Doubly-Robust Quantile Treatment Effect Estimators

Benjamin I. Miller

June 12, 2022

Abstract

I develop a doubly-robust estimator of the quantile treatment effect on the treated (QTT). I modify the Callaway and Li (2019) conditional estimator of the QTT to obtain consistent estimates using either the propensity score or the conditional cdf of the first-differenced untreated outcomes. Aside from the benefits of obtaining consistent estimates of a QTT when a nuisance function is misspecified, there are also efficiency gains. In addition, assumptions on the smoothness of the nuisance parameters can be relaxed when the estimator is doubly-robust. Finally, I demonstrate via simulations that the Callaway and Li estimator can produce estimates of the QTT that are not statistically different from 0 when the true values at particular quantiles are significantly different from 0. This problem is averted once the estimator is modified to be doubly-robust.

1 Introduction

There has been a recent reevaluation of the effectiveness of difference-in-difference estimators. Estimates of the average treatment effect on the treated (ATT) can be sensitive to nuisance functions that are based upon the probability of treatment given covariates, known as the propensity score, or the conditional mean of the difference in untreated outcomes before and after treatment for the untreated subpopulation. Identification of the ATT relies upon the parallel trends assumption

and the overlap assumption. A relaxation of these assumptions, which cannot be falsified, involves conditioning on covariates when that may change either the expected treatment status or the change in untreated outcomes. These separate approaches can be combined to obtain a doubly-robust estimator of the ATT.

If we are willing to strengthen these assumptions, then we can go further than estimation of the ATT. The QTT can be estimated at any desired quantile. How useful this is depends upon the topic under study. If a researcher is concerned with the median outcome of a variable post-intervention, or outcomes at the tail of a variable's distribution, then an estimator of the QTT would be desired. The assumptions that are necessary go beyond the parallel trends and overlap assumptions. The parallel trends assumption is strengthened from an assumption on conditional mean independence to an assumption of independence conditioned on covariates. More assumptions are required.

In this paper, I demonstrate that the strength of the assumptions from Callaway and Li (2019) purchase more than the authors may have realized. Using their assumptions, it is possible to generate a doubly-robust estimator of the QTT. This estimator then allows for a relaxation of the assumptions placed upon the nuisance functions. In this case, the nuisance functions are a propensity score and a conditional cdf of the difference in untreated outcomes from before and after treatment. The double-robustness property allows for a reduced rate of convergence of the both functions. If the nuisance functions are estimated nonparametrically, and depending upon the type of nonparametric estimator that is used, the double-robustness result has another beneficially property. Subject to minimal smoothness conditions on the nuisance functions, the rate of convergence itself may not matter **Note to committee: This can be proven by generalizing the results in Theorem x through considering a separate case where the rate of convergence for the propensity score is not derived in a separate theorem ,and the conditions of that theorem are relaxed. If you feel that this is necessary, I believe that I can make such a generalization**. Asymptotic normality and $\sqrt{n}$ consistency can still be achieved under minimal assumptions thanks to the double-robustness property.

First, in this paper I provide the identification result that guarantees double-robustness. This

result is an extension of the key identification result found in Callaway and Li (2019). The propensity score and the relevant conditional cdf can be misspecified, but not simultaneously. The properties of this estimator are studied. The properties are broken up into subcases, considering if the nuisance parameters are estimated nonparametrically or otherwise.

As a prerequisite to studying the properties of the QTT estimator, I derive the efficient influence function of the doubly-robust portion of my estimator. The portion that I am referring to is the cdf of the difference in untreated outcomes from the pre and post-treatment periods for the treated subpopulation at an arbitrary real number. It is upon the estimation of this parameter that the double-robustness result is applied. The efficient influence function is considered in the panel data setting.

With the efficient influence function in hand, I then demonstrate that the doubly-robust estimator achieves the semiparametric efficiency bound in the panel data setting. This is shown through two separate cases, where in the first case the nuisance functions are estimated parametrically. In the second case they are estimated nonparametrically, with the propensity estimated using the sieve logit estimator and the conditional cdf estimated using a kernel estimator. Both estimators could be chosen to be sieves or kernel estimator, and this would only change the restrictions on the smoothness of the nuisance functions.

Finally, I demonstrate through simulations that without the double-robustness correction the confidence intervals around the estimate of the QTT can be so large that the estimates are not statistically different from 0 at several quantiles. This erroneously makes the effect treatment effect seem insignificant around the median but significant at the tails of the outcome distribution for the treated subpopulation. When compared with an estimate of the ATT, this could make it seem that the treatment has no effect on a population except at high and low quantiles.

1.1 Literature Review

This paper draws directly from Callaway and Li (2019) in order to establish the identification result and the form of the QTT estimator, but the idea of a doubly-robust estimator in this context was

inspired by other papers in the treatment effects and missing data literature. Sant'Anna and Zhao (2020) examines the properties of the doubly-robust ATT estimator, and Callaway and Sant'Anna (2021) extends these results to the staggered treatment setting while also examining the asymptotic properties of various weighted ATT estimators. The doubly-robust ATT estimator has existed in the econometrics literature as an example of a doubly-robust estimator in Rothe and Firpo (2013), along with most of its properties in the single treatment period, panel data setting. The doubly-robust estimator combines the regression approach of Heckman, Ichimura, and Todd (1998) and Heckman et al. (1998), along with the propensity score matching approach of Abadie (2005), which itself is based upon Horvitz and Thompson (1952). My estimator is similar in that combines two approaches that are analogous to separate regression and propensity score approachesAll of these approaches take place within the difference-in-difference framework popularized by Card (1990) and Card and Krueger (1994).

The literature on quantile treatment effects, when considering either selection on observables or a panel data setting, prominently includes Firpo (2007), Athey and Imbens (2006), Bonhomme and Sauder (2011), and Chernozhukov et al. (2013). These approaches do not consider double-robustness, due to the M-estimation approaches within those papers. For example, Firpo (2007) sets up an M-estimation problem with weights based upon propensity score matching. The non-parametric logit sieve estimator of Hirano, Imbens, and Ridder (2003) is applied, but the conditions needed for asymptotic normality are considerably restrictive. A result similar to Firpo (2007) in the missing data setting is found in Wooldridge (2007).

The double-robustness properties of my estimator are based upon more than the aforementioned doubly-robust estimators of the ATT. A general weighting result for treatment effects is presented in Słoczyński and Wooldridge (2018), and the basis for constructing doubly-robust moments conditions is outlined in Chernozhukov et al. (2016). The latter work is closely related to the doubly-robust estimator in the missing data setting of Muris (2020). The construction of my double-robustness estimator is partly based upon the estimators in Sued, Valdora, and Yohai (2020) , though they consider a missing data setting and do not discuss the issue of inference. Rothe and

Firpo (2019) and Rothe and Firpo (2013) consider the asymptotic properties of double-robust estimators when the nuisance functions are estimated using kernel density methods. The work of Fan et al. (2016) is particularly important for this paper, since it considers a doubly-robust estimators with nuisance parameters that are estimated via sieves. I also make one final point about a doubly-robust QTT estimator that exists in the literature. Caracciolo and Furno (2017) proposes an estimator of the quantile treatment effect (QTE) that involves taking a quantile of a random variable that is a function of the propensity score and fitted values; however, this estimator only identifies the quantile of interest in a very restrictive case, using unstated assumptions. Their approach builds off of Machado and Mata (2005), but this approach involves obtaining unconditional quantiles directly through a random sample over conditional quantiles.

1.2 Structure of the Paper

The paper is structured as follows. Section 2 lays presents the framework, assumptions and identification result. Section 3 presents the estimator and considers estimation of the nuisance parameters. Section 4 considers the large-sample properties of the estimator. Section 5 contains a Monte Carlo study that examines the small sample properties of my estimator at various quantiles in comparison to other estimators. Section 6 concludes the paper. The mathematical proofs are contained in the appendix.

2 Assumptions and Background

2.1 Background

The setting that I am considering is the panel data setting. As in Callaway and Li (2019), I assume that the data consists of at least three periods, with treatment period t and pre-treatment periods $t-1$ and $t-2$. No unit receives the binary treatment before time t. $D = 1$ for unit i if treated at time t. D=0 if an individual is never treated. The outcomes Y_t, Y_{t-1}, and Y_{t-2} are observed, along

with covariates X.

Each unit i have the potential outcomes Y_{0t} and Y_{1t}, but these outcomes cannot be observed simultaneously for each unit i. The observed outcome Y_t is then expressed as

$$Y_t = DY_{1t} + (1 - D)Y_{0t}$$

Untreated outcomes are observed in the previous periods, since treatment does not take place until period t. Then $Y_{t-1} = Y_{0t-1}$ and $Y_{t-2} = Y_{0t-2}$.

Let q_τ denote the τ-quantile for some random variable Z, where

$$q_\tau = F_Z^{-1}(\tau) \coloneqq inf\{z : F_Z(z) \geq \tau\}$$

and $F_Z(z)$ is the cumulative distribution function (cdf) of Z. $F_{Y1t|D=1}$ and $F_{Y0t|D=1}$ denote the cdfs of the treated and untreated outcomes for the treated subpopulation, respectively. The QTT is then defined as,

$$QTT(\tau) = F^{-1}_{Y_{1t}|D=1}(\tau) - F^{-1}_{Y_{0t}|D=1}(\tau)$$

Interest in the QTT stems from the ability to identify the effect of an intervention on a treated group, compared to the counterfactual outcome. For example, suppose that a portion of a population receives a Covid-19 vaccine, with the outcome variable being gross income. While we may wish to know the treatment effect at the median ($\tau = 0.5$) for the entire population, the identification assumptions may too strong to identify such a parameter [1]. With weaker assumptions, we can identify QTT(0.5). That is to say, we can identify the median effect of Covid-19 vaccination on gross income for the treated subpopulation.

[1] In the case of the average treatment effect vs the average treatment effect on the treated, see Wooldridge (2010)

2.2 Assumptions

These assumptions are directly from Callaway and Li (2019). When presenting these assumptions, I will note how they are comparatively mild when compared to other assumptions in the econometrics literature. Note that these assumptions are the distributional equivalent of the traditional diff-in-diff assumptions.

The first assumption is known as the "Copula Stability Assumption"

Assumption ID.1 (Copula Stability Assumption)**.** $C_{\Delta Y_{0t}, Y_{0t-1}|D=1,X}(\cdot,\cdot) = C_{\Delta Y_{0t-1}, Y_{0t-2}|D=1,X}(\cdot,\cdot)$

This assumption is the most controversial assumption used for identification. As explained in Callaway and Li (2019), this assumption requires both the panel data setting and data over three time periods. This assumption is not placing any restrictions on any of the marginal distributions of the variables involved. Instead, by placing restrictions on the copula we are restricting the joint distribution of the variables in question based upon their joint distribution in prior periods. We cannot observe the untreated outcome for the treated subpopulation, but we can jointly observe the outcomes in the periods before treatment for the treated subpopulation. By an application of Skylar's Theorem and by writing $Y_{0t}|D = 1$ as $Y_{0t} - Y_{0t-1} + Y_{0t-1}|D = 1$, the cdf of $Y_{0t}|D = 1$ can be identified using the joint distribution of $Y_{0t-1}|D = 1$ and $Y_{0t-2}|D = 1$.

This assumption is similar to, and perhaps even weaker than, the assumption of stationarity in the time-series setting. No claim is being made that a sequence of random variables has a constant joint distribution over some shift in time. All that is being claimed is that a feature of the joint distribution, the joint dependence between the random variables, is fixed over a limited time period. A form of this assumption has been applied in the measurement error literature. In Cameron et al. (2004), the copula is used to model the difference in count variables, where each count variable represents different measurements of the same outcome.

The second assumption is directly exploited to obtain double-robustness of the estimator.

Assumption ID.2. $\Delta Y_{0t} \perp\!\!\!\perp D|X$

This assumption takes the parallel trends assumption of the difference-in-difference literature, $E[\Delta Y_{0t}|D = 1, X] = E[\Delta Y_{0t}|D = 0, X]$ and strengthens it over the entire distribution. The assumption states that the distribution of the untreated outcomes is unaffected by the treatment effect, conditional on the covariates. This assumption is necessary to obtain the estimator of $F_{\Delta Y_{0t}|D=1}(y)$. This assumption, as I will show, also makes the doubly-robust estimator of $F_{\Delta Y_{0t}|D=1}(y)$ the most efficient estimator in the panel data setting. Furthermore, as strong as this assumption is, it is weaker than the Strong Ignorability assumption of Rosenbaum and Rubin (1983), where $Y_{0t}, Y_{1t} \perp\!\!\!\perp D|X$, that is applied in Firpo (2007).

Now, let $\Delta Y_t = Y_t - Y_{t-1}$. Then, consider the following assumptions.

Assumption ID.3. Each of the random variables ΔY_t for the treated group and $\Delta Y_{t-1}, Y_{t-1}, Y_{t-2}$ for the treated group are continuously distributed on their support with densities that are uniformly bounded from above and bounded away from 0.

Assumption ID.4. The observed data $\{Y_{it}, Y_{it-1}, Y_{t-2}, X_i, D_i\}_{i=1}^n$ are independent and identically distributed draws from the joint distribution $F_{Y_t,Y_{t-1},Y_{t-2},X,D}$. In addition, $Y_{it} = D_i Y_{1it} + (1 - D_i)Y_{0it}$, $Y_{it-1} = Y_{0it-1}, Y_{it-2} = Y_{0it-2}$.

Assumption ID.3 ensures the uniqueness of the copula by restricting the outcomes to be continuous. Assumption ID.4 restricts the data to be panel data. If the copula is not unique, then even with the Copula Stability Assumption we may not be able to identify the cdf of $Y_{0t}|D = 1$.

The next assumption is the final assumption that I use for identification. Let the support of X be denoted by $\mathcal{X}$.

Assumption ID.5. $p := P(D = 1) > 0$ and, for all $x \in \mathcal{X}$, $p(x) := P(D = 1|X = x) < 1$.

The first part of the assumption ensures that there is some positive probability of treatment. The second assumption is the "overlap" assumption that is common in the difference-in-difference literature. This guarantees that for any value of X in $\mathcal{X}$ there is a positive probability of that value appearing in both the control and treatment groups. Without this assumption, the QTT cannot be identified since $F_{\Delta Y_{0t}|D=1}(y)$ could not be identified for a population that contains that contains

those values of X for which the assumption is violated. Note that the overlap assumption as stated here is not enough when the estimation of the propensity score is nonparametric. In that case, the propensity score requires sharp upper and lower bounds away from 0 and 1.

2.3 Identification

With the identification assumptions in hand, I now present the identification result.

Theorem 1. Under assumptions ID.1-ID.5, and assuming that $\pi(x) = p(x)$or $\tilde{P}(Y \leq y|X) = P(Y \leq y|X)$,

$$\begin{aligned} &F_{Y_{0t}|D=1}(y) \\ &= E\left[\mathbb{1}\{F^{-1}_{\Delta Y_{0t}|D=1}(F_{\Delta Y_{t-1}|D=1}(\Delta Y_{t-1})) \leq y - F^{-1}_{Y_{t-1}|D=1}(F_{Y_{t-2}|D=1}(Y_{t-2}))\}|D=1\right] \end{aligned}$$

where,

$$F_{\Delta Y_{0t}|D=1}(y) = E\left[\left(\frac{1-D}{p}\frac{\pi(x)}{1-\pi(x)}\right)\mathbb{1}\{\Delta Y_t \leq y\} - \left(\frac{1-D}{p}\frac{\pi(x)}{1-\pi(x)} - \frac{D}{p}\right)\tilde{P}(\Delta Y_{0t} \leq y|X)\right] \quad (1)$$

if $\pi(x) = p(x)$ a.c., or $\tilde{P}(\Delta Y_{0t} \leq y|X) = P(\Delta Y_{0t} \leq y|X)$ a.c. Then,

$$QTT(\tau) = F_{Y_{1t}|D=1}(\tau)^{-1} - F_{Y_{0t}|D=1}(\tau)^{-1}$$

The proof of the first part of this result is provided in Callaway and Li (2019). For the sake of making this paper as self-contained as possible, I will outline their argument. First, note that since $E[\mathbb{1}_{Y_{0t}|D=1 \leq y}] = E[\mathbb{1}_{Y_{0t}|D=1 - Y_{0t-1}|D=1 + Y_{0t-1}|D=1 \leq y}] = E[\mathbb{1}_{\Delta Y_{0t}|D=1 + Y_{0t-1}|D=1}]$. Since this expectation is over the joint distribution of the random variables $\Delta Y_{0t}|D=1$ and $Y_{0t-1}|D=1$, and since this joint distribution can be written in terms of the copula and the marginal distributions of Y_{0t-1} and Y_{0t-2} by assumption ID.1 and Sklar's Theorem. The result then follows from a change of variables.

The second part of the theorem is the basis for the double-robustness property of the estimator. Either the propensity score or the conditional cdf of ΔY_{0t} needs to be correctly specified so

$F_{\Delta Y_{0t}|D=1}(\delta)$ will be correctly identified. $F_{\Delta Y_{0t}|D=1}(\delta)$ is needed to identify $F_{Y_{0t}|D=1}(y)$. The intuition behind the double-robustness result is that if the propensity score is correctly specified, then the information provided by conditional cdf of ΔY_{0t} becomes redundant, at least for identification. If $\tilde{P}(\Delta Y_{0t} \leq y|X)$ is correctly specified, then the weight that is applied to this conditional cdf filters out the incorrect information that is left over from misspecification of the propensity score.

3 Estimation

This section will present the estimators that can be used to obtain the QTT under the identification assumptions. There are two different estimators that I present, with an asymptotically negligible difference. These estimators differ in that they calculate the weights differently for estimation of $F_{\Delta Y_{0t}|D=1}(y)$. The first estimator is,

$$\hat{QTT}(\tau) = \hat{F}_{Y_{1t}|D=1}(\tau)^{-1} - \hat{F}_{Y_{0t}|D=1}(\tau)^{-1}$$

where

$$\hat{F}^{-1}_{Y_{1t}|D=1}(\tau) = inf\{y : \hat{F}_{Y_{1t}|D=1}(y) \geq \tau\}$$
$$\hat{F}^{-1}_{Y_{0t}|D=1}(\tau) = inf\{y : \hat{F}_{Y_{0t}|D=1}(y) \geq \tau\}$$

$$\begin{aligned}&\hat{F}_{Y_{0t}|D=1}(y)\\ &= n_{\mathcal{D}}^{-1} \sum_{i\in\mathcal{D}} [\mathbb{1}\{\hat{F}^{-1}_{\Delta Y_{0t}|D=1}(\hat{F}_{\Delta Y_{t-1}|D=1}(\Delta Y_{t-1})) \leq y - \hat{F}^{-1}_{Y_{t-1}|D=1}(\hat{F}_{Y_{t-2}|D=1}(Y_{t-2}))]\end{aligned}$$

where $n_{\mathcal{D}}$ denotes the number of treated observations and

$$\hat{F}_{\Delta Y_{0t}|D=1}(y) = n^{-1} \sum_{i=1}^{n} \left[\left(\frac{1-D_i}{\frac{\sum_{k=1}^{n} D_k}{n}} \frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)} \right) \mathbb{1}\{\Delta Y_t \leq y\} - \left(\frac{1-D_i}{\frac{\sum_{k=1}^{n} D_k}{n}} \frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)} - \frac{D_i}{\frac{\sum_{k=1}^{n} D_k}{n}} \right) \hat{P}(\Delta Y_{0t} \leq y|X) \right] \tag{2}$$

An alternative estimator of $F_{\Delta Y_{0t}|D=1}(y)$ is,

$$\hat{F}_{\Delta Y_{0t}|D=1}(y) = n^{-1}\sum_{i=1}^{n}\left[\left(\frac{1-D_i}{l_0(\pi(x_i))}\frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)}\right)\mathbb{1}\{\Delta Y_t \leq y\} - \left(\frac{1-D_i}{l_0(\pi(x_i))}\frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)} - \frac{D_i}{\frac{\sum_{k=1}^{n} D_k}{n}}\right)\hat{P}(\Delta Y_{0t} \leq y|X)\right] \tag{3}$$

where $l_0(\pi(x)) = \sum_{i=1}^{n}\frac{\pi(x_i)(1-D_i)}{1-\pi(x_i)}$. The interesting point here is the estimation of the nuisance functions. If the nuisance functions are estimated parametrically, then by standard assumptions in the appendix the nuisance functions will converge rapidly enough in probability to guarantee asymptotic normality, since the parameters that index the functions will converge at a sufficiently fast rate. The issue here is that it is unlikely for the nuisance functions to be correctly specified.

If the nuisance functions are estimated nonparametrically, then estimation depends upon exactly how they are estimated. For example, suppose that both nuisance functions are estimated using kernel-based methods. The advantage of this, as seen in Rothe and Firpo (2013), is that the kernel estimation permits the estimator to be decomposed into a bias term, a first-order stochastic term, and a second-order stochastic term. Depending upon how the bandwidth is chosen, the estimator can converge in probability at a fast enough rate to ensure asymptotic normality, but at a slower rate than would otherwise be necessary due to the double-robustness property.

Note to committee: The next paragraph may be inserted if necessary. If the nuisance functions are estimated using a sieve approach, then the rate of convergence can be even slower than is necessary when compared to a kernel approach. This is for two reasons. First, this is due to the double-robustness property. Second, the search for the solution to the sieve estimation problem is taking place over a space of sufficiently well-behaved functions so that as long as the sieve estimator is consistent, the rate of convergence does not matter on this space.

The double-robustness property here does not only mean that the identifying moment is doubly-robust. An implication of this is that the higher-order derivatives are also doubly-robust. This implies that like the identifying moment, the derivatives also equal 0 if at least one of the nuisance parameters is correctly specified. This is useful for both sieve and kernel estimation. In

the case of kernel estimation, this implies that the bias term in the kernel decomposition equals 0. In the case of sieve estimation, the usefulness of this is that terms in the asymptotic expansion of the doubly-robust estimator can then be bounded by the bracket integral with respect to the L_2 norm over the function space.

4 Asymptotic Properties

The key asymptotic properties of the estimator revolve around the asymptotic behavior of the estimator of $F_{\Delta Y_{0t}|D=1}(y)$, in addition to the asymptotic behavior of $\pi(x)$ and $\hat{P}(\Delta Y_{0t} \leq y|X)$. The limiting behavior of the QTT estimator is unchanged by the doubly-robust estimator. Before the asymptotic behavior of the estimator can be discussed, the following assumption will be introduced:

Assumption NP.1.

$$\sup_{x\in\mathcal{X}}|\hat{\pi}(X) - \pi(X)| = o_p(n^{-1/4})$$

$$\sup_{x\in\mathcal{X}}||\hat{P}(\Delta Y_{0t} \leq \delta|x) - P(\Delta Y_{0t} \leq \delta|x)|| = o_p(n^{-1/4})$$

This is the conventional uniform convergence assumption in the literature on nonparametric rates of convergence. It is not necessary when the estimator is doubly robust, but it is sufficient[2]. For parametric estimation of the nuisance functions, I will make the following assumptions,

Assumption P.1. (i)$G(x;\gamma)$ is a parametric model for $p(x)$, where $\gamma \in \Gamma \subset \mathbb{R}^M$ and $G(x,\gamma) > 0$, all $x \in X, \gamma \in \Gamma$, where Γ is compact. (ii) There exists $\gamma_0 \in \Gamma$ such that $p(x) = G(x,\gamma_0), \gamma_0 \in int(\Gamma)$. (iii)$G(X;\gamma)$ is a.s. twice continuously differentiable in a neighborhood of $\gamma_0, \Gamma^* \subset \Gamma$. (iv)$\hat{\gamma}$ is a consistent estimator of γ_0 and $n^{1/2}(\hat{\gamma} - \gamma_0) = n^{-1/2}\sum_{i=1}^n l_\gamma(W_i;\gamma_0) + o_p(1)$, where $W_i = (Y_{01}, Y_{i1}, D_i, X_i)$, $E[l_\gamma(W_i;\gamma_0)] = 0$, $E[l_\gamma(W_i;\gamma_0)l_\gamma(W_i;\gamma_0)']$ exists and is positive definite and $\lim_{\delta\to 0}$ $\mathrm{E}[\sup_{\gamma\in\Gamma^*,||\gamma-\gamma^*||\leq\delta}||l_\gamma(W_i;\gamma) - l_\gamma(W_i;\gamma_0)||^2]$ =0. (vi) For some $\epsilon > 0, 0 < P(x;\gamma) \leq 1-\epsilon$ a.s. for all $\gamma \in int(\Gamma)$.

[2]This assumption is mentioned in a footnote of Callaway and Sant'Anna (2021) when the nuisance parameters are estimated nonparametrically, even though their estimator is doubly-robust. It is not necessary.

Assumption P.2. (i)$g(x) = g(x;\beta)$ is a parametric model for the conditional mean of Y_{0t}, where $\beta \in \Theta \subset \mathbb{R}^k, \Theta$ being compact; (ii) $g(X,\beta)$ is a.c. continuous at each $\beta \in \Theta$; (iii) there exists a unique pseudo-true parameter $\beta^* \in int(\Theta)$;(iv)$g(X,\beta)$ is a.c. twice continuously differentiable in a neighborhood of $\beta^*, \Theta^* \subset \Theta$; (iv) the estimator $\hat{\beta}$ is strongly consistent for the β^* and satisfies the following linear expansion:

$$\sqrt{n}(\hat{\beta} - \beta) = n^{-1/2} \sum_{i=1}^{n} l_\beta(W_i;\beta^*) + o_p(1)$$

where $l_\beta(\cdot;\beta)$ is such that $E[l_g(W;\beta^*)] = 0, E[l_g(W;\beta^*)l_g(W;\beta^*)']$ exists and is positive definite and $lim_{a\to 0} E\left[sup_{\beta\in\Theta^*,\|\beta-\beta^*\leq a}\|l_\beta(W;\beta) - l_\beta(W;\beta^*)\|\right] = 0$.

Assumption P.3. $E[\|h(W;\beta,\gamma)\|^2] < \infty$ and $E\left[sup_{\beta\in\Theta_s,\gamma\in\Gamma_s}|\dot{h}(W;\beta,\gamma)\right] < \infty$ where Θ^s, Γ_s is a small neighborhood of β^*, γ^*, and

$$h(W;\beta,\gamma) = (w_0(D,X;\gamma))\mathbb{1}\{\Delta Y_{ti} \leq \delta\} - (w_0(D,X;\gamma) - w_1(D))P(\Delta Y_{0t} \leq \delta, X;\beta)$$

These are the standard assumptions found in the literature, such as in Sant'Anna and Zhao (2020). Assumptions P.2 and P.2 imply that the parameters which index $p(x;\gamma)$ and $P(\Delta Y_{0t} \leq \delta|x;\beta)$ are sufficiently smooth and are $\sqrt{n}$-asymptotically linear. Assumption P.3 is an integrability condition. Assumptions P.1 and P.2 are stronger than Assumption NP, while Assumption P.3 is necessary to apply the Weak Law of Large Numbers along with the Central Limit Theorem.

Before I establish the asymptotic properties of the estimator, the efficient influence function needs to be found. Besides claiming that the estimator of $F_{\Delta Y_{0t}|D=1}(y)$ is doubly-robust, the efficient influence function will allow us to determine whether the estimator is the most efficient estimator of $F_{\Delta Y_{0t}|D=1}(y)$. Note that this assumption of efficiency is only being made under the identification assumptions. If these assumptions do not hold, then the efficiency result fails. The efficient influence function should also be the identifying moment condition to estimate $F_{\Delta Y_{0t}|D=1}(y)$ at a fixed value y. When the nuisance functions are nonparametrically estimated, asymptotic normality will depend upon a Taylor expansion of the efficient influence functions.

The efficient influence functions is presented in the following theorem:

Theorem 2. Under assumptions ID.1-ID.5, the efficient influence function is,

$$\frac{(1-D)p(x)}{p(1-p(x))}\mathbb{1}_{\{\Delta Y_3 \leq \delta\}} - \frac{(1-D)p(x)}{p(1-p(x))}P(\Delta Y_{0t} \leq \delta|X)$$
$$+\frac{D}{p}P(\Delta Y_{0t} \leq \delta|X) - \frac{D}{p}F_{\Delta Y_{0t}|D=1}(\delta)$$

Note that if we were to take the expected value of this function, $\frac{E[D]}{p} = 1$, so in expectation the efficient influence function reduces to the identifying moment condition of $F_{\Delta Y_{0t}|D=1}(y)$ induced by $\hat{F}_{\Delta Y_{0t}|D=1}(y)$. With the efficient influence function in hand, we can proceed to describe the asymptotic behavior of $\hat{F}_{\Delta Y_{0t}|D=1}(y)$. Before that takes place, it should be noted how exactly each of the nuisance functions are estimated. I consider as a nonparametric estimator of the propensity score the sieve logit estimator of Hirano, Imbens, and Ridder (2003), though the proof and assumptions that I am placing on that estimator are different, and in some sense relaxed, compared to the conditions in Hirano, Imbens, and Ridder (2003) that are used to prove that the estimator converges uniformly to the true function at $o_p(n^{-1/4})$. When using this estimator, the goal is to approximate $p(\mathbf{x})$, the log odds ratio, using a series approximation such that

$$p(\mathbf{x}) \approx \tilde{r}^{\tilde{K}}(\mathbf{x})'\Gamma_{\tilde{K}}$$

where

$$\tilde{r}^{\tilde{K}}(\mathbf{x})' = (r_{1\tilde{K}}(\mathbf{x}), ..., r_{\tilde{K}\tilde{K}}(\mathbf{x}))'$$
$$K = \tilde{K} + 1$$

The estimator is given by

$$\hat{\pi^*} = argmax_{\pi \in \Pi_n} \frac{1}{n}\sum_{i=1}^{n}[D_i log L(\pi(\mathbf{x}_i)) + (1-D_i)log(1 - L(\pi^*(\mathbf{x}_i))]$$

where $L(a) = \frac{exp(a)}{1+exp(a)}$ and Π_n denotes the sieve space. Let the sieve space be over the p-smooth class of functions which I will denote by,

$$\Pi = \Lambda_c^p(\mathcal{X}) = \left\{h \in C^m(\mathcal{X}) : \sup_{[\alpha]\leq m}\sup_{x\in\mathcal{X}}|D^\alpha h(x)| \leq c, \sup_{[\alpha]=m}\sup_{x,y\in\mathcal{X},x\neq y}\frac{D^\alpha h(x) - D^\alpha h(y)}{|x-y|_e^\gamma} \leq c\right\}$$

where $C^m(\mathcal{X})$ denotes the space of all m-times continuously differentiable functions on $\mathcal{X}$, and $|\cdot|$ denotes the Euclidean norm. Furthermore, let $\mathcal{H}_n = \left\{h \in \Pi_n : h(x^*) = 0, |h|_s \leq c\right\}$, and $\Pi_n = \text{Pol}(k_n)$ where $\text{Pol}(k_n) = \left\{\sum_{k=0}^{k_n} a_k x^k, x \in \mathcal{X} : a_k \in \mathbb{R}\right\}$.

I will let $||\hat{\pi} - p||_s = \sup_{\mathbf{x}\in\mathcal{X}}|\hat{\pi}(\mathbf{x}) - p(\mathbf{x})|$ and $\ell(\pi, \mathbf{x}_i) = D_i logL(\pi^*(\mathbf{x}_i)) + (1 - D_i)log(1 - L(\pi(\mathbf{x}_i))$. Let $H(w, \mathcal{F}_n, ||\cdot||_r)) \coloneqq log(N(w, \mathcal{F}_n, ||\cdot||_r))$, where $N(w, \mathcal{F}_n, ||\cdot||_r)$ is the minimal number of w-balls that cover $\mathcal{F}_n$ under $||\cdot||_r$, and

$\mathcal{F}_n = \{\ell(\pi^*, \mathbf{x}_i) - \ell(p, \mathbf{x}_i) : ||\pi - p|| \leq \delta, \pi \in \Pi_n\}$. In addition,

$\delta_n = \inf\left\{\delta \in (0,1) : \frac{1}{\sqrt{n}\delta^2}\int_{b\delta^2}^{\delta}\sqrt{H(w, \mathcal{F}_n, ||\cdot||_r))}dw \leq const.\right\}$, where $b > 0$ is a constant. The assumptions below are sufficient for the consistency of the sieve logit estimator and to satisfy Assumption NP.1. They are based upon conditions in Chen (2007):

Assumption NP.2. (i)$E[D(logL(p(\mathbf{x}))) + (1 - D)log(1 - L(p(\mathbf{x}))] > -\infty$, and

if $E[D(logL(p(\mathbf{x}))) + (1 - D)log(1 - L(p(\mathbf{x}))] = \infty$

then $E[D(logL(p(\mathbf{x}))) + (1 - D)log(1 - L(p(\mathbf{x}))] < \infty$ for all $\pi \in \Pi_K \backslash p$ for all $k \geq 1$

(ii)There are functions d() and t(), where d() is a non-increasing positive function and t() is a positive function such that for all $\epsilon > 0$ and for all $k \geq 1$,

$E[D(logL(p(\mathbf{x}))) + (1 - D)log(1 - L(p(\mathbf{x}))]$

- $sup_{\pi\in\Pi_n:||\pi-p||_s\geq\epsilon}E[D(logL(\pi(\mathbf{x}))) + (1 - D)log(1 - L(\pi(\mathbf{x}))]$

$\geq d(k)t(\epsilon) > 0$

Assumption NP.3. $\Pi_k \subset \Pi_{k+1} \subset \Pi$ for all $k \geq 1$, and there exists a sequence $\sigma_k p \in \Pi_k$ such that $||\sigma_k p - p||_s \to 0$ as $k \to \infty$.

Assumption NP.4. (i)The sieve spaces Π_k are compact under $||\pi_1 - \pi_2||_s$, where $\pi_1, \pi_2 \in \Pi_k$

(ii)$\lim inf_{k(n)} d(k(n)) > 0$, $E[\ell(\pi, \mathbf{x}_i]$ is continuous at $\pi = p \in \Pi$, and $E[sup_{\pi \in \Pi_n} |\ell(\pi, \mathbf{x}_i)]$ is bounded.
(iii)$E[||\mathbf{x}_i||] < \infty$

Assumption NP.5. $log(N(\delta, \Pi_n, ||\cdot||)) = o(n)$ for all $\delta > 0$.

Assumption NP.6. There exist $\underline{p}$ and $\overline{p}$ such that $0 < \underline{p} \leq p(\mathbf{x}) \leq \overline{p} < 1$.

Assumption NP.7. $\frac{2(p)^2}{(2p+d)^2} > \frac{1}{4}$, where d is the dimension of X, and $p = (s + \alpha)$, where $p(\mathbf{x})$ is s times continuously differentiable and $|p(\mathbf{x}) - p(\mathbf{y})| \leq ||\mathbf{x} - \mathbf{y}||_e^{\alpha}, \mathbf{x}, \mathbf{y} \in \mathcal{X}$ under the Euclidean norm $||\cdot||_e$ for $0 < \alpha \leq 1$.

Assumptions NP.2 consists of regularity conditions on the objective function. Assumption NP.3 implies that there exists some sequence of functions such that on subsets of the entire function space there exists some sequence of function that uniformly converge to 0 as the subspaces grow in size. Assumption NP.4 implies the existence of a solution at which the objective function is minimized. Assumption NP.5 ensures that the function space does not grow too fast as the sample size increases. Assumption NP.6 strengthens Assumption ID.5 so that the propensity score has upper and lower bounds away from 0 and 1. This is necessary so that the log odds ratio is finite for all $x \in \mathcal{X}$. Assumption NP.7 is a restriction on the differentiability and smoothness of the propensity score relative to the dimension of $\mathbf{x}$. This is a weakening of the smoothness assumption in Hirano, Imbens, and Ridder (2003).

Using the previous assumptions, I obtain the following result:

Theorem 3. Under assumptions NP.1-NP.6, $||\pi_n^* - p^*||_s \xrightarrow{p} 0$, and $||\pi_n^* - p^*||_s = o_p(n^{-1/4})$.

The next theorem concerns the asymptotic behavior of the estimator of $\hat{P}(\Delta Y_{0t} \leq y|X)$. This estimator is a kernel density estimator, though a sieve estimator could also be chosen. The estimator that I have chosen is based upon the estimator of Li and Racine (2008). The assumptions that are needed include, (from Li and Racine (2008))

Assumption C.1. Both $\mu(x)$ and $F(y|x)$ have continuous second-order partial derivatives with respect to x^c, where x^c denotes the vector of continuous random variables. For fixed values of y and x, $\mu(x) > 0, 0 < F(y|x) < 1$

Assumption C.2. $w(\cdot)$ is a symmetric, bounded, and compactly supported density function, and $w(\cdot)$ is a Lipschitz function on the compact set D.

Assumption C.3. As $n\to\infty, h_s\to\infty$ for $s = 1..., q, \lambda_s\to 0$ for $s = 1, ..., r$, and $(nh_1...h_q)\to\infty$, and as $n\to\infty, h_0\to 0$.

Assumption C.4. $F(y|x)$ is twice continuously differentiable in (y, x^c).

Let $|h| = \sum_{i=1}^{q} h_s$, $|\lambda| = \sum_{i=1}^{r} \lambda_s$, where $0 \leq \lambda_s \leq 1$ and $W_h(X_i^c, x_i^c) = \prod_{s=1}^{q} h_s^{-1} w((X_{is}^c - x_s^c)/h_s)$. Let $\hat{\mu}(x) = n^{-1} \sum_{i=1}^{n} W_h(X_i^c, x_i^c)$ Let $\tilde{F}(y|x^c) = \frac{n^{-1} \sum_{i=1}^{n} G(\frac{y-Y_i}{h_0}) W_h(X_i^c, x_i^c)}{\hat{\mu}(x)}$. $G(\cdot)$ is the distribution function with corresponding density function $w(\cdot)$. h_s is the bandwidth associated with the continuous variable x_s^c, and h_0 is the bandwidth associated with Y_i[3].

We then have the following theorem:

Theorem 4. Suppose Assumptions C.1-C.4 hold and $h_0 = h$. Then, $sup_{x\in D}|\tilde{F}(y|x^c) - F(y|x^c)| = O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(h^2)$.

The restriction that $h_0 = h$ can be used to achieve the rate of convergence of $o_p(n^{-1/4})$, but this is not the optimal rate of convergence that minimizes the cross-validation objective function. In fact, the optimal choice of bandwidth as given in Li and Racine (2007) will not achieve this rate of convergence, though this rate can still be forced through the choice of bandwidth. Now that I have two nonparametric estimators of the nuisance functions, I can proceed to the first major distributional result. I will now proceed proving the consistency and asymptotic normality of $\hat{F}_{\Delta Y_{0t}|D=1}(\delta)$.

Theorem 5. Suppose that assumptions ID.1-ID.5, and NP.1, or P.1-P.3 hold. Then $\hat{F}_{\Delta Y_{0t}|D=1}(\delta) \xrightarrow{p} F_{\Delta Y_{0t}|D=1}(\delta)$, and

[3]In principle, the estimator here could be the estimator of Li and Racine (2008), where the covariates can be discrete and ordered; however, in order to cite particular theorems from Rothe and Firpo (2013) I am only considering an estimator that allows for continuous covariates, though as noted by Rothe and Firpo (2019), the results could be modified to allow for discrete covariates.

$\sqrt{n}\left(\hat{F}_{\Delta Y_{0t}|D=1}(\delta) - F_{\Delta Y_{0t}|D=1}(\delta)\right) \xrightarrow{d} N(0, E[\psi(D_i, p(x_i), P(\Delta Y_{0t} \leq \delta|X_i), \hat{p})]^2)$, where

$$\psi(D_i, p(x_i), P(\Delta Y_{0t} \leq \delta|X_i), \hat{p}) = \left[w_0(D_i, X_i; \gamma^*)(\mathbb{1}\{\Delta Y_{ti} \leq \delta\} - (w_0(D_i, X_i; \gamma^*) - w_1(D_i, X_i; \gamma^*))P(\Delta Y_{0ti} \leq \delta|X_I; \beta^*) - w_1(D_i)F_{\Delta Y_{0t}|D=1}(\delta)\right]$$

and

$$w_0(D, X; \gamma^*) = \left(\frac{1 - D}{\frac{\sum_{k=1}^n D_k}{n}} \frac{\pi(\mathbf{x}; \gamma^*)}{1 - \pi(\mathbf{x}; \gamma^*)}\right)$$

$$w_1(D) = \frac{D}{\frac{\sum_{k=1}^n D_k}{n}}$$

Note that the result shows that the estimator attains the semiparametric efficient lower bound, since the asymptotic variance equals the second moment of the efficient influence function of $F_{\Delta Y_{0t}|D=1}(\delta)$.

Now, I will present a central limit theorem result, based upon a similar result in Callaway and Li (2019), which I will use to later establish the limiting behavior of the QTT estimator. Note the following result:

Theorem 6. Suppose assumptions ID.1-ID.5, and Assumption NP.1 or assumption P.1-P.3 hold. Then,

$$\left(\hat{G}_{\Delta Y_{0t}|D=1}, \hat{G}_{\Delta Y_{t-1}|D=1}, \hat{G}_{Y_{0t}|D=1}, \hat{G}_{Y_t|D=1}, \hat{G}_{Y_{t-1}|D=1}, \hat{G}_{Y_{t-2}|D=1}\right)$$
$$\xrightarrow{d} (\mathbb{W}_1, \mathbb{W}_2, \mathbb{V}_0, \mathbb{V}_1, \mathbb{W}_3, \mathbb{W}_4)$$

In the space $S = l^\infty(\Delta\mathcal{Y}_{0t|D=1}) \times l^\infty(\Delta\mathcal{Y}_{t-1|D=1}) \times l^\infty(\mathcal{Y}_{0t|D=1}) \times l^\infty(\mathcal{Y}_{t|D=1}) \times l^\infty(\mathcal{Y}_{t-1|D=1}) \times l^\infty(\mathcal{Y}_{t-2|D=1})$ where $(\mathbb{W}_1, \mathbb{W}_2, \mathbb{V}_0, \mathbb{V}_1, \mathbb{W}_3, \mathbb{W}_4)$ is a tight Gaussian process with mean 0 and covariance $V(y', y) =$

$E[\eta(y)'\eta(y)]$ for $y = (y_1, y_2, y_3, y_4, y_5, y_6) \in S$ and with $\eta(y)$ given by

$$\eta(y) = \begin{bmatrix} \psi(D, X, Y_0, Y_1) \\ \frac{D}{p}\mathbb{1}\{\Delta Y_{t-1} \leq y_2\} - F_{\Delta Y_{t-1}|D=1}(y_2) \\ \frac{D}{p}\mathbb{1}\{\tilde{Y}_t \leq y_3\} - F_{Y_{0t}|D=1}(y_3) \\ \frac{D}{p}\mathbb{1}\{Y_t \leq y_4\} - F_{Y_t|D=1}(y_4) \\ \frac{D}{p}\mathbb{1}\{Y_{t-1} \leq y_5\} - F_{Y_{t-1}|D=1}(y_5) \\ \frac{D}{p}\mathbb{1}\{Y_{t-2} \leq y_6\} - F_{Y_{t-2}|D=1}(y_6) \end{bmatrix}$$

where

$$\begin{aligned}
\hat{G}_{\Delta Y_{0t}|D=1}(\delta) &= \sqrt{n}\left(\hat{F}_{\Delta Y_{0t}|D=1}(\delta) - F_{\Delta Y_{0t}|D=1}(\delta)\right) \\
\tilde{Y}_{it} &= F^{-1}_{\Delta Y_{0t}|D=1}\left(F_{\Delta Y_{t-1}|D=1}(\Delta Y_{it-1})\right) + F^{-1}_{\Delta Y_{t-1}|D=1}\left(F_{\Delta Y_{t-2}|D=1}(\Delta Y_{it-2})\right) \\
\tilde{F}_{Y_{0t}|D=1}(y) &= \frac{1}{n_D}\sum_{i\in\mathcal{D}} \mathbb{1}\{\tilde{Y}_{it} \leq y\} \\
\tilde{G}_{Y_{0t}|D=1}(y) &= \sqrt{n}\left(\tilde{F}_{Y_{0t}|D=1}(y) - F_{Y_{0t}|D=1}(y)\right)
\end{aligned}$$

Proposition SA2 and Theorem SA1 still hold from Callaway and Li (2019). I reproduce them here, in order to account for the change in notation and numbering of the assumptions. Proposition 1 is used to establish the result in Theorem 7.

Proposition 1. Let $\hat{G}_0(y) = \sqrt{n}(\hat{F}_{Y_{0t}|D=1}(y)) - F_{Y_{0t}|D=1}(y))$ and let $\hat{G}_1(y) = \sqrt{n}(\hat{F}_{Y_t|D=1}(y)) - F_{Y_t|D=1}(y))$. Suppose assumptions ID.1-ID.5, Assumption NP.1 or assumption P.1-P.3 hold. Then,

$$(\hat{G}_0, \hat{G}_1) \xrightarrow{d} (\mathbb{G}_0, \mathbb{G}_1)$$

where $\mathbb{G}_0$ and $\mathbb{G}_1$ are tight Gaussian processes with mean 0 with almost surely uniformly continuous paths on the space $\mathcal{Y}_{0t|D=1} \times \mathcal{Y}_{t|D=1}$ given by

$$\mathbb{G}_1 = \mathbb{V}_1$$

and

$$\mathbb{G}_0 = \mathbb{V}_0 + \int \left[\mathbb{W}_1 \circ K_2(y,v) - f_{\Delta Y_{0t}|D=1}\left(y - F^{-1}_{Y_{t-1}|D=1} \circ F_{Y_{t-2}|D=1} \circ \frac{\mathbb{W}_4 - \mathbb{W}_1 \circ K_1(v)}{f_{Y_{t-1}|D=1} \circ K_1(v)} - \mathbb{W}_2 \circ K_3(y,v) \right) \right] \times \frac{f_{\Delta Y_{t-1}|Yt-2,D=1}(K_3(y,v)|v)}{f_{\Delta Y_{t-1}|D=1}(K_3(y,v))} dF_{Y_{t-2}|D=1}(v),$$

where $K_1(v) \coloneqq F^{-1}_{Y_{t-1}|D=1} \circ F_{Y_{t-2}|D=1}(v)$, $K_2(y,v) \coloneqq y - K_1(v)$,and $K_3(y,v) \coloneqq F^{-1}_{\Delta Y_{t-1}|D=1} \circ F_{\Delta Y_{0t}|D=1}(K_2(y,v))$

Theorem 7. Suppose $F_{Y_{0t}|D=1}$ admits a positive continuous density $f_{Y_{0t}|D=1}$ on an interval $[a,b]$ containing an ϵ-enlargement of the set $\{F^{-1}_{Y_{0t}|D=1}(\tau) : \tau \in \mathcal{T}\}$. Suppose assumptions ID.1-ID.5, Assumption NP.1 or assumption P.1-P.3 hold. Then,

$$\sqrt{n}(Q\hat{T}T(\tau) - QTT(\tau)) \xrightarrow{d} \bar{\mathbb{G}}_1(\tau) - \bar{\mathbb{G}}_0(\tau)$$

where $(\bar{\mathbb{G}}_0(\tau), \bar{\mathbb{G}}_1(\tau))$ is a stochastic process in the metric space $(\ell^\infty(\mathcal{T}))^2$ with

$$\bar{\mathbb{G}}_0(\tau) = \frac{\mathbb{G}_0(F^{-1}_{Y_{0t}|D=1}(\tau))}{f_{Y_{0t}|D=1}(F^{-1}_{Y_{0t}|D=1}(\tau))} \qquad \bar{\mathbb{G}}_1(\tau) = \frac{\mathbb{G}_1(F^{-1}_{Y_t|D=1}(\tau))}{f_{Y_t|D=1}(F^{-1}_{Y_t|D=1}(\tau))}$$

Theorem 7 gives the limiting behavior of the QTT estimator. The result is unchanged from Callaway and Li (2019), since the asymptotic distribution of $\hat{G}_0(y) = \sqrt{n}(\hat{F}_{Y_{0t}|D=1}(y)) - F_{Y_{0t}|D=1}(y))$ and $\hat{G}_1(y) = \sqrt{n}(\hat{F}_{Y_t|D=1}(y)) - F_{Y_t|D=1}(y))$ is unchanged by the doubly-robust estimator of $F_{\Delta Y_{0t}|D=1}(y)$. This is analogous to a doubly-robust estimator having the same distribution as other semiparametric two-step estimators [4].

5 Simulations

In this section I will simulate the estimation of the QTT at $\tau = 0.5$ and $\tau = 0.75$. The goal is to demonstrate not only how my estimator performs in small samples, but also how it compares to the estimator of Callaway and Li (2019). The data generating process is as follows: I generate the

[4] See the introduction of Rothe and Firpo (2019)

following data generating process with $N = 1000$ and $T = 3$ for 200 iterations.

$$v \sim Normal(0, 1)$$

$$\eta|D = 0 \sim Normal(0, 1) \qquad \eta|D = 1 \sim Normal(1, 1)$$

$$x_1 \sim Normal(0, 1) \qquad x_2 \sim Normal(0, 2)$$

$$x_3 \sim Normal(0, 3) \qquad x_4 \sim Normal(0, 4)$$

$$Y_{t-2} = 0.25x_1 + 0.5x_2 + 0.75x_3 + x_4 + \eta + v \qquad Y_{t-1} = 1 + 0.5x_1 + 0.75x_2 + x_3 + 1.5x_4 + \eta + v$$

$$Y_{0t} = 2 + 0.25x_1 + 0.5x_2 + 0.75x_3 + x_4 + \eta + v \qquad Y_{1t} = 1.5x_1 + x_2 + 1.5x_3 + x_4 + \eta + v$$

$$Y_t = D \times Y_{1t} + (1 - D) \times Y_{0t}$$

$$P(X) = \frac{e^{\beta X'}}{1 + e^{\beta' X}} \qquad \beta = [-0.25, -0.5, -0.75, 1]$$

The data generating process is based upon Example 3 in Callaway and Li (2019). The estimators applied in these simulations are the estimators from (1) and (2). This is done to compare the small sample performance of each estimator to each other and the conditional estimator of Callaway and Li (2019). In the table below, $\hat{QTT}_{dr}$ denotes the estimates when both nuisance functions are correctly specified. $\hat{QTT}_{p}$ denotes the estimates when the propensity score is correctly specified. $\hat{QTT}_{cdf}$ denotes the estimates propensity score when the conditional cdf of ΔY_{0t} is correctly specified. $\hat{QTT}_{cl}$ denotes the estimates using the original Callaway and Li estimator. Estimator (1) is the estimator of equation (1), and Estimator (2) is the estimator of equation (2).

Table 1: $QTT(\tau)$ Estimates

	$\hat{QTT}_{dr}$	$\hat{QTT}_{p}$	$\hat{QTT}_{cdf}$	$\hat{QTT}_{cl}$
τ = 0.75 with Estimator (1)				
QTT	-3.331	-3.331	-3.331	-3.331
Estimate	-4.622	-4.656	-4.108	-3.713
se	0.9192	0.7794	0.4758	1.815
Replications	200	200	200	200
RMSE	1.683	1.536	0.9101	1.851
N	1000	1000	1000	1000
T	3	3	3	3
τ = 0.75 with Estimator (2)				
QTT	-3.331	-3.331	-3.331	-3.331
Estimate	-4.7342	-4.7387	-3.7663	-3.713
se	0.7543	0.6891	0.2645	1.815
Replications	200	200	200	200
RMSE	1.5919	1.566	0.63	1.851
N	1000	1000	1000	1000
T	3	3	3	3
τ = 0.5 with Estimator (1)				
QTT	-5.167	-5.167	-5.167	-5.167
Estimate	-5.195	-5.225	-4.2594	-4.739
se	0.9434	0.6805	0.2717	1.800
Replications	200	200	200	200
RMSE	0.9415	0.6813	0.9468	1.845
N	1000	1000	1000	1000
T	3	3	3	3
τ = 0.5 with Estimator (2)				
QTT	-5.167	-5.167	-5.167	-5.167
Estimate	-5.281	-5.248	-4.260	-4.739
se	0.6868	0.5907	0.2711	1.800
Replications	200	200	200	200
RMSE	0.6946	0.5948	0.9463	1.845
N	1000	1000	1000	1000
T	3	3	3	3

Here, misspecification of the propensity score is when the logit function is correctly chosen

in all cases, but only x_1 is considered to influence the probability of treatment. Misspecification of the cdf nuisance function in this case concerns a misspecification of the functional form itself, but the variables are correctly chosen. In this instance, the function is chosen to be the cdf of a uniform (0,1) random variable.

The estimator that I propose in this paper outperforms the Callaway and Li estimator when comparing the RMSE. Across the estimators, it would seem that if only one of the nuisance functions is correctly specified it can the estimator when both functions are correctly specified. Estimator (2) generally produces smaller standard errors than Estimator (1), though the bias of the estimator in this small sample does not seem to display a pattern across the two estimators. At the median, the bias of the doubly-robust estimator, regardless of which nuisance functions are correctly specified, outperforms the Callaway and Li estimator.

6 Conclusion

I have provided a doubly-robust estimator of the quantile treatment effect on the treated. This estimator relaxes the assumptions on the nuisance functions, allowing for a slower rate of convergence in order to achieve the limiting distribution of the QTT estimator. This causes nonparametric estimation of each of the nuisance functions to be much more viable, since nonparametric estimation requires assumptions upon the differentiable of the nuisance functions, which in turn affects the rate of convergence. As my simulations demonstrate, this leads to a lower RMSE in small samples, particularly when estimating the QTT at the median. Without the double-robustness property, confidence intervals could be so large that the QTT is not statistically different from 0 except at the extremes of the distribution of the difference in treated and untreated outcomes for the treated subpopulation.

It is important to recognize what this estimator is not. It is not a substitute for the doubly-robust estimator of the ATT that is presented in Sant'Anna and Zhao (2020). The assumptions necessary for identification are relaxed. There is no conditional copula assumption, and the par-

allel trends assumption is weaker than the conditional independence assumption on the difference in untreated outcomes. Instead, the two estimators should be used to complement each other. The ATT should be estimated along with quantile treatment effects on the treated at a variety of quantiles. If the estimate of the ATT is inconsistent with the results that are being presented across the information that is summarized by the QTTs, then perhaps either the conditional copula assumption or the conditional independence assumption does not hold.

What this estimator should be seen as is part of a middle ground between some of the more nonparametric estimators and estimators that rely entirely upon propensity score matching. In particular, the optimal transport methods of Gunsilius and Xu (2021) and Torous, Gunsilius, and Rigollet (2021) avoid the Curse of Dimensionality that is common with nonparametric estimation of the propensity score when estimating treatment effects; however, a doubly-robust estimator will relax the smoothness assumptions on the propensity score function in relation to the dimension of the covariate matrix. When supplemented with other doubly-robust estimators in the causal inference literature, my QTT estimator becomes part of a battery of doubly-robust estimators that increase the feasibility of propensity score matching.

References

Abadie, Alberto (2005). "Semiparametric difference-in-differences estimators". In: *The Review of Economic Studies* 72.1, pp. 1–19.

Athey, Susan and Guido W Imbens (2006). "Identification and inference in nonlinear difference-in-differences models". In: *Econometrica* 74.2, pp. 431–497.

Bonhomme, Stéphane and Ulrich Sauder (2011). "Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling". In: *Review of Economics and Statistics* 93.2, pp. 479–494.

Callaway, Brantly and Tong Li (2019). “Quantile treatment effects in difference in differences models with panel data”. In: *Quantitative Economics* 10.4, pp. 1579–1618.

Callaway, Brantly and Pedro HC Sant’Anna (2021). “Difference-in-differences with multiple time periods”. In: *Journal of Econometrics* 225.2, pp. 200–230.

Cameron, A Colin et al. (2004). “Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts”. In: *The Econometrics Journal* 7.2, pp. 566–584.

Caracciolo, Francesco and Marilena Furno (2017). “Quantile treatment effect and double robust estimators: an appraisal on the Italian labor market”. In: *Journal of Economic Studies*.

Card, David (1990). “The impact of the Mariel boatlift on the Miami labor market”. In: *ILR Review* 43.2, pp. 245–257.

Card, David and Alan Krueger (1994). *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*.

Chen, Xiaohong (2007). “Large sample sieve estimation of semi-nonparametric models”. In: *Handbook of econometrics* 6, pp. 5549–5632.

Chen, Xiaohong and Xiaotong Shen (1998). “Sieve extremum estimates for weakly dependent data”. In: *Econometrica*, pp. 289–314.

Chernozhukov, Victor et al. (2013). “Average and quantile effects in nonseparable panel models”. In: *Econometrica* 81.2, pp. 535–580.

Chernozhukov, Victor et al. (2016). “Locally robust semiparametric estimation”. In: *arXiv preprint arXiv:1608.00033*.

Fan, Jianqing et al. (2016). *Improving covariate balancing propensity score: A doubly robust and efficient approach*. Tech. rep. Technical report, Princeton University.

Firpo, Sergio (2007). "Efficient semiparametric estimation of quantile treatment effects". In: *Econometrica* 75.1, pp. 259–276.

Gunsilius, Florian and Yuliang Xu (2021). "Matching for causal effects via multimarginal optimal transport". In: *arXiv preprint arXiv:2112.04398*.

Hansen, Bruce E (2008). "Uniform convergence rates for kernel estimation with dependent data". In: *Econometric Theory* 24.3, pp. 726–748.

Heckman, James J, Hidehiko Ichimura, and Petra Todd (1998). "Matching as an econometric evaluation estimator". In: *The review of economic studies* 65.2, pp. 261–294.

Heckman, James J et al. (1998). *Characterizing selection bias using experimental data*.

Hirano, Keisuke, Guido W Imbens, and Geert Ridder (2003). "Efficient estimation of average treatment effects using the estimated propensity score". In: *Econometrica* 71.4, pp. 1161–1189.

Horvitz, Daniel G and Donovan J Thompson (1952). "A generalization of sampling without replacement from a finite universe". In: *Journal of the American statistical Association* 47.260, pp. 663–685.

Li, Qi and Jeffrey S Racine (2008). "Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data". In: *Journal of Business & Economic Statistics* 26.4, pp. 423–434.

Li, Qi and Jeffrey Scott Racine (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Lorentz, GG (1966). *Approximation of Functions, Athena Series*. Holt, Rinehart and Winston, New York.

Machado, José AF and José Mata (2005). "Counterfactual decomposition of changes in wage distributions using quantile regression". In: *Journal of applied Econometrics* 20.4, pp. 445–465.

Masry, Elias (1996). "Multivariate local polynomial regression for time series: uniform strong consistency and rates". In: *Journal of Time Series Analysis* 17.6, pp. 571–599.

Muris, Chris (2020). "Efficient GMM estimation with incomplete data". In: *Review of Economics and Statistics* 102.3, pp. 518–530.

Newey, Whitney K (1990). "Semiparametric efficiency bounds". In: *Journal of applied econometrics* 5.2, pp. 99–135.

Rosenbaum, Paul R and Donald B Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55.

Rothe, Christoph and Sergio Firpo (2013). "Semiparametric estimation and inference using doubly robust moment conditions". In:

— (2019). "Properties of doubly robust estimators when nuisance functions are estimated nonparametrically". In: *Econometric Theory* 35.5, pp. 1048–1087.

Sant'Anna, Pedro HC and Jun Zhao (2020). "Doubly robust difference-in-differences estimators". In: *Journal of Econometrics* 219.1, pp. 101–122.

Słoczyński, Tymon and Jeffrey M Wooldridge (2018). "A general double robustness result for estimating average treatment effects". In: *Econometric Theory* 34.1, pp. 112–133.

Sued, Mariela, Marina Valdora, and Víctor Yohai (2020). "Robust doubly protected estimators for quantiles with missing data". In: *TEST* 29.3, pp. 819–843.

Torous, William, Florian Gunsilius, and Philippe Rigollet (2021). "An Optimal Transport Approach to Causal Inference". In: *arXiv preprint arXiv:2108.05858*.

Vaart, Aad W and Jon A Wellner (1996). "Weak convergence". In: *Weak convergence and empirical processes*. Springer, pp. 16–28.

Wooldridge, Jeffrey M (2007). "Inverse probability weighted estimation for general missing data problems". In: *Journal of econometrics* 141.2, pp. 1281–1301.

— (2010). *Econometric analysis of cross section and panel data*. MIT press.

Appendices

The first proof is of the identification result in Theorem 1.

Proof of Theorem 1

Proof. Note that by Theorem 1 in Callaway and Li (2019), the first portion of the result is proven. All that remains is to show that $F_{\Delta Y_{0t}|D=1}(\delta) = E\left[\left(\frac{1-D}{p}\frac{\pi(x)}{1-\pi(x)}\right)\mathbb{1}\{\Delta Y_t \leq \delta\} - \left(\frac{1-D}{p}\frac{\pi(x)}{1-\pi(x)} - \frac{D}{p}\right)\tilde{P}(\Delta Y_{0t} \leq \delta|X)\right]$ if $\pi(x) = p(x)$ a.c., or $\tilde{P}(\Delta Y_{0t} \leq \delta|X) = P(\Delta Y_{0t} \leq \delta|X)$ a.c. Suppose $\pi(x) = p(x)$ a.c. Then,

$$
\begin{aligned}
&E\left[\left(\frac{(1-D)p(x)}{p(1-p(x))}\right)\mathbb{1}_{\Delta Y_t \leq \delta} - \left(\frac{(1-D)p(x)}{p(1-p(x))}\right) - \frac{D}{p}\right)\tilde{P}(\Delta Y_{0t} \leq \delta|X)\right] \\
&= E\left[\frac{p(x)E[(1-D)\mathbb{1}_{\Delta Y_t \leq \delta|D=0,X}}{p}\right] - E\left[\frac{p(x)\tilde{P}(\Delta Y_{0t}|X, D=0)}{p}\right] + E\left[\frac{p(x)\tilde{P}(\Delta Y_{0t} \leq \delta|X, D=1)}{p}\right] \\
&= E\left[\frac{p(x)P(\Delta Y_{0t} \leq \delta|X, D=0)}{p}\right] - E\left[\frac{p(x)\tilde{P}(\Delta Y_{0t} \leq \delta|X, D=1)}{p}\right] + E\left[\frac{p(x)\tilde{P}(\Delta Y_{0t} \leq \delta|X, D=1)}{p}\right] \\
&= E\left[\frac{p(x)P(\Delta Y_{0t} \leq \delta|X, D=1)}{p}\right] \\
&= E\left[\frac{P(\Delta Y_{0t} \leq \delta, D=1|X)}{p}\right] \\
&= P(\Delta Y_{0t} \leq \delta|D=1) \\
&= F_{\Delta Y_{0t}|D=1}(\delta)
\end{aligned}
$$

Now, suppose $\pi(x) \neq p(x)$ a.c. and $\tilde{P}(\Delta Y_{0t} \leq \delta|X) = P(\Delta Y_{0t} \leq \delta|X)$ a.c. Then,

$$
\begin{aligned}
&E\left[\left(\frac{(1-D)\pi(x)}{p(1-\pi(x))}\right)\mathbb{1}_{\Delta Y_t \leq \delta} - \left(\frac{(1-D)\pi(x)}{p(1-\pi(x))}\right) - \frac{D}{p}\right)P(\Delta Y_{0t} \leq \delta|X)\right] \\
&= E\left[\frac{p(x)P(\Delta Y_{0t} \leq \delta|X, D=1)}{p}\right] - E\left[\frac{(1-p(x))\pi(x)P(\Delta Y_{0t} \leq \delta|X, D=0)}{p(1-\pi(x))}\right] + E\left[\frac{(1-p(x))\pi(x)P(\Delta Y_{0t} \leq \delta|X, D=0)}{p(1-\pi(x))}\right] \\
&= E\left[\frac{P(\Delta Y_{0t} \leq \delta, D=1|X)}{p}\right] \\
&= P(\Delta Y_{0t} \leq \delta|D=1) \\
&= F_{\Delta Y_{0t}|D=1}(\delta)
\end{aligned}
$$

□

The proof holds in either the parametric or nonparametric subcase, though this proof is for a parametric submodel, while the nonparametric submodel will proceed similarly. For the purpose of estimation of $F_{\Delta Y_{0t}|D=1}(\delta)$, only the period of treatment and the period prior needs to be considered. If there are additional pre-treatment and post-treatment periods, they are not relevant to the density of the data that is used to estimate $F_{\Delta Y_{0t}|D=1}(\delta)$. The proof itself is similar to a result in Sant'Anna and Zhao (2020).

Proof of Theorem 2:

Proof. The density of $(y_t(1), y_t(0), y_{t-1}(0), d, x)$ with respect to some sigma-finite measure on $\mathcal{L} \in \mathbb{R}^3 \times \{0,1\} \times \mathbb{R}^k$ is given by

$$\bar{f}(y_t(1), y_t(0), y_{t-1}(0), d, x) = \bar{f}(y_t(1), y_t(0), y_{t-1}(0)|D=1, x)^d p(x)^d \bar{f}(y_t(1), y_t(0), y_{t-1}(0)|D=0, x)^{1-d}(1-p(x))^{1-d} f(x)$$

The density of the observed data is,

$$f(y_t, y_{t-1}, d, x) = f_1(y_t, y_{t-1}|D=1, x)^d p(x)^d f_0(y_t, y_{t-1}|D=0, x)^{1-d}(1-p(x))^{1-d} f(x)$$

where

$$f_1(\cdot,\cdot|D=1,x) = \int \bar{f}(\cdot, y_t(0), \cdot|D=1,x)dy_t(0)$$
$$f_0(\cdot,\cdot|D=0,x) = \int \bar{f}(y_t(1), \cdot, \cdot|D=0,x)dy_t(1)$$

Consider a parametric submodel indexed by a parameter θ,

$$f_\theta(y_t, y_{t-1}, d, x) = f_{1,\theta}(y_t, y_{t-1}|D=1,x)^d p_\theta(x)^d f_{0,\theta}(y_t, y_{t-1}|D=0,x)^{1-d}(1-p_\theta(x))^{1-d} f_\theta(x)$$

which equals $f(y_t, y_{t-1}, d, X)$ when $\theta = \theta_0$. The score is

$$s_\theta(y_t, y_{t-1}, d, x) = d s_{1\theta}(y_t, y_{t-1}|D=1,x) + (1-d)s_{0\theta}(y_t, y_{t-1}|D=0,x) + \frac{d-p_\theta(x)}{p_\theta(x)(1-p_\theta(x))}\dot{p}_\theta(x) + t_\theta(x)$$

where, for $d = 0, 1$

$$s_{d\theta}(y_t, y_{t-1}|D=d,x) = \frac{d}{d\theta} log f_{d,\theta}(y_t, y_{t-1}|D=d,X),\ \dot{p}_\theta(x)\frac{d}{dx}p_\theta(x),\ \text{and } t_\theta(x) = \frac{d}{d\theta} log f_\theta(x)$$

Then the tangent space is

$$\mathcal{F} = \{d s_1(y_t, y_{t-1}|D=1,x) + (1-d)s_0(y_t, y_{t-1}|D=0,x) + a(x)(d-p(x)) + t(x)\}$$

where $\iint s_d(y_t, y_{t-1}|D=d,X) f_d(y_t, y_{t-1}|D=d,x)dy_y dy_{t-1} = 0\ \forall x, d = 0, 1$, $\int t(x)f(x)dx = 0$ and $a(x)$ is any square integrable function of x. Under the assumption that $\Delta Y_{0t} \perp\!\!\!\perp D|X$, note that $\tau = F_{\Delta Y_t|D=1}(\delta)$,

$$\tau = E[E[\mathbb{1}_{\Delta Y_{0t}\leq\delta}|D=1,X]|D=1] = E[E[\mathbb{1}_{\Delta Y_{0t}\leq\delta}|D=0,X]|D=1]$$

For the parametric submodel under consideration, I note that

$$\tau(\theta) = \frac{\iiint \mathbb{1}_{y_t \leq \delta + y_{t-1}} p_\theta(x) f_{0,\theta}(y_t, y_{t-1}|D=0,x) f_\theta(x) dy_3 dy_2 dx}{\int p_\theta(x) f_\theta(x) dx}$$

Then,

$$\begin{aligned}\frac{\partial \tau(\theta_0)}{\partial \theta} = & \frac{\iiint \mathbb{1}_{y_t \leq \delta + y_{t-1}} p(x) s_0(y_t, y_{t-1}|D=0,x) f_0(y_t, y_{t-1}|D=0,x) f(x) dy_3 dy_2 dx}{p} \\ & + \frac{\int P(\Delta Y_{0t} \leq \delta|x, D=0) p(x) t(x) f(x) dx}{p} + \frac{\int P(\Delta Y_{0t} \leq \delta|x, D=0) p(x) \dot{p}(x) f(x) dx}{p} \\ & - \frac{\tau[\int \dot{p}(x) + p(x)t(x)] f(x) dx]}{p}\end{aligned}$$

Let the initial choice of an influence function be,

$$\begin{aligned}F_\tau(Y_t, Y_{t-1}, D, X) = & \frac{(1-D)p(x)}{p(1-p(x))} \mathbb{1}_{\Delta Y_t \leq \delta} - \frac{(1-D)p(x)}{p(1-p(x))} P(\Delta Y_t \leq \delta|X, D=0) \\ & + \frac{D - p(x)}{p} P(\Delta Y_t \leq \delta|X, D=0) + \frac{p(x)}{p} P(\Delta Y_t \leq \delta|X|D=0) - \frac{D}{p}\tau \\ = & \frac{(1-D)p(x)}{p(1-p(x))} \mathbb{1}_{\Delta Y_t \leq \delta} - \frac{(1-D)p(x)}{p(1-p(x))} P(\Delta Y_{0t} \leq \delta|X) \\ & + \frac{D}{p} P(\Delta Y_{0t} \leq \delta|X) - \frac{D}{p}\tau\end{aligned}$$

Note that for the parametric submodel with score $s_\theta(y_1, y_0, d, x)$, I can conclude that τ is a differentiable parameter since

$$\frac{\partial \tau(\theta_0)}{\partial \theta} = E[F_\tau(Y_t, Y_{t-1}, D, X) s_\theta(Y_1, Y_0, D, X)]$$

Since $F_\tau \in \mathcal{F}$, then by Theorem 3.1 of Newey (1990), $F_\tau(Y_t, Y_{t-1}, D, X)$ is the efficient influence function for $F_{\Delta Y_{0t}|D=1}(\delta)$. □

The next proof will prove the results for the nonparametric logit sieve estimator as in Hirano, Imbens, and Ridder (2003), but starting from different assumptions. The proof itself is broken up

in two parts, first under a set of regularity conditions I prove that the estimator is consistent. Then, I prove that it achieves the desired rate of convergence.

Lemma 1. Suppose b,c are arbitrary constants such that $b, c > 0$ and $b \neq c$. Then sign($log(\frac{b}{1+b}) - log(\frac{c}{1+c})$) ≠sign($log(\frac{1}{1+b}) - log(\frac{1}{1+c})$)

Proof. Suppose $b > c$. Then $log(\frac{1}{1+b}) - log(\frac{1}{1+c}) = log(\frac{1+c}{1+b})$. Since $b > c > 0$, then $0 < \frac{1+c}{1+b} < 1$, so $log(\frac{1+c}{1+b}) < 0$. Now, suppose that $log(\frac{b}{1+b}) - log(\frac{c}{1+c}) < 0$. Then $\frac{b}{1+b} < \frac{c}{1+c}$, which implies that $b < c$. This is a contradiction, so $log(\frac{b}{1+b}) - log(\frac{c}{1+c}) > 0$.
Now, suppose $c < b$. Then $log(\frac{1+c}{1+b}) > 1$. If $log(\frac{b}{1+b}) - log(\frac{c}{1+c}) > 0$, then $b > c$. This is a contradiction, so $log(\frac{b}{1+b}) - log(\frac{c}{1+c}) < 0$ □

Proof of Theorem 3:

Proof. Note that

$$\begin{aligned}
|\ell(\pi, \mathbf{x}_i) - \ell(\pi', \mathbf{x}_i)| = \Big| & D_i log(\frac{exp(\pi(\mathbf{x}_i))}{1 + exp(\pi(\mathbf{x}_i))}) + (1 - D_i) log(\frac{1}{1 + exp(\pi(\mathbf{x}_i))}) \\
& - D_i log(\frac{exp(\hat{\pi}(\mathbf{x}_i))}{1 + exp(\hat{\pi}(\mathbf{x}_i))}) - (1 - D_i) log(\frac{1}{1 + exp(\hat{\pi}(\mathbf{x}_i))}) \Big| \\
= \Big| & D_i [log(\frac{exp(\pi(\mathbf{x}_i))}{1 + exp(\pi(\mathbf{x}_i))}) - log(\frac{exp(\hat{\pi}(\mathbf{x}_i))}{1 + exp(\hat{\pi}(\mathbf{x}_i))})] \\
& + (1 - D_i)[log(\frac{1}{1 + exp(\pi(\mathbf{x}_i))}) - log(\frac{1}{1 + exp(\hat{\pi}(\mathbf{x}_i))})] \Big|
\end{aligned}$$

By the preceding lemma,

$$\begin{aligned}
& \Big| D_i [log(\frac{exp(\pi(\mathbf{x}_i))}{1 + exp(\pi(\mathbf{x}_i))}) - log(\frac{exp(\hat{\pi}(\mathbf{x}_i))}{1 + exp(\hat{\pi}(\mathbf{x}_i))})] + (1 - D_i)[log(\frac{1}{1 + exp(\pi(\mathbf{x}_i))}) - log(\frac{1}{1 + exp(\hat{\pi}(\mathbf{x}_i))})] \Big| \\
& \leq \Big| [log(\frac{exp(\pi(\mathbf{x}_i))}{1 + exp(\pi(\mathbf{x}_i))}) - log(\frac{exp(\hat{\pi}(\mathbf{x}_i))}{1 + exp(\hat{\pi}(\mathbf{x}_i))})] + [log(\frac{1}{1 + exp(\pi(\mathbf{x}_i))}) - log(\frac{1}{1 + exp(\hat{\pi}(\mathbf{x}_i))})] \Big| \\
& = |\pi(\mathbf{x}_i) - \pi(\mathbf{x}_i)'|
\end{aligned}$$

Then, $sup_{\pi,\pi' \; in \Pi : \|\pi - \pi'\| \leq \delta} |\ell(\pi, \mathbf{x}_i) - \ell(\pi', \mathbf{x}_i)| \leq \delta$. Hence, Condition (ii) of Theorem 3.5M in Chen (2007) is satisfied. Then by Condition 3.5M, $\pi_n \xrightarrow{p} p$ under $\|\cdot\|_s$.

Now, to prove the second part of the theorem consider the ℓ^2 metric defined by $||\pi - p|| = E\left[\left(\frac{d\ell(\pi,\mathbf{x}_i)}{d\pi}[\pi - p]\right)^2\right]$. Suppose $||\pi - p|| \leq \epsilon^2$. Note that by the mean value theorem, $\ell(\pi, \mathbf{x}_i) - \ell(p, \mathbf{x}_i) = \frac{d\ell(\tilde{\pi},\mathbf{x}_i)}{d\pi}[\pi - p]$, where $\tilde{\pi}$ lies between α and p. Then,

$$
\begin{aligned}
&sup_{||\pi-p||\leq\epsilon} E[[(\ell(\pi, \mathbf{x}_i) - \ell(p, \mathbf{x}_i))^2] \\
=&sup_{||\pi-p||\leq\epsilon} E\left[\left[\frac{\tilde{\pi}(\mathbf{x}_i)(y_i + y_i exp(\tilde{\pi}(\mathbf{x}_i)) - exp(\tilde{\pi}'(\mathbf{x}_i)))}{1 + exp(\tilde{\pi}(\mathbf{x}_i))}[\pi(\mathbf{x}_i) - p(\mathbf{x}_i)]\right]^2\right] \\
=&sup_{||\pi-p||\leq\epsilon} E\left[\left[\frac{C_1\pi(\mathbf{x}_i)(y_i + y_i exp(\pi(\mathbf{x}_i)) - exp(\pi(\mathbf{x}_i)))}{1 + exp(\pi(\mathbf{x}_i))}[\pi(\mathbf{x}_i) - p(\mathbf{x}_i)]\right]^2\right] \\
\leq&C_1\epsilon^2
\end{aligned}
$$

where C_1 is a constant and $C_1 > 0$. Then Condition 3.7 of Chen (2007) is satisfied By Lemma 2 in Chen and Shen (1998), $||\pi - p||_s \leq C_1||\pi - p||^{2p/(2p+d)}$. Then condition 3.8 of Chen (2007) is satisfied with $sup_{||\pi-p||\leq\delta}|\ell(\pi, \mathbf{x}_i) - \ell(p, \mathbf{x}_i)| \leq \delta^{2p/(2p+d)}C_1$. Then by Theorem 3.2 in Chen (2007), $||\pi - p|| = O_p(\epsilon_n)$, with $\epsilon = max\{\delta_n, |p - \sigma_n p|\}$. Let $u_\pi = sup_{\pi\in\Pi_n}||\pi||_\infty$, and $||\pi||_\infty = sup_{\mathbf{x}_i\in X}|\pi(\mathbf{x}_i)|$. Then for all $0 < \frac{\epsilon}{C_1^2} \leq \delta < 1, log(\frac{\epsilon}{C_1^2}, \mathcal{H}_n, ||\cdot||_\infty) \leq const \cdot k_n \cdot log(1 + \frac{4u_\pi}{\epsilon})$ by Lemma 2.5 in van der Geer (2000), where $k_n \uparrow$ as $n\to\infty$, but $\frac{k_n}{n}\to 0$. Then,

$$
\begin{aligned}
&\frac{1}{\sqrt{n}\delta_n^2}\int_{b\delta_n^2}^{\delta_n}\sqrt{H_\square(\epsilon, \mathcal{F}_n, ||\cdot||)}d\epsilon \\
&\leq \frac{1}{\sqrt{n}\delta_n^2}\int_{b\delta_n^2}^{\delta_n}\sqrt{k_n \cdot log\left(1 + \frac{4u_\pi}{\epsilon}\right)}d\epsilon \\
&\leq \frac{1}{\sqrt{n}\delta_n^2}\sqrt{k_n}\int_{b\delta_n^2}^{\delta_n} log\left(1 + \frac{4u_\pi}{\epsilon}\right)d\epsilon \\
&\leq \frac{1}{\sqrt{n}\delta_n^2}\sqrt{k_n}\delta_n \\
&\leq const
\end{aligned}
$$

Then $\delta_n \asymp \sqrt{\frac{k_n}{n}}$, and $||\sigma_k p - p||_s = O(k_n^{-p})$ by Lorentz (1966). Let $\delta_n \asymp ||\sigma_k p - p||_s$. Then the optimal rate is with $k_n = o(n^{1/(2p+d)})$. Then $||\pi_n - p|| = O_p(n^{-p/(2p+d)})$. Now, note that $||\pi_n - p||^{2p/(2p+d)} = o_p(n^{-2p^2/(2p+d)^2})$. Since $||\pi - p||_s \leq C_1||\pi - p||^{2p/(2p+d)}$ and $||\pi - p||_s = o_p(1)$, then

$\|\pi - p\|_s = o_p(n^{(-2p^2/(2p+d)^2)})$. This implies that $\|\pi - p\|_s = o_p(n^{-1/4})$. □

The next proof will tackle the case for nonparametric estimation of the conditional CDF.

The following lemma is similar to Lemma 1 in Li and Racine (2008).

Lemma 2. Under assumptions C.1-C.3, $E[\hat{\mu}(x)] = \mu(x) + O(|h^2|)$

Proof.

$$
\begin{aligned}
E[\hat{\mu}(x)] &= \int \mu(x_i^c) W\left(\frac{X_i^c - x_i^c}{h}\right) dx_i^c (nh_1 \dots h_q)^{-1} \\
&= \int \mu(x_i^c) k\left(\frac{x_{i1} - x_1}{h_1}\right) \times \cdots \times k\left(\frac{x_{iq} - x_q}{h_q}\right) \\
&= \int \mu(x_i^c + hv) k(v) dv \\
&= \int [\mu(x_i^c) + \sum_{s=1}^{q} \mu_s(x_i^c) h_s v_s + \frac{1}{2} \sum_{s=1}^{q} \sum_{\ell=1}^{q} \mu(x^c) h_s h_\ell v_s v_\ell] k(v) dv + O(|h|^3) \\
&= \mu(x^c) + \frac{\kappa}{2} \sum_{s=1}^{q} \mu_{ss}(x^c) h_s^2 + O(|h|^3) \\
&= \mu(x^c) + O(|h|^2)
\end{aligned}
$$

□

where $\kappa = \int v^2 k(v) dv$.

Lemma 3. Under Assumptions C.1-C.4, $E[\hat{\mu}(x)\tilde{F}(y|x^c)] = \mu(x)F(y|x^c) + \mu(x) \sum_{i=1}^{q} h_s^2 B_s(y, x) + o(|h|^2) + o(h_0^2)$

Proof. See Theorem 6.2 (i) in Li and Racine (2007) □

Now, the rate of uniform convergence proof largely follows Masry (1996). Furthermore, since by Li and Racine (2008) it is shown that at the optimal (to minimize the integrated mean square error) occurs when $h_1, \dots, h_q$ all converge to 0 at the same rate. I will denote this common h by h_{min}. **Note to committee: I might simply refer to the proof made in Masry (1996). This proof does not really differ from that proof.**

Lemma 4. Under assumptions C.1-C.4, $sup_{x\in D}|\hat{\mu}(x) - \mu(x)| = O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(|h|^2)$.

Proof. Note that by the Triangle Inequality,

$$\begin{aligned}|\hat{\mu}(x) - \mu(x)| &= |\hat{\mu}(x) - \mu(x) - E[\hat{\mu}(x)] + E[\hat{\mu}(x)]| \\ &\leq |\mu(x) - E[\hat{\mu}(x)]| + |E[\hat{\mu}(x)] - \hat{\mu}(x)|\end{aligned}$$

By Lemma 2 and Lemma 3, $|\mu(x) - E[\hat{\mu}(x)]| = O(|h|^2)$. Then it is sufficient to find the rate for $|E[\hat{\mu}(x)] - \hat{\mu}(x)|$. Since D is compact, it can be covered by a finite number $L = L(n)$ of cubes $I_k = I_{n,k}$ with centers $x_k = x_{n,k}$ having sides of length ℓ_n for $k = 1, ..., L(n)$. Clearly $\ell_n = constant/L^{1/d}(n)$. Since D is compact, write

$$\begin{aligned}sup_{x\in D}|E[\hat{\mu}(x)] - \hat{\mu}(x)| &= max_{1\leq k\leq L_n} sup_{x\in D\cap I_k}|E[\hat{\mu}(x)] - \hat{\mu}(x)| \\ &\leq max_{1\leq k\leq L_n} sup_{x\in D\cap I_k}|\hat{\mu}(x) - \hat{\mu}(x_{k,n})| \\ &+ max_{1\leq k\leq L_n}|E[\hat{\mu}(x_{k,n})] - \hat{\mu}(x_{k,n})| \\ &+ max_{1\leq k\leq L_n} sup_{x\in D\cap I_k}|E[\hat{\mu}(x_{k,n})] - E[\hat{\mu}(x)]| \\ &:= Q_1 + Q_2 + Q_3\end{aligned}$$

Since each kernel is Lispschitz, and the product of Lipschitz functions is a Lipschitz function,

$$\begin{aligned}Q_1 &= |\hat{\mu}(x) - \hat{\mu}(x_{k,n})| \\ &\leq |W_h(X_i^c, x^c) - W_h(X_i^c, x_{k,n}^c)| \\ &\leq (C_2/h_{min}^{q+1}) sup_{x\in D\cap I_k}|x - x_{k,n}| \\ &\leq C_2\ell_n/h_{min}^{q+1}.\end{aligned}$$

Let $\ell_n = (ln(n))^{1/2}h^{(q+2)/2}/n^{1/2}$. Then $Q_1 = O((ln(n)/(nh^q))^{1/2})$. Similarly, $Q_3 = O((ln(n)/(nh^q))^{1/2})$

Now let $W_n(x) = \hat{\mu}(x) - E[\hat{\mu}(x)] = \sum_{i=1} Z_{n,i}$ where,

$$Z_{n,i} = (nh_{min}^q)^{-1} \left[W_h(X_i^c, x_i^c)] - E[W_h(X_i^c, x_i^c)] \right]$$

For $\eta > 0$, we have

$$\begin{aligned}
P[Q_2 > \eta] &\leq P[max_{1 \leq k \leq L_n} W_n(x_{k,n}) > \eta] \\
&\leq P[W_n(x_{1,n}) > \eta \text{ or } W_n(x_{2,n}) > \eta, \ldots, \text{ or } W_n(x_{L(n),n}) > \eta] \\
&\leq P[W_n(x_{1,n}) > \eta + W_n(x_{2,n}) > \eta, \ldots, + W_n(x_{L(n),n}) > \eta] \\
&\leq sup_{x \in S} P[|W_n(x)| > \eta]
\end{aligned}$$

Since $\hat{\mu}(\cdot)$ is bounded, and letting $A = sup_{x \in D} |\hat{\mu}(x)|$, we have $|Z_{n,i}| \leq 2A/nh_{min}^q$ for all $i = 1, ..., n$. Define $\gamma_n = (nh_{min}^q ln(n))^{1/2}$. Then $\gamma_n |Z_{n,i}| \leq 2A(ln(n))/(nh_{min}^q]^{1/2} \leq 1/2$ for all $i = 1, ..., n$ for n sufficiently large. Using the inequality $e^x \leq 1 + x + x^2$ for $|x| \leq 1/2$, we have $e^{\gamma_n Z_{n,i}} \leq 1 + \gamma_n Z_{n,i} + \gamma_n^2 Z_{n,i}^2$. Hence, $E[e^{\gamma_n Z_{n,i}}] \leq 1 + \gamma_n^2 E[Z_{n,i}^2] \leq e^{E[\gamma_n^2 Z_{n,i}^2]}$. Then,

$$\begin{aligned}
P[|W_n(x)| > \eta] &= P[|\sum_{i=1}^{n} Z_{n,i}| > \eta] \\
&= P[\sum_{i=1}^{n} Z_{n,i} > \eta] + P[\sum_{i=1}^{n} Z_{n,i} < -\eta] \\
&\leq P[\sum_{i=1}^{n} Z_{n,i} > \eta] + P[-\sum_{i=1}^{n} Z_{n,i} > \eta] \\
&\leq E[e^{\gamma_n \sum_{i=1}^{n} Z_{n,i}}] + E[e^{-\gamma_n \sum_{i=1}^{n} Z_{n,i}}] \\
&\leq 2e^{-\gamma_n} e^{\gamma_n^2 \sum_{i=1}^{n} E(Z_{n,i}^2)} \\
&\leq 2e^{-\gamma_n} e^{A\gamma_n^2/(nh_{min}^q)}
\end{aligned}$$

Then $sup_{x \in D} P[|W_n(x)| > \eta] \leq 2e^{-\gamma_n \eta + \frac{A\gamma_n^2}{nh_{min}^q}}$. Let $\gamma_n \eta = C_3 ln(n)$. Choose $\gamma_n = [(nh_{min}^q ln(n)]^{1/2}$. Then $-\gamma_n \eta/\alpha + A\gamma_n^2/(nh_{min}^q) = -C_3 ln(n) + Aln(n) = -\alpha ln(n)$, where $\alpha = C_3 - A$. Since $sup_{x \in D} P[|W_n(x)| > \eta] \leq 2e^{-\gamma_n \eta + \frac{A\gamma_n^2}{nh_{min}^q}}$ and $P[Q_2 > \eta] \leq L(n) sup_{x \in D} P[|W_n(x)| > \eta]$, then $P[Q_2 > \eta_n] \leq 2L(n)/n^\alpha$. Choose

C_3 sufficiently large and $L(n)$ such that $\sum_{n=1}^{\infty} P[|Q_2/\eta_n| > 1] \leq 4\sum_{n=1}^{\infty} L(n)/n^{\alpha} < \infty$. Then by the Borel-Cantelli lemma, $Q_2 = O_p((ln(n))^{1/2}/(nh_{min}^q)^{1/2})$. □

Similarly, by Lemma 3, and by a similar result to Theorem 8, $sup_{x\in D}|\hat{\mu}(x)\tilde{F}(y|x^c) - \mu(x)F(y|x^c)| = O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(h_0^2) + O_p(|h|^2)$. Then I have the following theorem,

Proof of Theorem 4

Proof. Note that $\tilde{F}(y|x^c) = \frac{\hat{\mu}(x)\tilde{F}(y|x^c)}{\hat{\mu}(x)} = \frac{\hat{\mu}(x)\tilde{F}(y|x^c)/\mu(x)}{\hat{\mu}(x)/\mu(x)}$. By Theorem 8,

$$sup_{x\in D}|\hat{\mu}(x) - \mu(x)| = O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(|h|^2)$$

Then, by Lemma 4

$$\begin{aligned} sup_{x\in D}|\frac{\hat{\mu}(x)}{\mu(x)} - 1| &= sup_{x\in D}|\frac{\hat{\mu}(x) - \mu(x)}{\mu(x)}| \\ &\leq \frac{O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(|h|^2)}{sup_{x\in D}\mu(x)} \\ &= O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(|h|^2) \end{aligned}$$

Similarly, since

$$sup_{x\in D}|\hat{\mu}(x)\tilde{F}(y|x^c) - \mu(x)F(y|x^c)| = O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(h_0^2) + O_p(|h|^2)$$

Then,

$$\begin{aligned} sup_{x\in D}|\frac{|\hat{\mu}(x)\tilde{F}(y|x^c)}{\mu(x)} - F(y|x^c)| &\leq \frac{O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(h_0^2) + O_p(|h|^2)}{sup_{x\in D}\mu(x)} \\ &\leq O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(h_0^2) + O_p(|h|^2) \end{aligned}$$

Then,

$$\begin{aligned}
\tilde{F}(y|x^c) &= \frac{\hat{\mu}(x)\tilde{F}(y|x^c)/\mu(x)}{\hat{\mu}(x)/\mu(x)} \\
&= \frac{F(y|x^c) + O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(h_0^2) + O_p(|h|^2)}{1 + O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(h_0^2) + O_p(|h|^2)} \\
&= \frac{F(y|x^c) + O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(h_0^2) + O_p(|h|^2)}{1 + O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(|h|^2)} \\
&= F(y|x^c) + O_p(\frac{ln(n)^{1/2}}{(nh^q)^{1/2}}) + O_p(h^2)
\end{aligned}$$

□

Proof of Theorem 5:

Proof. **Consistency of estimator, nonparametric case**:

$$\hat{F}_{\Delta Y_{0t}|D=1}(\delta) = n^{-1}\sum_{i=1}^{n}\left[\left(\frac{\frac{1-D_i}{\frac{\sum_{k=1}^{n}D_k}{n}}\frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)}\right)\mathbb{1}\{\Delta Y_t \leq \delta\} - \left(\frac{\frac{1-D_i}{\frac{\sum_{k=1}^{N}D_k}{n}}\frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)} - \frac{\frac{D_i}{\sum_{k=1}^{n}D_k}}{n}\right)\hat{P}(\Delta Y_{0t} \leq \delta|X)\right]$$

Suppose that $\hat{\pi}(\mathbf{x}) \xrightarrow{p} \pi(\mathbf{x})$ Furthermore, $\sum_{k=1}^{n}\frac{D_k}{n} \xrightarrow{p} p$. Then by the WLLN and the Continuous Mapping Theorem,

$$n^{-1}\sum_{i=1}^{n}\left(\frac{\frac{1-D_i}{\frac{\sum_{k=1}^{n}D_k}{n}}\frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)}\right)\mathbb{1}\{\Delta Y_{ti} \leq \delta\} \xrightarrow{p} E\left(\frac{1-D}{p}\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\mathbb{1}\{\Delta Y_t \leq \delta\}\right)$$

Then,

$$n^{-1}\sum_{i=1}^{n}\left[-\left(\frac{\frac{1-D_i}{\frac{\sum_{k=1}^{N}D_k}{n}}\frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)} - \frac{\frac{D_i}{\sum_{k=1}^{n}D_k}}{n}\right)\hat{P}(\Delta Y_{0t} \leq \delta|X)\right] \xrightarrow{p} E\left[-\left(\frac{1-D_i}{p}\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} - \frac{D_i}{p}\right)\tilde{P}(\Delta Y_{0t} \leq \delta)\right]$$

This implies that $\hat{F}_{\Delta Y_{0t}|D=1}(\delta) \xrightarrow{p} E\left[\frac{1-D}{p}\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\mathbb{1}\{\Delta Y_t \leq \delta\}\right] + E\left[-\left(\frac{1-D_i}{p}\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} - \frac{D_i}{p}\right)\tilde{P}(\Delta Y_{0t} \leq \delta)\right]$. If

$\pi(\mathbf{x}) = p(\mathbf{x})$ a.c. or $\tilde{P}(\Delta Y_{0t} \leq \delta|X) = P(\Delta Y_{0t} \leq \delta|X)$ a.c., then by the previous theorem

$$E\left[\frac{1-D}{p}\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\mathbb{1}\{\Delta Y_t \leq \delta\}\right] + E\left[-\left(\frac{1-D_i}{p}\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} - \frac{D_i}{p}\right)P(\Delta Y_{0t} \leq \delta)\right] = F_{\Delta Y_{0t}|D=1}(\delta)$$

$$\hat{F}_{\Delta Y_{0t}|D=1}(\delta) - F_{\Delta Y_{0t}|D=1}(\delta) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\frac{1-D_i}{\frac{\sum_{k=1}^{n}D_k}{n}}\frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)}}\right)\mathbb{1}\{\Delta Y_t \leq \delta\} - \left(\frac{\frac{1-D_i}{\frac{\sum_{k=1}^{N}D_k}{n}}\frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)} - \frac{D_i}{\frac{\sum_{k=1}^{n}D_k}{n}}}\right)\left[\hat{P}(\Delta Y_{0t} \leq \delta|X_i)\right]\right] - F_{\Delta Y_{0t}|D=1}(\delta)$$

The next proof follows partly from the proof of Theorem 2(b) in Rothe and Firpo (2019). In particular, the object is to expand the doubly robust moment condition and demonstrate that each term converges in probability to 0 at the desired rate. In the parametric case, the proof is very similar to Sant'Anna and Zhao (2020). The proof in nonparametric case is also similar to Fan et al. (2016) when the nuisance function is estimated using a sieve approach. Now, I will expand

$$\hat{F}_{\Delta Y_{0t}|D=1}(\delta) - F_{\Delta Y_{0t}|D=1}(\delta) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\frac{1-D_i}{\frac{\sum_{k=1}^{n}D_k}{n}}\frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)}}\right)\mathbb{1}\{\Delta Y_t \leq \delta\} - \left(\frac{\frac{1-D_i}{\frac{\sum_{k=1}^{N}D_k}{n}}\frac{\hat{\pi}(\mathbf{x}_i)}{1-\hat{\pi}(\mathbf{x}_i)} - \frac{D_i}{\frac{\sum_{k=1}^{n}D_k}{n}}}\right)\left[\hat{P}(\Delta Y_{0t} \leq \delta|X_i)\right]\right] - F_{\Delta Y_{0t}|D=1}(\delta)$$

Let

$$\begin{aligned}
\psi_i^2 &= \left(\frac{1-D_i}{p}\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} - \frac{D_i}{p}\right)\\
\psi_i^{22} &= 0\\
\psi_i^1 &= \left(\frac{1-D_i}{p}\frac{1}{(1-\hat{\pi}(\mathbf{x}_i))^2}\right)\mathbb{1}\{\Delta Y_t \leq \delta\} - \left(\frac{1-D_i}{p}\frac{1}{(1-\hat{\pi}(\mathbf{x}_i))^2}\right)[P(\Delta Y_{0t} \leq \delta|X_i)]\\
\psi_i^{11} &= \left(\frac{1-D_i}{p}\frac{2\hat{\pi}(\mathbf{x}_i)}{(1-\hat{\pi}(\mathbf{x}_i))^3}\right)\mathbb{1}\{\Delta Y_t \leq \delta\} - \left(\frac{1-D_i}{p}\frac{2\hat{\pi}(\mathbf{x}_i)}{(1-\hat{\pi}(\mathbf{x}_i))^3}\right)[P(\Delta Y_{0t} \leq \delta|X_i)]\\
\psi_i^{12} &= \left(\frac{1-D_i}{p}\frac{1}{(1-\hat{\pi}(\mathbf{x}_i))^2}\right)\\
\psi_i^{13} &= \left(\frac{D_i-1}{\hat{p}^2}\frac{1}{(1-\pi(\mathbf{x}_i))^2}\right)\mathbb{1}\{\Delta Y_t \leq \delta\} - \left(\frac{D_i-1}{\hat{p}^2}\frac{1}{(1-\pi(\mathbf{x}_i))^2}\right)[P(\Delta Y_{0t} \leq \delta|X_i)]\\
\psi_i^{23} &= \left(\frac{D_i-1}{\hat{p}^2}\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} + \frac{D_i}{\hat{p}^2}\right)\\
\psi_i^3 &= \left[\left(\frac{D_i-1}{\hat{p}^2}\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right)\mathbb{1}\{\Delta Y_t \leq \delta\} - \left(\frac{D_i-1}{\hat{p}^2}\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} + \frac{D_i}{\hat{p}^2}\right)[P(\Delta Y_{0t} \leq \delta|X_i)]\right]
\end{aligned}$$

$$\psi_i^{33} = \left[\left(\frac{2(1-D_i)}{\hat{p}^3}\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right)\mathbb{1}\{\Delta Y_t \leq \delta\} - \left(\frac{2(1-D_i)}{\hat{p}^3}\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} - \frac{2D_i}{\hat{p}^3}\right)[P(\Delta Y_{0t} \leq \delta|X_i)]\right]$$

$$\Psi_n(\hat{p},\hat{\pi},\hat{P}) = \frac{1}{n}\sum_{i=1}^{n}\psi(D_i, p(x_i), P(\Delta Y_{0t} \leq \delta|X_i), \hat{p})$$

Then,

$$\begin{aligned}
\Psi_n(\hat{p},\hat{\pi},\hat{P}) - \Psi_n(p,\pi,P) = {} & \frac{1}{n}\sum_{i=1}^{n}\psi_i^1(\hat{\pi}(X_i) - \pi(X_i)) + \frac{1}{n}\sum_{i=1}^{n}\psi_i^2(\hat{P}(\Delta Y_{0t} \leq \delta|X_i) - P(\Delta Y_{0t} \leq \delta|X_i)) \\
& + \frac{1}{n}\sum_{i=1}^{n}\psi_i^3(\hat{p} - p) + \frac{1}{n}\sum_{i=1}^{n}\psi_i^{11}(\hat{\pi}(X_i) - \pi(X_i)^2 \\
& + \frac{1}{n}\sum_{i=1}^{n}\psi_i^{12}(\hat{P}(\Delta Y_{0t} \leq \delta|X_i) - P(\Delta Y_{0t} \leq \delta|X_i))(\hat{\pi}(X_i) - \pi(X_i)) \\
& + \frac{1}{n}\sum_{i=1}^{n}\psi_i^{13}(\hat{\pi}(X_i) - \pi(X_i))(\hat{p} - p) \\
& + \frac{1}{n}\sum_{i=1}^{n}\psi_i^{23}(\hat{P}(\Delta Y_{0t} \leq \delta|X_i) - P(\Delta Y_{0t} \leq \delta|X_i))(\hat{p} - p) \\
& + \frac{1}{n}\sum_{i=1}^{n}\psi_i^{22}(\hat{P}(\Delta Y_{0t} \leq \delta|X_i) - P(\Delta Y_{0t} \leq \delta|X_i))^2 + \frac{1}{n}\sum_{i=1}^{n}\psi_i^{33}(\hat{p} - p)^2 \\
& + O_p(||\hat{\pi} - \pi||_\infty^3) + O_p(||\hat{P} - P||_\infty^3) + O_p(||\hat{p} - p||_\infty^3)
\end{aligned}$$

It remains to show that each term is $o_p(n^{1/2})$. Each term outside of the first term converges converges either due to Firpo and Rothe (2018) or due to the fact that $|\hat{\pi}(x_i) - \pi(x_i)| = o_p(n^{1/4})$. For the first term, let $\mathbb{G}_n(f_0) = n^{1/2}(\mathbb{P}_n - \mathbb{P})f_0(D, \Delta Y_t, X)$, where $\mathcal{P}_n$ is the empirical measure, $\mathcal{P}$ is the expectation, and

$$f_0(D, \Delta Y_t, X) = \frac{(1-D)(\mathbb{1}_{\Delta Y - t \leq y} - P(\Delta Y_{0t} \leq y|x))}{p(1-\pi(x))}$$

Since $sup_{x\in\mathcal{X}}|\hat{\pi}(x) - \pi(x)| \lesssim o_p((\frac{k_n^{1/2}}{n^{1/2}})^{1/2} + k_n^{\frac{-2p^2}{2p+d}}) = o_p(1)$ by Theorem 3.2 in Chen (2007) and the previous proof, then define $\mathcal{F} = \{f_0 : ||\hat{\pi}(x) - \pi(x)||_\infty \leq \delta_n\}$, where $\delta_n = C(\frac{k_n^{1/2}}{n^{1/2}})^{1/2} + k_n^{\frac{-2p^2}{2p+d}})$ for some $C > 0$. By Lemma 3 in Firpo/Rothe (2013), $\mathbb{P}f_0(D.\Delta Y_t, X) = 0$. By the Markov inequality and

Corollary 19.35 of Van der Vaart (2000),

$$\frac{\sum_{i=1}^{n} \psi_i^1(\hat{\pi}(x) - \pi(x))}{n^{1/2}} \leq sup_{f_0 \in \mathcal{F}} \mathbb{G}_n(f_0) \lesssim J_{[]}(||F_0||_{p,2}, \mathcal{F}, L_2(p))$$

where $J_{[]}(||F_0||_{p,2}, \mathcal{F}, L_2(p))$ is the bracketing integral, and F_0 is the envelope function. Since p and $\pi(x)$ is bounded away from 0, then

$$|f_0(D, \Delta Y_t, X)| \lesssim \delta_n |\mathbb{1}_{\Delta Y - t \leq y} - P(\Delta Y_{0t} \leq y|x)| \coloneqq F_0$$

Then since $\mathbb{1}_{\Delta Y - t \leq y} - P(\Delta Y_{0t} \leq y|x))$ is bounded by 1, $||F_0||_{p,2} \leq \delta_n$. Then,

$$logN_{[]}(\epsilon, \mathcal{F}, L_2(p)) \lesssim logN_{[]}(\epsilon, \mathcal{F}_0 \delta_n, L_2(p) = logN_{[]}(\epsilon/\delta_n, \mathcal{F}_0, L_2(p) \lesssim logN_{[]}(\epsilon/\delta_n, \Lambda_c^p(\mathcal{X}), L_2(p) \lesssim (\delta_n/\epsilon)^{d/p}$$

where the last inequality follows by Corollary 2.7.2 in Vaart and Wellner (1996). Then,

$$J_{[]}(||F_0||_{p,2}, \mathcal{F}, L_2(p)) \lesssim \int_0^{\delta_n} \sqrt{logN_{[]}(\epsilon, \mathcal{F}, L_2(p))} d\epsilon \lesssim \int_0^{\delta_n} (\delta_n/\epsilon)^{d/p} d\epsilon \xrightarrow{\delta_n \xrightarrow{n \to \infty} 0} 0$$

Then $\frac{\sum_{i=1}^{n} \psi_i^1(\hat{\pi}(x) - \pi(x))}{n^{1/2}} = o_p(1)$.

Consistency of estimator, parametric case:

$$\hat{F}_{\Delta Y_{0t}|D=1}(\delta) = n^{-1} \sum_{i=1}^{n} \left[\left(\frac{1-D_i}{\frac{\sum_{k=1}^{n} D_k}{n}} \frac{\pi(\mathbf{x}_i; \hat{\gamma})}{1 - \pi(\mathbf{x}_i; \hat{\gamma})} \right) \mathbb{1}\{\Delta Y_t \leq \delta\} - \left(\frac{1-D_i}{\frac{\sum_{k=1}^{N} D_k}{n}} \frac{\pi(\mathbf{x}_i; \hat{\gamma})}{1 - \pi(\mathbf{x}_i; \gamma)} - \frac{D_i}{\frac{\sum_{k=1}^{n} D_k}{n}} \right) \hat{P}(\Delta Y_{0t} \leq \delta | X; \hat{\beta}) \right]$$

Suppose that $\hat{\gamma} \xrightarrow{p} \gamma^*$ Furthermore, $\sum_{k=1}^{n} \frac{D_k}{n} \xrightarrow{p} p$. Then by the WLLN and the Continuous Mapping Theorem,

$$n^{-1} \sum_{i=1}^{n} \left(\frac{1 - D_i}{\frac{\sum_{k=1}^{n} D_k}{n}} \frac{\pi(\mathbf{x}_i; \hat{\gamma})}{1 - \pi(\mathbf{x}_i; \hat{\gamma})} \right) \mathbb{1}\{\Delta Y_{0t} \leq \delta\} \xrightarrow{p} E\left(\frac{1 - D}{p} \frac{\pi(\mathbf{x}_i; \gamma^*)}{1 - \pi(\mathbf{x}_i; \gamma^*)} \mathbb{1}\{\Delta Y_{0t} \leq \delta\} \right)$$

Assume that $\hat{\beta} \xrightarrow{p} \beta^*$. Then,

$$n^{-1}\sum_{i=1}^{n}\left[-\left(\frac{\frac{1-D_i}{\sum_{k=1}^{N}D_k}\frac{\pi(\mathbf{x}_i;\hat{\gamma})}{1-\pi(\mathbf{x}_i;\hat{\gamma})}-\frac{D_i}{\frac{\sum_{k=1}^{n}D_k}{n}}\right)\hat{P}(\Delta Y_{0t}\leq\delta|X_i;\hat{\beta})\right]\xrightarrow{p}E\left[-\left(\frac{1-D_i}{p}\frac{\pi(\mathbf{x}_i;\gamma^*)}{1-\pi(\mathbf{x}_i;\gamma^*)}-\frac{D_i}{p}\right)\tilde{P}(\Delta Y_{0t}\leq\delta)\right]$$

This implies that $\hat{F}_{\Delta Y_{0t}|D=1}(\delta)\xrightarrow{p}E\left[\frac{1-D}{p}\frac{\pi(\mathbf{x};\gamma^*)}{1-\pi(\mathbf{x};\gamma)}\mathbb{1}\{\Delta Y_t\leq\delta\}\right]+E\left[-\left(\frac{1-D_i}{p}\frac{\pi(\mathbf{x}_i;\gamma^*)}{1-\pi(\mathbf{x}_i;\gamma^*)}-\frac{D_i}{p}\right)P(\Delta Y_{0t}\leq\delta)\right]$.

If $\pi(\mathbf{x};\gamma)=p(\mathbf{x};\gamma^*)$ a.c. or $\tilde{P}(\Delta Y_{0t}\leq\delta|X;\beta)=P(\Delta Y_{0t}\leq\delta|X;\beta^*)$ a.c., then by the previous theorem

$$E\left[\frac{1-D}{p}\frac{\pi(\mathbf{x};\gamma^*)}{1-\pi(\mathbf{x};\gamma^*)}\mathbb{1}\{\Delta Y_t\leq\delta\}\right]+E\left[-\left(\frac{1-D_i}{p}\frac{\pi(\mathbf{x}_i;\gamma^*)}{1-\pi(\mathbf{x}_i;\gamma^*)}-\frac{D_i}{p}\right)P(\Delta Y_{0t}\leq\delta)\right]=F_{\Delta Y_{0t}|D=1}(\delta)$$

Note that,

$$\begin{aligned}
&\hat{F}_{\Delta Y_{0t}|D=1}(\delta)-F_{\Delta Y_{0t}|D=1}(\delta)\\
&\qquad=n^{-1}\sum_{i=1}^{n}\left(\frac{1-D_i}{\frac{\sum_{k=1}^{n}D_k}{n}}\frac{\pi(\mathbf{x}_i;\hat{\gamma})}{1-\pi(\mathbf{x}_i;\hat{\gamma})}\right)\mathbb{1}\{\Delta Y_{ti}\leq\delta\}-E\left(\frac{1-D}{p}\frac{\pi(\mathbf{x};\gamma)}{1-\pi(\mathbf{x};\gamma)}\mathbb{1}\{\Delta Y_t\leq\delta\}\right)\\
&\qquad-n^{-1}\sum_{i,j=1}^{n}\left[\left(\frac{\frac{1-D_i}{\sum_{k=1}^{N}D_k}\frac{\pi(\mathbf{x}_i;\hat{\gamma})}{1-\hat{\pi}(\mathbf{x}_i;\hat{\gamma})}-\frac{D_i}{\frac{\sum_{k=1}^{n}D_k}{n}}}{n_{1-D}}\right)\left[\frac{(1-D_j)(\mathbb{1}\{\Delta\mu_{0t}(\mathbf{x}_i;\hat{\beta})+\Delta\hat{u}_{0tj}\})}{n_{1-D}}\right]\right]-E\left[\left(\frac{1-D_i}{p}\frac{\tilde{p}(\mathbf{x}_i;\gamma)}{1-\pi(\mathbf{x}_i;\gamma)}-\frac{D_i}{p}\right)\tilde{P}(\Delta Y_{0t}\leq\delta);\beta\right]\\
&\qquad=(C\hat{D}F^1-CDF^1)-(C\hat{D}F^2-CDF^2)
\end{aligned}$$

Note that since

$$\sum_{k=1}^{n}\frac{D_k}{n}-p=O_p(n^{-1/2})$$

$$n^{-1}\sum_{k=1}^{n}\frac{D_k}{p}\mathbb{1}\{\Delta Y_{ti}\leq\delta\}-E[\frac{D}{p}\mathbb{1}\{\Delta Y_{ti}\leq\delta\}]=O_p(n^{-1/2})=o_p(1)$$

$$(\sum_{k=1}^{n}\frac{D_k}{n}-p)^2=O_p(n^{-1})=o_p(n^{-1/2})$$

Let $w_0(D,X;\hat{\gamma})=\left(\frac{\frac{1-D}{\sum_{k=1}^{n}D_k}\frac{\pi(\mathbf{x};\hat{\gamma})}{1-\pi(\mathbf{x};\hat{\gamma})}}{n}\right)$.

Then,

$$
\begin{aligned}
&\sqrt{n}(C\hat{D}F^1 - CDF^1) \\
&= n^{-1/2}\sum_{i=1}^{n}(w_0(D_i, X_i; \hat{\gamma})\mathbb{1}\{\Delta Y_{ti} \le \delta\} - E[w_0(D, X; \gamma^*)\mathbb{1}\{\Delta Y_t \le \delta\}]) \\
&= n^{-1/2}\sum_{i=1}^{n}(\tilde{w}_0(D_i, X_i; \hat{\gamma})\mathbb{1}\{\Delta Y_{ti} \le \delta\} - E[w_0(D, X; \gamma^*)\mathbb{1}\{\Delta Y_t \le \delta\}]) \\
&- \sqrt{n}\sum_{i=1}^{n}\left((1-D_i)\frac{\pi(\mathbf{x}_i;\hat{\gamma})}{1-\pi(\mathbf{x};\hat{\gamma})}\right) - E\left[(1-D)\frac{\pi(\mathbf{x};\gamma^*)}{1-\pi(\mathbf{x};\gamma^*)}\right]\cdot\frac{E\left[(1-D)\frac{\pi(\mathbf{x}_i;\gamma^*)}{1-\pi(\mathbf{x};\gamma^*)}\mathbb{1}\{\Delta Y_t \le \delta\}\right]}{E\left[(1-D)\frac{\pi(\mathbf{x};\gamma^*)}{1-\pi(\mathbf{x};\gamma^*)}\right]^2} + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^{n}(\tilde{w}_0(D_i, X_i; \hat{\gamma})\mathbb{1}\{\Delta Y_{ti} \le \delta\} - E[w_0(D, X; \gamma^*)\mathbb{1}\{\Delta Y_t \le \delta\}]) \\
&- n^{-1/2}\sum_{i=1}^{n}((\tilde{w}_0(D_i, X_i; \hat{\gamma}) - 1)E[w_0(D, X; \gamma^*)\mathbb{1}\{\Delta Y_t \le \delta\}]) + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^{n}((\tilde{w}_0(D_i, X_i; \hat{\gamma})(\mathbb{1}\{\Delta Y_{ti} \le \delta\} - E[w_0(D, X; \gamma^*)\mathbb{1}\{\Delta Y_{ti} \le \delta\}]) + o_p(1)
\end{aligned}
$$

where

$$
\tilde{w}_0(D, X; \hat{\gamma}) = \frac{\pi(X;\hat{\gamma})(1-D)}{1-\pi(X;\hat{\gamma})}\Bigg/E\left[\frac{\pi(X;\gamma^*)(1-D)}{1-\pi(X;\gamma^*)}\right]
$$

Then, I do a second-order Taylor expansion around γ^*, so that

$$
\begin{aligned}
&\sqrt{n}(C\hat{D}F^1 - CDF^1) \\
&= n^{-1/2}\sum_{i=1}^{n} w_0(D_i, X_i; \gamma*)(\mathbb{1}\{\Delta Y_{ti} \le \delta\} - E[w_0(D, X; \gamma*)\mathbb{1}\{\Delta Y_t \le \delta\}]) \\
&+ (\hat{\gamma} - \gamma*)' \cdot n^{-1/2}\sum_{i=1}^{n}\dot{w}_0(D_i, X_i; \gamma*)(\mathbb{1}\{\Delta Y_{ti} \le \delta\} - E[w_0(D, X; \gamma*)\mathbb{1}\{\Delta Y_t \le \delta\}] + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^{n} w_0(D_i, X_i; \gamma*)(\mathbb{1}\{\Delta Y_{ti} \le \delta\} - E[w_0(D, X; \gamma*)\mathbb{1}\{\Delta Y_t \le \delta\}]) \\
&+ \sqrt{n}(\hat{\gamma} - \gamma*)' \cdot n^{-1}\sum_{i=1}^{n}\dot{w}_0(D_i, X_i; \gamma*)(\mathbb{1}\{\Delta Y_{ti} \le \delta\} - E[w_0(D, X; \gamma*)\mathbb{1}\{\Delta Y_t \le \delta\}] + o_p(1)
\end{aligned}
$$

$$= n^{-1/2} \sum_{i=1}^{n} w_0(D_i, X_i; \gamma*)(\mathbb{1}\{\Delta Y_{ti} \leq \delta\} - E[w_0(D, X; \gamma*)\mathbb{1}\{\Delta Y_t \leq \delta\}])$$

$$+ n^{-1/2} \sum_{i=1}^{n} l_{\gamma*}(W_i)' \cdot E[\dot{w}_0(D_i, X_i; \gamma*)(\mathbb{1}\{\Delta Y_{ti} \leq \delta\} - E[w_0(D, X; \gamma*)\mathbb{1}\{\Delta Y_t \leq \delta\}])] + o_p(1)$$

$$= n^{-1/2} \sum_{i=1}^{n} w_0(D_i, X_i; \gamma*)(\mathbb{1}\{\Delta Y_{ti} \leq \delta\} - E[w_0(D, X; \gamma*)\mathbb{1}\{\Delta Y_t \leq \delta\}])$$

$$+ n^{-1/2} \sum_{i=1}^{n} l_{\gamma*}(W_i)' \cdot E[\alpha(D_i, X_i; \gamma*)(\mathbb{1}\{\Delta Y_{ti} \leq \delta\} - E[w_0(D, X; \gamma*)\mathbb{1}\{\Delta Y_t \leq \delta\}])\dot{\pi}(X; \gamma*)] + o_p(1)$$

$$= n^{-1/2} \sum_{i=1}^{n} (w_0(D_i, X_i; \gamma*)(\mathbb{1}\{\Delta Y_{ti} \leq \delta\} - E[w_0(D, X; \gamma*)\mathbb{1}\{\Delta Y_t \leq \delta\}])$$

$$+ l_{\gamma*}(W_i)' \cdot E[\alpha(D_i, X_i; \gamma*)(\mathbb{1}\{\Delta Y_{ti} \leq \delta\} - E[w_0(D, X; \gamma*)\mathbb{1}\{\Delta Y_t \leq \delta\}])\dot{\pi}(X; \gamma*)]) + o_p(1)$$

where

$$\dot{w}(D, X; \gamma) = \alpha(D, X; \gamma)\dot{\pi}(X; \gamma)$$

$$\alpha(D, X; \gamma) = \frac{1 - D}{(1 - \pi(X; \gamma))^2} \Big/ E\left[\frac{\pi(X; \gamma^*)(1 - D)}{1 - \pi(X; \gamma^*)}\right]$$

Observe that,

$$\begin{aligned} C\hat{D}F^2 - CDF^2 &= n^{-1} \sum_{i,j=1}^{n} \left[\left(\frac{1-D_i}{\frac{\sum_{k=1}^{N} D_k}{n}} \frac{\pi(\mathbf{x}_i;\hat{\gamma})}{1-\hat{\pi}(\mathbf{x}_i;\hat{\gamma})} \left[\frac{(1-D_j)(\mathbb{1}\{\Delta\mu_{0t}(\mathbf{x}_i;\hat{\beta}) + \Delta\hat{u}_{0tj}\})}{n_{1-D}}\right]\right]\right] - E\left[\left(\frac{1-D_i}{p} \frac{\tilde{p}(\mathbf{x}_i;\gamma)}{1-\pi(\mathbf{x}_i;\gamma)}\right) \tilde{P}(\Delta Y_{0t} \leq \delta | X; \beta^*)\right] \\ &\quad - n^{-1} \sum_{i,j=1}^{n} \left[\left(\frac{D_i}{\frac{\sum_{k=1}^{N} D_k}{n}} \left[\frac{(1-D_j)(\mathbb{1}\{\Delta\mu_{0t}(\mathbf{x}_i;\hat{\beta}) + \Delta\hat{u}_{0tj}\})}{n_{1-D}}\right]\right]\right] - E\left[\left(\frac{D}{p}\right) \tilde{P}(\Delta Y_{0t} \leq \delta | X; \beta^*)\right] \\ &= (C\hat{D}F^{21} - CDF^{21}) - (C\hat{D}F^{22} - CDF^{22}) \end{aligned}$$

Similarly note that,

$$\begin{aligned} &\sqrt{n}(C\hat{D}F^{22} - CDF^{22}) \\ &= n^{-1/2} \sum_{i=1}^{n} w_1(D_i) \left(\tilde{P}(\Delta Y_{0ti} \leq \delta | X_i; \beta^*) - E[w_1(D)\tilde{P}(\Delta Y_{0t} \leq \delta | X; \beta^*)]\right) \end{aligned}$$

$$
\begin{aligned}
&+ \sqrt{n}(\hat{\beta}-\beta^*)' \cdot n^{-1}\sum_{i=1}^{n}(w_1(D_i)\dot{\tilde{P}}(\Delta Y_{0t}\le\delta|X_i;\beta^*)) + o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n} w_1(D_i)\left(\tilde{P}(\Delta Y_{0ti}\le\delta|X_i;\beta^*) - E[w_1(D)\tilde{P}(\Delta Y_{0t}\le\delta|X;\beta^*)]\right)\\
&+ n^{-1/2}\sum_{i=1}^{n} l(W_i;\beta^*)'E[w_1(D_i)\dot{\tilde{P}}(\Delta Y_{0t}\le\delta|X;\beta^*)] + o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n}(w_1(D_i)\left(\tilde{P}(\Delta Y_{0ti}\le\delta|X_i;\beta^*) - E[w_1(D)\tilde{P}(\Delta Y_{0t}\le\delta|X;\beta^*)]\right)\\
&+ l_{\beta^*}(W_i)'E[w_1(D_i)\dot{\tilde{P}}(\Delta Y_{0t}\le\delta|X;\beta^*)]) + o_p(1)
\end{aligned}
$$

where $w_1(D) = \frac{D}{\frac{\sum_{k=1}^{n} D_k}{n}}$ Furthermore, note that

$$
\begin{aligned}
&\sqrt{n}(C\hat{D}F^{21} - CDF^{21})\\
&= n^{-1/2}\sum_{i=1}^{n}\tilde{w}_1(D_i,X_i;\hat{\gamma})(\tilde{P}(\Delta Y_{0ti}\le\delta;|X_i\hat{\beta}) - E[w_0(D,X\,\gamma^*)\tilde{P}(\Delta Y_{0ti}\le\delta|X;\beta^*)] + o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n} w_1(D_i,X_i;\gamma^*)(\tilde{P}(\Delta Y_{0ti}\le\delta|X_I;\beta^*) - E[w_0(D,X;\gamma)\tilde{P}(\Delta Y_{0ti}\le\delta|X;\beta^*)]\\
&+ \sqrt{n}(\hat{\gamma}-\gamma^*)' \cdot E\left[\alpha(D,X;\gamma^*)\left(\tilde{P}(\Delta Y_{0t}\le\delta|X;\beta^*) - E[w_0(D,X;\gamma^*)\tilde{P}(\Delta Y_{0ti}\le\delta|X;\beta^*)]\right)\dot{\pi}(X;\gamma^*)\right]\\
&+ \sqrt{n}(\hat{\beta}-\beta^*)' \cdot E[w_0(D,X;\gamma^*)\dot{\tilde{P}}(\Delta Y_{0ti}\le\delta|X;\beta^*)] + o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n}(w_0(D_i,X_i;\gamma^*)(\tilde{P}(\Delta Y_{0ti}\le\delta|X_I;\beta^*) - E[w_0(D,X;\gamma^*)P(\Delta Y_{0ti}\le\delta|X;\beta^*)])\\
&+ l_{\gamma^*}(W_i)' \cdot E\left[\alpha(D,X;\gamma^*)\left(\tilde{P}(\Delta Y_{0t}\le\delta|X;\beta^*) - E[w_0(D,X;\gamma^*)\tilde{P}(\Delta Y_{0ti}\le\delta|X;\beta^*)]\right)\dot{\pi}(X;\gamma^*)\right]\\
&+ l_{\beta^*}(W_i)' \cdot E[w_0(D,X;\gamma^*)\dot{\tilde{P}}(\Delta Y_{0ti}\le\delta|X;\beta^*)]) + o_p(1)
\end{aligned}
$$

Then by combining all the asymptotic expansions, I obtain

$$
\begin{aligned}
&\sqrt{n}(\hat{F}_{\Delta Y_{0t}|D=1}(\delta) - F_{\Delta Y_{0t}|D=1}(\delta))\\
&= n^{-1/2}\sum_{i=1}^{n}(w_0(D_i,X_i;\gamma^*)(\mathbb{1}\{\Delta Y_{ti}\le\delta\} - E[w_0(D,X;\gamma^*)\mathbb{1}\{\Delta Y_t\le\delta\}])
\end{aligned}
$$

$$
\begin{aligned}
&+ l_{\gamma^*}(W_i)' \cdot E[\alpha(D_i, X_i; \gamma^*)(\mathbb{1}\{\Delta Y_{ti} \le \delta\} - E[w_0(D, X; \gamma^*)\mathbb{1}\{\Delta Y_t \le \delta\}])\dot{\pi}(X; \gamma^*)] \\
&+ [w_1(D_i)\left(\tilde{P}(\Delta Y_{0ti} \le \delta | X_I; \beta^*) - E[w_1(D)\tilde{P}(\Delta Y_{0t} \le \delta | X; \beta^*)]\right) \\
&+ l_{\beta^*}(W_i)' E[w_1(D_i)\dot{\tilde{P}}(\Delta Y_{0t} \le \delta | X; \beta^*)]] \\
&- [(w_0(D_i, X_i; \gamma^*)(\tilde{P}(\Delta Y_{0ti} \le \delta | X_i; \beta^*) - E[w_0(D, X; \gamma^*)\tilde{P}(\Delta Y_{0ti} \le \delta | X; \beta^*)]) \\
&+ l_{\gamma^*}(W_i)' \cdot E\left[\alpha(D, X; \gamma^*)\left(\tilde{P}(\Delta Y_{0t} \le \delta | X; \beta^*) - E[w_0(D, X; \gamma^*)\tilde{P}(\Delta Y_{0ti} \le \delta | X; \beta^*)]\right)\dot{\pi}(X; \gamma^*)\right] \\
&+ l_{\beta}(W_i)' \cdot E[w_0(D, X; \gamma^*)\dot{\tilde{P}}(\Delta Y_{0ti} \le \delta | X; \beta^*)])] + o_p(1)
\end{aligned}
$$

After simplification, I obtain,

$$
\begin{aligned}
&\sqrt{n}(\hat{F}_{\Delta Y_{0t}|D=1}(\delta) - F_{\Delta Y_{0t}|D=1}(\delta)) \\
&= n^{-1/2} \sum_{i=1}^{n} (w_0(D_i, X_i; \gamma^*)(\mathbb{1}\{\Delta Y_{ti} \le \delta\} - \tilde{P}(\Delta Y_{0ti} \le \delta | X_i; \beta^*) + E[w_0(\gamma)\tilde{P}(\Delta Y_{0ti} \le \delta | X; \beta^*)] - E[w_0 \mathbb{1}\{\Delta Y_t \le \delta\}]) \\
&+ [w_1(D_i)\left(\tilde{P}(\Delta Y_{0ti} \le \delta | X_I; \beta^*) - E[w_1 \tilde{P}(\Delta Y_{0t} \le \delta | X; \beta^*)]\right) \\
&+ l_{\beta^*}(W_i)' E[(\dot{\tilde{P}}(\Delta Y_{0ti} \le \delta | X; \beta^*))(w_1 - w_0)] \\
&+ l_{\gamma^*}(W_i)' E[\alpha(\gamma^*)(\mathbb{1}\{\Delta Y_{ti} \le \delta\} + \tilde{P}(\Delta Y_{0ti} \le \delta | X; \beta^*)) - E[w_0(\mathbb{1}\{\Delta Y_{ti} \le \delta\} - \tilde{P}(\Delta Y_{0ti} \le \delta | X; \beta^*)])\dot{\pi}(\gamma^*)]] + o_p(1)
\end{aligned}
$$

Now, suppose that the propensity score and the CDF of $\Delta Y_{0ti}|X$ are correctly specified.

Note that $l_{\beta^*}(W_i)' E[(\dot{P}(\Delta Y_{0ti} \le \delta; \beta^*))(w_1 - w_0)] = 0$,

$l_{\gamma^*}(W_i)' E[\alpha(\gamma^*)(\mathbb{1}\{\Delta Y_{ti} \le \delta\} + P(\Delta Y_{0ti} \le \delta | X; \beta^*)) - E[w_0(\mathbb{1}\{\Delta Y_{ti} \le \delta\} - P(\Delta Y_{0ti} \le \delta | X; \beta^*)])\dot{\pi}(\gamma^*)]] =$

0, and $E[w_0(\gamma^*)P(\Delta Y_{0ti} \le \delta | X; \beta^*)] - E[w_0 \mathbb{1}\{\Delta Y_t \le \delta\}] = 0$. Then,

$$
\begin{aligned}
&\sqrt{n}(\hat{F}_{\Delta Y_{0t}|D=1}(\delta) - F_{\Delta Y_{0t}|D=1}(\delta)) \\
&= n^{-1/2} \sum_{i=1}^{n} \left[w_0(D_i, X_i; \gamma^*)(\mathbb{1}\{\Delta Y_{ti} \le \delta\} - (w_0(D_i, X_i; \gamma^*) - w_1(D_i, X_i; \gamma^*))P(\Delta Y_{0ti} \le \delta | X_I; \beta^*) - w_1(D_i)F_{\Delta Y_{0t}|D=1}(\delta)\right] \\
&+ o_p(1) \\
&= n^{-1/2} \sum_{i=1}^{n} \psi(D_i, X_i, Y_{0i}, Y_{1i}) + o_p(1)
\end{aligned}
$$

□

Proof of Theorem 6:

Proof. Note that $\mathbb{1}\{\Delta y_t \leq y_1\}|y_1 \in \Delta\mathcal{Y}_{0t|D=1}\}$ is Donsker. Also note that $\frac{D}{p} - \frac{1-D}{p}\frac{\pi(x)}{1-\pi(x)}$ is a uniformly bounded and measurable function. Furthermore, $P(\Delta Y_{0ti} \leq \delta)$ is trivially Donsker. Then by Vaart and Wellner (1996), Example 2.10.10, and Vaart and Wellner (1996), Example 2.10.7, the result follows from the functional central limit theorem for empirical distribution functions. □

Proof of Proposition 1: See the proof of Proposition SA2 in the appendix of Callaway and Li (2019).

Proof of Theorem 7: See the proof of Theorem SA1 in the appendix of Callaway and Li (2019).