a=0}^{1} \left(\|\widehat{R}_a^{f,(k)} - R_a^{f,*}\|\|\widehat{e}^{(k)} - e^*\| + \|\widehat{\beta}_a^{(k)} - \beta_a^*\|^2 + \|\widehat{\lambda}_a^{(k)} - \lambda_a^*\|^2 + \|\widehat{\beta}_a^{(k)} - \beta_a^*\|\|\widehat{\lambda}_a^{(k)} - \lambda_a^*\| \right)$$

Proof. From Example 3.3, we have that the CfRTE is identified by the following conditional moments:

$$\begin{aligned} &\mathbb{E}[m(Z, \beta_a, \lambda_a;\delta) - R_a^f(x;\delta) \mid X = x, A = a] = 0 \\ &\mathbb{E}\left[\frac{\partial}{\partial \beta_a} m(Z, \beta_a, \lambda_a;\delta) \middle| X = x, A = a \right] = 0 \\ &\mathbb{E}\left[\frac{\partial}{\partial \lambda_a} m(Z, \beta_a, \lambda_a;\delta) \middle| X = x, A = a \right] = 0 \end{aligned}$$

where it is understood that β_a and λ_a are functions of X. Thus, the Jacobian $J_a^*(X)$ and its inverse if given by:

$$J_a^*(X) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -\mathbb{E}\left[\frac{(Y-\lambda_a(X))^2}{\beta_a(X)^3} (f^*)''\left(\frac{Y-\lambda_1(X)}{\beta_a(X)}\right) \middle| X=x, A=1 \right] & \mathbb{E}\left[\frac{Y-\lambda_a(X)}{\beta_a(X)^2} (f^*)''\left(\frac{Y-\lambda_a(X)}{\beta_a(X)}\right) \middle| X=x, A=1 \right] \\ 0 & \mathbb{E}\left[\frac{Y-\lambda_a(X)}{\beta_a(X)^2} (f^*)''\left(\frac{Y-\lambda_a(X)}{\beta_a(X)}\right) \middle| X=x, A=1 \right] & -\mathbb{E}\left[\frac{1}{\beta_a(X)} (f^*)''\left(\frac{Y-\lambda_a(X)}{\beta_a(X)}\right) \middle| X=x, A=1 \right] \end{pmatrix}$$

$$(J_a^*(X))^{-1} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & \dots & \dots \\ 0 & \dots & \dots \end{pmatrix}$$

By Definition 3 with $\alpha_a = (-1, 0, 0)$, $\nu_a = (R_a^f, \beta_a, \lambda_a)$, the pseudo-outcome is given by:

$$\psi^{\mathrm{C}f\mathrm{RTE}}(Z, e, R, \beta, \lambda) = R_1^f(X;\delta) - R_0^f(X;\delta) + \frac{A - e(X)}{e(X)(1-e(X))}(m(Y, \beta_A, \lambda_A;\delta) - R_A^f(X;\delta))$$

We now apply Theorem 1. We first note that the binary matrices G and H are given by:

$$G = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad H = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

Moreover, $\|\widehat{\alpha}_{a,i}^{(k)} - \alpha_{a,i}^*\| = 0$ since the $\alpha_{a,i}$'s are constants. Therefore, the cross-products $\|\widehat{\alpha}_{a,i}^{(k)} - \alpha_{a,i}^*\|\|\widehat{\nu}_{a,j}^{(k)} - \nu_{a,j}^*\|$ vanish. Furthermore, given H, the only surviving terms in the cross-products $\|\widehat{\nu}_{a,i}^{(k)} - \nu_{a,i}^*\|\|\widehat{\nu}_{a,j}^{(k)} - \nu_{a,j}^*\|$ involve only β_a and λ_a. Thus, the oracle deviation is given by:

$$\mathcal{E} \leq \sum_{k=1}^{K} \sum_{a=0}^{1} \left(\|\widehat{R}_a^{f,(k)} - R_a^*\|\|\widehat{e}^{(k)} - e^*\| + \|\widehat{\beta}_a^{(k)} - \beta_a^*\|^2 + \|\widehat{\lambda}_a^{(k)} - \lambda_a^*\|^2 + \|\widehat{\beta}_a^{(k)} - \beta_a^*\|\|\widehat{\lambda}_a^{(k)} - \lambda_a^*\| \right)$$

□

We further provide a specific example of CfRTE for $f(x) = x \log x$ which corresponds to the Kullback–Leibler (KL) divergence. The associated risk is known as the entropic value-at-risk (EVaR). We term this conditional f-risk treatment effect the CKLRTE. The convex conjugate of f is given by $f^*(x^*) = e^{x^*-1}$. From the first-order optimality conditions, we obtain:

$$\beta_a^*(x;\delta) = \arg\inf_{\beta_a(x;\delta)\geq 0} \beta_a(x;\delta) \left(\log \mathbb{E}\left[e^{\frac{Y}{\beta_a(x;\delta)}} \middle| X = x, A = a \right] + \delta \right) \tag{17}$$

$$\begin{aligned} R^{KL}(x;\delta) &= \beta_a^*(x;\delta) \left(\log \mathbb{E}\left[e^{\frac{Y}{\beta_a^*(x;\delta)}} \middle| X = x, A = a \right] + \delta \right) \\ \lambda_a^*(x;\delta) &= \beta_a^*(x;\delta) \left(\log \mathbb{E}\left[e^{\frac{Y}{\beta_a^*(x;\delta)}} \middle| X = x, A = a \right] - 1 \right) \\ &= R^{KL}(x;\delta) - \beta_a^*(x;\delta)(\delta + 1) \end{aligned} \tag{18}$$

Thus, the optimization problem contains only one variable of interest, $\beta_a(x;\delta)$. The pseudo-outcome for CKLRTE at level δ is then given by:

$$\begin{aligned} \psi^{\text{CKLRTE}}(Z, e, R, \beta, \lambda) = R_1^{KL}(X;\delta) - R_0^{KL}(X;\delta) \\ + \frac{A - e(X)}{e(X)(1-e(X))} \left(\delta\beta_A(X;\delta) + \lambda_A(X;\delta) + \beta_A(X;\delta) e^{\frac{Y - \lambda_A(X;\delta)}{\beta_A(X;\delta)} - 1} - R_A^{KL}(X;\delta) \right) \end{aligned}$$

Remark 5. An interesting problem is the estimation of $\beta_a^*(X;\delta)$. We observe that the empirical analog of Eq. (17) requires fitting a large number of regression functions $\widehat{\mathbb{E}}_n\left[e^{\frac{Y}{\beta_a(x;\delta)}} \middle| X = x, A = a \right]$, one for each candidate $\beta_a(x;\delta)$. This can be mitigated by instead learning similarity weights between x and the training data and then replacing $\widehat{\mathbb{E}}_n\left[e^{\frac{Y}{\beta_a(x;\delta)}} \middle| X = x, A = a \right]$ with a weighted sum. We can then use this approximation with any convex optimization method since the argument of Eq. (17) is a convex function in β_a.

C ADDITIONAL EXPERIMENTAL RESULTS

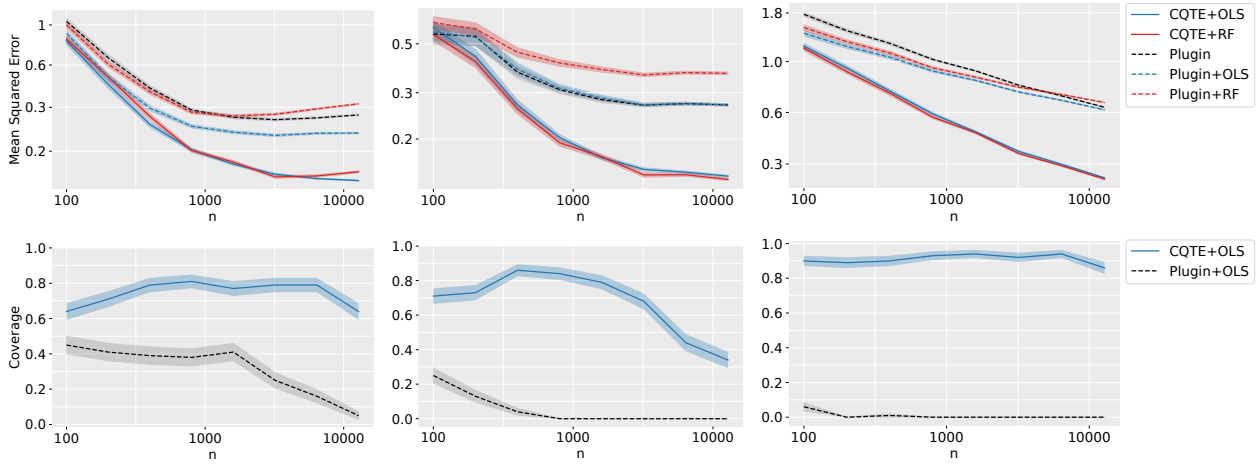
The replication code is distributed under an MIT license. The results in Section 8 were obtained using an Amazon Web Services instance with 32 vCPUs and 64 GiB of RAM.

C.1 Simulation Study

In this section, we provide additional results on simulated data for CQTEs (Section 3.1) and CfRTE (Section 3.3). Our analysis relies on the same data generating process and setup described in Section 8. For nuisance and second-stage estimation, we use Python's `scikit-learn` library (Pedregosa et al., 2011) for logistic and linear regression, random forest regression (RF) and linear quantile regression (LQR). We employ the extension `sklearn-quantile` for quantile random forest regression (QRF). We use the estimators' default hyperparameters except for the RFs where we set the minimum leaf size to $n/20$ to control overfitting. For each comparison, we run 100 simulations for each sample size $n = 100, 200, \ldots, 12800$ and evaluate the mean squared error (MSE) over a fixed set of 500 random X values.

CQTE Experiments. We measure the conditional quantile treatment effect at level $\tau = 0.75$. For the given DGP, the true CQTE at level τ is given by $q_1(X;\tau) - q_0(X;\tau) \simeq 1.14(e^{X_0+X_1} - e^{X_0})$, a heterogenous function of X. Learning CQTEs requires estimating three nuisances: the propensity score $e(X)$, the conditional quantile $q_a(X;\tau)$, as well as the density at the conditional quantile, $f_a(X)$.

We estimate $e(X)$ using logistic regression. Likewise, we learn $f_a(X)$ using the method described in Section 5.2 with a Gaussian kernel (bandwidth $b = 1$) for the estimates ω_i and a random forest (RF) regressor as the final estimator



(a) Flexible $\hat{q}$ (b) Misspecified $\hat{q}$ (c) Slow $\hat{q}$

Figure 4: Mean squared error (MSE) and confidence 95% interval coverage for different CQTE learners. Shaded regions depict plus/minus one standard error over 100 simulations.

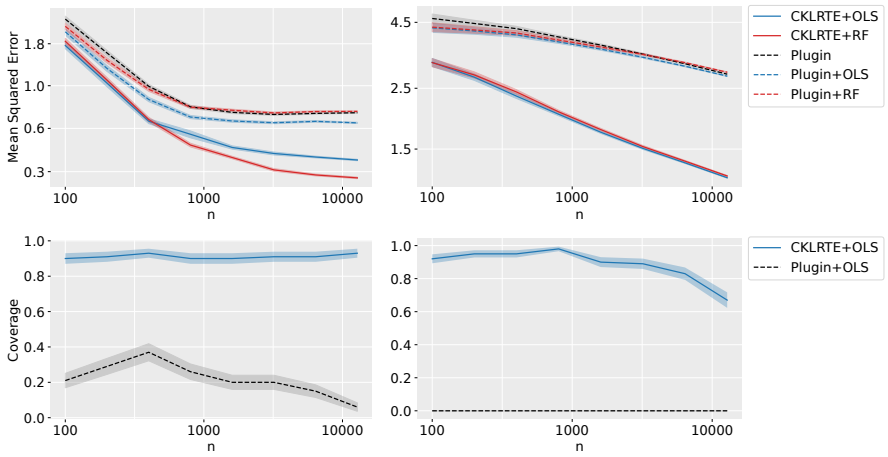
$\widehat{\mathbb{E}}_n[\omega \mid X = x, A = a]$. Finally, we consider three methods for estimating $q_a(X;\tau)$. First, we consider a quantile RF (QRF) (Meinshausen and Ridgeway, 2006) as the *flexible* learner. Then, we consider a "*misspecified*" model such as the linear quantile regressor (LQR) (Koenker, 2005). Lastly, we consider an estimator that uses a Gaussian kernel for calculating weights and then computes $\widehat{q}_a$ using the weighted version of the moment in Eq. (3). We choose the kernel bandwidth by Silverman's rule (Silverman, 2018). This will be our *slow* estimator as we expect it to suffer from the curse of dimensionality. For the final stage, we use an ordinary least squares model (CQTE+OLS) or a RF (CQTE+RF).

For the performance comparisons, we proceed similarly to Section 8. Thus, we compare the out-of-sample MSE of the CQTE estimator with that of the plug-in estimator from Eq. 7. We also construct Plugin+OLS and Plugin+RF given by running an additional OLS/RF model on the cross-fitted plug-in predictions. We do so to account for additional smoothing from the last stage regressor. When the second stage algorithm is OLS, we check whether the 95%-confidence interval OLS returns for the X_1 coefficient contains the coefficient from the true projection. The results are shown in Figure 4. Our CQTE learner provides uniformly better MSE performance that the plugin counterparts, and the results show this is not just a consequence of the second-stage regression. For inference, we achieve better coverage than the plug-in approaches. However, for the flexible and "misspecified" estimator, the coverage does not reach the desired level of 95%. This is most likely due to misspecifcation in the conditional density learner. *Note:* we put *misspecified* in quotes since the MSE of this estimator is better than that of the *flexible* estimator. We attribute this to the fact that for small t we can approximate $e^t \simeq 1 + t$ which makes the true CQTE close to a linear approximation.

CfRTE Experiments. We now turn to estimating the conditional f-risk treatment effects when we set the f-divergence to be the KL divergence. This f-risk measures the difference between conditional entropic values-at-risk (EVaRs). We call this risk treatment effect the CKLRTE (see Appendix B.3). We now wish to measure the CKLRTE at level $\tau = 0.75$ ($\delta = -\log(1-\tau)$). We note that our DGP does not admit EVaRs as the moment generating function of a lognormal distribution diverges. Thus, we modify the DGP to truncate any values above the 99$^{\text{th}}$ conditional quantile. For the truncated DGP, the true CKLRTE at level δ is given by $R_1^{KL}(X;\delta) - R_0^{KL}(X;\delta) \simeq 1.42(e^{X_0+X_1} - e^{X_0})$, a heterogenous function of X. Learning CKLRTEs requires estimating the following nuisances: the propensity score $e(X)$, the conditional EVaR $R_a^{KL}(X;\delta)$ and the $\beta_a(X;\delta)$ optimization parameter.

As before, we estimate $e(X)$ using logistic regression. We learn $R_a^{KL}(X;\delta)$ and the $\beta_a(X;\delta)$ optimization parameter jointly via the procedure described in Appendix B.3. Namely, we learn the weights necessary for computing a weighted average version of $\widehat{\mathbb{E}}_n\left[e^{\frac{Y}{\beta_a(x;\delta)}} \middle| X = x, A = a\right]$ and we then solve the convex optimization problem in Eq. (17). We use two models to calculate the weights: a random forest (*flexible* learner, most likely to capture complex functions) and a Gaussian kernel (*slow* learner, most likely suffers from the curse of dimensionality). We choose the Gaussian kernel bandwidth by Silverman's rule (Silverman, 2018). For the final stage, we use an OLS model (CKLRTE+OLS) or a RF (CKLRTE+RF).

Similar to our other experimental benchmarks, we compare the out-of-sample MSE of the CKLRTE estimator with that of the plug-in estimator from Eq. 7. We also construct Plugin+OLS and Plugin+RF given by running an additional OLS/RF model on the cross-fitted plug-in predictions. This will account for any additional smoothing from the last stage regressor.



(a) Flexible $\widehat{R}^{KL}$ (b) Slow $\widehat{R}^{KL}$

Figure 5: Mean squared error (MSE) and 95% confidence interval coverage for different CKLRTE learners. Shaded regions depict plus/minus one standard error over 100 simulations.

Table 2: Features of 401(k) dataset.

Name	Description	Type
age	age	continuous
inc	income	continuous
educ	years of completed education	continuous
fsize	familiy size	continuous
marr	marital status	binary
two_earn	whether dual-earning household	binary
db	defined benefit pension status	binary
pira	IRA participation	bianry
hown	home ownership	binary
e401	401 (k) eligibility	binary
net_tfa	net financial assets	continuous

When the second stage algorithm is OLS, we check whether the 95%-confidence interval for the X_1 coefficient contains the coefficient from the true projection. We display the results in Figure 5. Our CKLRTE learner provides uniformly better MSE performance than the plugin counterparts, regardless of the second stage regression. For inference, we achieve good coverage whereas plug-in approaches yield little to no coverage.

C.2 Impact of 401(k) Eligibility on Financial Wealth

We provide a detailed description the 401(k) dataset features in Table 2. We then compare the feature importances given by the random forest final stage of the CSQTE estimators and the DR-Learner. Figure 6 shows that the features driving the treatment effects for the three estimators have the same importance profile across tasks. For example, income, age and education are the most important features when determining the conditional average treatment effect, as well as the conditional average effects in the left 25% tail and right 25% tail.

D PRACTICAL CONSIDERATIONS FOR AND LIMITATIONS OF CDTE ESTIMATION

One of the limitations of our work is that the interpretation of our estimands as causal effects only hold when the unconfoundedness assumption holds. Whether it holds or does not, our estimands are differences in distributional statistics between the distributions of $Y \mid X, A = 1$ and $Y \mid X, A = 0$. If it does hold, this coincides with the differences in distributional statistics between the distributions of $Y(1) \mid X$ and $Y(0) \mid X$. While unconfoundedness can usually be guaranteed in experiments barring any non-compliance, there is no way to test for it in observational data. Thus, the onus is on the researcher to determine whether unconfoundedness is plausible in their data and to interpret the estimands carefully.

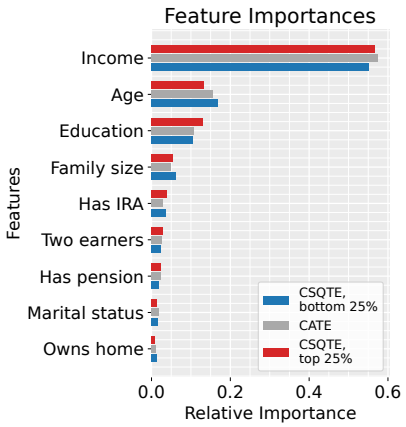


Figure 6: Feature importances given by the final stage algorithm for CSQTE on the bottom 25% financial asset holders, DR-Learner and CSQTE on the top 25% financial asset holders

Another possible concern in practice is the possibility for biased selection, where the sample may not be representative of the population of interest. To the extent that any unrepresentativeness is explained by X (known as, selection at random), the CDTEs given X will be unaffected. Therefore, we can alleviate this issue in the application of our method by learning CDTEs with a second-stage regression that is a universal approximator (like nonparametric regression, forests, or deep networks) so it learns the true CDTEs, thus adjusting for X fully. If there is unrepresentativeness not reflected in X (known as, selection *not* at random), then even the true CDTEs given $X = x$ only reflect effects on the data-generating subpopulation with $X = x$. This may be alleviated by considering richer X so as to minimize the discrepancy and/or by acknowledging possible biases. Finally, whether selection is at random or not, any best-in-class prediction guarantees (such as for simple second-stage regressions like OLS) only hold with respect to the data-generating population, and therefore "best-performance" may be unrepresentative of performance in the population of interest. If selection is at random, this is nonetheless easily fixable by reweighting.

Yet another practical concern is whether the outcome we observe truly reflects the appropriate notion of benefit/disbenefit. Otherwise, CDTEs similarly may not reflect the right risk measure, and our estimators could further propagate undesirable biases and choices encoded in the data (Passi and Barocas, 2019).

Lastly, there may be issues of calibrating risk profiles to human preferences, so as to make CDTEs informative for decision making. In particular, there may be conflicting risk preferences between the decision making entity (*e.g.*, doctor, policymaker, product manager) and the individual (*e.g.*, patient, citizen, platform user). If the risk measure employed is in conflict with individual preferences, it is possible to have a negative impact on individuals from their own perspective. Nonetheless, using the outputs of our models using different risk measures on an individualized basis could potentially be used to make policy decisions for individuals based on their own risk preferences.