Estimating Counterfactual Distribution Functions via Optimal Distribution Balancing with Applications*†‡

Zongwu Cai[a], Ying Fang[b,c], Ming Lin[b,c], and Yaqian Wu[d,†]

[a]Department of Economics, University of Kansas, Lawrence, KS 66045, USA

[b]Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian 361005, China

[c]Department of Statistics and Data Sciences, Xiamen University, Xiamen, Fujian 361005, China

[d]School of Economics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

October 29, 2024

Abstract

In this paper, we propose a new method to estimate counterfactual distribution functions via the optimal distribution balancing weights, to avoid estimating the inverse propensity weights, which is sensitive to model specification and easily causes unstable estimates as well as often fails to adequately balance covariates. First, we demonstrate that the estimated weights exactly balance the estimated conditional distributions among the treated, untreated, and combined groups via a well-defined convex optimization problem. Secondly, we show that the resulting estimator of counterfactual distribution function converges weakly to a mean-zero Gaussian process at the parametric rate of the squared root n. Additionally, we show that a properly designed Bootstrap method can be used to obtain confidence intervals for conducting statistical inferences, together with its theoretical justification. Furthermore, with the estimates of counterfactual distribution functions, we provide methods to estimate the quantile treatment effects and test the stochastic dominance relationship between the potential outcome distributions. Moreover, Monte Carlo simulations are conducted to illustrate that the finite sample performance for the proposed estimator is better than the inverse propensity score weighted estimators in many scenarios. Finally, our empirical study revisits the effect of maternal smoking on infant birth weight.

Keywords: Counterfactual distribution function; Covariate balance; Quantile treatment effect; Stochastic dominance; Weighting scheme.

*The authors acknowledge the partially financial supports from the National Science Fund of China (NSFC) key project grants #72033008 and #72133002, Basic Science Center Program of NSFC with grant #71988101.

†*Corresponding author.*

‡E-mail addresses: caiz@ku.edu (Z. Cai), yifst1@xmu.edu.cn (Y. Fang), linming50@xmu.edu.cn (M. Lin), wuyq2024@hust.edu.cn (Y. Wu)

1 Introduction

Estimating the treatment effects of one treatment or exposure is a major goal in causal inferences. Average treatment effect (ATE) is often of interest to researchers and there is an extensive literature on estimation methods, including, to name just a few, the papers by Rosenbaum and Rubin (1983), Hahn (1998), Heckman et al. (1997), Hirano et al. (2003), and references therein. However, the limitation of ATE is that it cannot explain the distributional effects of the treatment variable on the outcome variable. Sometimes researchers may be more interested in caring about the treatment's impact on the distribution of the outcome variable, especially when the distribution of outcome is skewed or multimodal or long-tailed. For example, the effect of maternal smoking on infant birth weight is different at the different levels of infant birth weight. In such scenarios, estimating the counterfactual distribution functions is more desirable and meaningful (LaLonde, 1995; Almond et al., 2005; Tang et al., 2021).

Existing methods for estimating counterfactual distribution functions generally depend on the calculation of inverse propensity score weights (IPW); see, for example, the papers by Firpo (2007) and Donald and Hsu (2014). The IPW approach is widely used due to its desirable asymptotic properties (Hirano et al., 2003; Firpo, 2007). However, it often fails to properly balance the observable covariates in finite samples, as balance is achieved only in expectation, as addressed by Imai and Ratkovic (2014). Ensuring sufficient covariate balance is more critical than achieving optimal treatment assignment prediction when estimating treatment effects. Moreover, a slight misspecification of the propensity score model might lead to a substantial bias of the estimated treatment effect as discussed in Kang and Schafer (2007). Additionally, this approach can result in unstable estimates when a few observations have very large weights; see, for instance, the paper by Zubizarreta (2015) for details.

In light of these challenges, some new approaches have been proposed in the literature to address these issues in estimating ATE by balancing pre-specified functions of covariates, which are typically the linear terms or quadratic terms of the covariates. Some methods realted to propensity score models are proposed with the aim to balance these covariate functions in finite samples; see, for instance, the papers by Imai and Ratkovic (2014), Li et al. (2018), Zhao (2019), Ning et al. (2020), Fan et al. (2023), and references therein. Instead of explicitly modeling the propensity score, other methods argue that estimating

propensity scores is an intermediate step in obtaining weights and suggest that the weights can be directly estimated by balancing the covariate functions while minimizing a measure of the dispersion of the weights; see, for example, the papers by Hainmueller (2012), Zubizarreta (2015), Chan et al. (2016), Athey et al. (2018), Wang and Zubizarreta (2020), Josey et al. (2021), and references therein. However, as pointed out by Chan et al. (2016), Zhao and Percival (2017), Wang and Zubizarreta (2020), and Fan et al. (2023), the theoretical validity of using such methods to estimate ATE relies on the assumption that the conditional means of the potential outcomes can be linearly represented by the pre-specified balancing functions. Therefore, the methods proposed for estimating ATE cannot be directly applied to estimate counterfactual distribution functions, since the conditional distribution functions of outcome variables are generally not well approximated by linear combinations of the balancing functions commonly used for ATE estimation. This gap motivates us to develop a novel covariate balancing method for the estimation of counterfactual distribution functions, designed to circumvent the aforementioned shortcomings inherent in the IPW approach.

This paper proposes a three-step procedure to estimate the counterfactual distribution functions without modeling the propensity score. In the first step, we use a nonparametric/semiparametric method to estimate the conditional cumulative distribution functions (CDF) of the outcome variable given the covariates for the treated and control groups, respectively. Then, we determine the weights by minimizing their dispersion, while ensuring exact balance of the estimated conditional CDF across the treated, untreated, and combined groups. We call these weights as the *optimal distribution balancing weights* (ODBW). Finally, we estimate the counterfactual distribution functions using the weighted empirical CDF. We show that the proposed estimator converges weakly to a mean-zero Gaussian process at the conventional parametric convergence rate of $\sqrt{n}$ under certain regular conditions. Furthermore, we propose a properly designed Bootstrap method that can be used for statistical inferences, together with its theoretical justification. With the estimates of the counterfactual distribution functions, we also provide methods to estimate the quantile treatment effects (QTE) and test the stochastic dominance relationship between the potential outcome distributions.

It is interesting to note that some existing papers are related to our work, including Rothe (2010) and Hsu et al. (2022). For example, Rothe (2010) proposed a fully nonparametric procedure to evaluate the effect of changes in the distribution of covariates on the unconditional

distribution of the outcome variable, whereas Hsu et al. (2022) considered extrapolation of quantile treatment effects estimated from a status quo population to a counterfactual population. Our method shares the same first step to estimate the conditional CDF of potential outcomes as in their procedures, but differs in the estimation of the unconditional potential outcome CDF. We propose a weighting scheme that achieves covariate balancing to obtain the unconditional CDF estimation instead of simply averaging the estimated conditional CDF. Since the estimates based on covariate balancing are stable as long as the conditional CDF of potential outcomes can be well approximated by linear combinations of the balancing functions, the proposed estimation method is not sensitive to the estimates of the conditional CDF in the first step, which is supported by the results of Monte Carlo simulations reported in Section 4.

The remainder of the article is organized as follows. We describe the model framework and the proposed estimation procedure in Section 2, which also presents the asymptotic theory under some regularity conditions and includes a Bootstrap inference procedure with its theoretical justification. Section 3 demonstrates the usefulness of applying the developed method to make statistical inferences for quantile treatment effect and testing stochastic dominance relationship between the potential outcome distributions. Section 4 collects simulation results to evaluate the finite sample performance of the proposed methods. In Section 5, we revisit the effects of maternal smoking on infant birth weight. Finally, Section 6 concludes the paper. All technical proofs are relegated to Appendix.

2 Estimation of Counterfactual Distributions

We adopt the potential outcome framework initiated by Rubin (1974). Let $D \in \{0, 1\}$ be a binary treatment indicator such that $D = 1$ if the individual receives treatment; $D = 0$ otherwise. Define $Y(1) \in \mathbb{R}$ as the potential outcome if the individual is assigned to the treated group and $Y(0) \in \mathbb{R}$ as that to the untreated group. Furthermore, let $X \in \mathcal{X} \subset \mathbb{R}^p$ be a vector of covariates. We consider the random sample $\{Y_i(0), Y_i(1), D_i, X_i\}_{i=1}^n$ in an independent and identically distributed (i.i.d.) fashion. Note that the observed sample is $\{Y_i, D_i, X_i\}_{i=1}^n$, where $Y_i = (1 - D_i)Y_i(0) + D_iY_i(1)$ is the observed outcome variable.

Let $F_d(\cdot)$ denote the marginal cumulative distribution function of $Y(d)$ for $d = 0, 1$. Here, $F_0(\cdot)$ and $F_1(\cdot)$ are counterfactual distribution functions, since they do not arise as

distributions from any observable population. Our focus is on estimating $F_0(\cdot)$ and $F_1(\cdot)$. To properly identify $F_0(\cdot)$ and $F_1(\cdot)$, the following assumption is often made in the literature (Rosenbaum and Rubin, 1983).

Assumption 1 (Strong Ignorability)**.** *(i) Unconfoundedness:* $\{Y(0), Y(1)\} \perp\!\!\!\perp D \,|\, X$*;*
(ii) Overlap: for all $x \in \mathcal{X}$*, the probability score function* $\pi(x) = \mathbb{P}(D = 1 \,|\, X = x)$ *is bounded away from 0 and 1.*

The unconfoundedness assumption requires that the treatment assignment is independent of the potential outcomes conditional on the observed covariates. It rules out the unobserved factors that simultaneously affect the treatment assignment and the potential outcomes. The overlap condition requires that the support of X to be the same across the treated and untreated groups. These two assumptions together are called "strong ignorability" in the econometrics and/or statistics literature.

2.1 Distribution Balancing Estimation Method

We consider using the weighting method to estimate $F_0(\cdot)$ and $F_1(\cdot)$. Let $w_i \geq 0$ be the weight associated with the observation (Y_i, D_i, X_i) for $i = 1, \cdots, n$, and define the weight vector $\boldsymbol{w} = (w_1, \cdots, w_n)^T$. Further, define $w_{di} = 1(D_i = d)w_i$ for $d = 0$ and 1, where $1(\cdot)$ is the indicator function. We estimate $F_d(y)$ using the following weighted empirical CDF

$$\widehat{F}_d(y) = \frac{1}{n} \sum_{i=1}^{n} w_{di} 1\left(Y_i \leq y\right) \tag{1}$$

for $d = 0$ and 1. Conventionally, the weights are obtained by first modeling the propensity score function $\pi(x)$ and then inverting the estimated propensity scores, which are called the inverse propensity score weights. More specifically, the IPWs are defined as $w_{0i}^{\text{IPW}} = 1(D_i = 0)/[1 - \pi(X_i)]$ and $w_{1i}^{\text{IPW}} = 1(D_i = 1)/\pi(X_i)$. Despite being widely used, the IPW approach suffers from some problems in practice as discussed in Section 1.

First, we consider the estimation error of $\widehat{F}_d(y)$ defined in (1). To this end, let $F_d(\cdot|x)$ denote the conditional CDF of $Y(d)$ given $X = x$ for $d = 0, 1$. Given the weights w_{di}, the

estimation error of $\widehat{F}_d(y)$ can be decomposed as

$$\begin{aligned}
&\widehat{F}_d(y) - F_d(y) = \frac{1}{n}\sum_{i=1}^{n} w_{di} 1\left(Y_i \leq y\right) - F_d(y) \\
=&\left[\frac{1}{n}\sum_{i=1}^{n} w_{di} F_d(y \,|\, X_i) - \frac{1}{n}\sum_{i=1}^{n} F_d(y \,|\, X_i)\right] + \left[\frac{1}{n}\sum_{i=1}^{n} F_d(y \,|\, X_i) - F_d(y)\right] \\
&+ \frac{1}{n}\sum_{i=1}^{n} w_{di}\left[1\left(Y_i \leq y\right) - F_d(y \,|\, X_i)\right] \\
:=&B_1 + B_2 + B_3.
\end{aligned}$$

One can see clearly that the second term B_2 and the third term B_3 go to zero by the law of large numbers under certain regularity conditions. If we choose the weights such that

$$\frac{1}{n}\sum_{i=1}^{n} w_{di} F_d\left(y \,|\, X_i\right) = \frac{1}{n}\sum_{i=1}^{n} F_d\left(y \,|\, X_i\right), \tag{2}$$

which is termed as *the distribution balancing condition*, then, $F_d(y \,|\, X_i)$ achieves balance between the treated/untreated group and full population after weighting. Therefore, the first term B_1 is zero and $\widehat{F}_d(y)$ becomes a consistent estimator of $F_d(y)$.

In practice, the balancing condition (2) cannot be directly applied. Firstly, the conditional CDF $F_d\left(y \,|\, x_i\right)$ is unknown and must be estimated. To this end, we propose using the kernel method to estimate $F_d\left(y \,|\, X_i\right)$. When the dimension of X_i is relatively large, some semiparametric estimation methods can also be employed, as described in Section 2.5. Secondly, to estimate $F_d(y)$ for all $y \in \mathcal{Y}$, where $\mathcal{Y}$ is the support of the outcome variable, it is infeasible to require that the balance condition (2) exactly holds for all $y \in \mathcal{Y}$. Without loss of generality, we assume that $\mathcal{Y}$ is a closed interval, denoted as $[y_l, y_u]$. Let $y_l = q_1 < \cdots < q_J = y_u$ be the equally spaced grid points on $[y_l, y_u]$. Then, according to the polynomial interpolation error formula from Süli and Mayers (2003), we have

$$F_d\left(y \,|\, X_i\right) = \sum_{j=1}^{J} c_j(y) F_d\left(q_j \,|\, X_i\right) + \frac{F_d^{(J)}\left(\xi_y \,|\, X_i\right)}{J!}\prod_{j=1}^{J}\left(y - q_j\right) \tag{3}$$

for any $y \in [y_l, y_u]$, where $c_j(y) = \prod_{k=1, k\neq j}^{J} \frac{y-q_k}{q_j - q_k}$, $\xi_y \in [y_l, y_u]$ depends on y and X_i, and $F_d^{(J)}\left(\xi_y \,|\, X_i\right) = \frac{\partial^J F_d(u \,|\, X_i)}{\partial u^J}\big|_{u=\xi_y}$. It is easy to see that $\left|\prod_{j=1}^{J}\left(y - q_j\right)\right| = \prod_{j=1}^{J}\left|y - q_j\right| \leq \frac{(J-1)!}{4}\left(\frac{y_u - y_l}{J-1}\right)^J$. If we assume that the absolute value of $F_d^{(J)}\left(\xi_y \,|\, X_i\right)$ is bounded by $C_0 > 0$

that does not depend on J, then, equation (3) leads to

$$\left| F_d\left(y \mid X_i\right) - \sum_{j=1}^{J} c_j(y) F_d\left(q_j \mid X_i\right) \right| \leq \frac{C_0}{4J} \left(\frac{y_u - y_l}{J-1} \right)^J.$$

Assume that the balance condition (2) exactly holds for $y = q_1, \cdots, q_J$. Since $F_d\left(q_J \mid X_i\right) \equiv 1$ by definition, $\frac{1}{n}\sum_{i=1}^{n} w_{di} F_d\left(q_J \mid X_i\right) = \frac{1}{n}\sum_{i=1}^{n} F_d\left(q_J \mid X_i\right)$ implies $\frac{1}{n}\sum_{i=1}^{n} w_{di} = 1$ for $d = 0$ and 1. Therefore, we always assume $\frac{1}{n}\sum_{i=1}^{n} w_{0i} = 1$ and $\frac{1}{n}\sum_{i=1}^{n} w_{1i} = 1$ are in the balance conditions. Notice that $c_j(y)$ does not depend on X_i. Thus, for any $y \in [y_l, y_u]$, the balance error is

$$\begin{aligned}
& \left| \frac{1}{n}\sum_{i=1}^{n} w_{di} F_d\left(y \mid X_i\right) - \frac{1}{n}\sum_{i=1}^{n} F_d\left(y \mid X_i\right) \right| \\
\leq & \left| \frac{1}{n}\sum_{i=1}^{n} w_{di} \sum_{j=1}^{J} c_j(y) F_d\left(q_j \mid X_i\right) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} c_j(y) F_d\left(q_j \mid X_i\right) \right| + \frac{C_0}{2J} \left(\frac{y_u - y_l}{J-1} \right)^J \\
= & \left| \sum_{j=1}^{J} c_j(y) \left[\frac{1}{n}\sum_{i=1}^{n} w_{di} F_d\left(q_j \mid X_i\right) \right] - \sum_{j=1}^{J} c_j(y) \left[\frac{1}{n}\sum_{i=1}^{n} F_d\left(q_j \mid X_i\right) \right] \right| + \frac{C_0}{2J} \left(\frac{y_u - y_l}{J-1} \right)^J \\
= & \frac{C_0}{2J} \left(\frac{y_u - y_l}{J-1} \right)^J.
\end{aligned}$$

It is clear that the balance error for any $y \in [y_l, y_u]$ tends to zero as the number of grid points J approaches infinity. We can control J so that the balance error is negligible relative to the asymptotic performance of $\widehat{F}_d(y)$.

Now, we present a three-step procedure for estimating the counterfactual distribution functions $\widehat{F}_0(y)$ and $\widehat{F}_1(y)$ as follows.

Step 1: We estimate the conditional CDF $F_d\left(y \mid X\right)$, $d = 0, 1$, by the Nadaraya-Watson estimator as

$$\widetilde{F}_d(y \mid x) = \frac{\sum_{i=1}^{n} 1(D_i = d) 1\left(Y_i \leq y\right) K_{h_d}\left(X_i - x\right)}{\sum_{i=1}^{n} 1(D_i = d) K_{h_d}\left(X_i - x\right)}, \quad d = 0, 1, \tag{4}$$

where $K(\cdot)$ is a kernel function, h_d is the bandwidth, and $K_{h_d}\left(X_i - x\right) = h_d^{-p} K\left(\left(X_i - x\right)/h_d\right)$.

Step 2: Compute the optimal distribution balancing weights $\widehat{\boldsymbol{w}} = (\widehat{w}_1, \cdots, \widehat{w}_n)^T$ by letting

$$\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \frac{1}{n}\sum_{i=1}^{n} \phi\left(w_i\right) \tag{5}$$

subject to $w_i \geq 0$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} 1(D_i = 0) w_i \widetilde{F}_0\left(q_j \mid X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{F}_0\left(q_j \mid X_i\right), \quad j = 1, \cdots, J, \\
&\frac{1}{n}\sum_{i=1}^{n} 1(D_i = 1) w_i \widetilde{F}_1\left(q_j \mid X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{F}_1\left(q_j \mid X_i\right), \quad j = 1, \cdots, J, \\
\text{and} \quad &\text{for } \iota = 1, \cdots, L, \\
&\frac{1}{n}\sum_{i=1}^{n} 1(D_i = 0) w_i u_\iota\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} 1(D_i = 1) w_i u_\iota\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} u_\iota\left(X_i\right),
\end{aligned}$$

where $\phi\left(w_i\right)$ is a non-negative, continuously differentiable and strictly convex function, which includes some special cases such as the entropy divergence as in Hainmueller (2012) with $\phi(w_i) = w_i \log(w_i)$, the stable balancing variance considered in Zubizarreta (2015) defined as $\phi(w_i) = 1(D_i = 0)(w_i - n/n_0)^2 + 1(D_i = 1)(w_i - n/n_1)^2$ with $n_0 = \sum_{i=1}^{n} 1(D_i = 0)$ and $n_1 = \sum_{i=1}^{n} 1(D_i = 1)$, and other distance measures as in Chan et al. (2016), Wang and Zubizarreta (2020) and Josey et al. (2021). The above objective function $\frac{1}{n}\sum_{i=1}^{n} \phi\left(w_i\right)$ measures the dispersion of the weights $w_1, \cdots, w_n$ and minimizing $\frac{1}{n}\sum_{i=1}^{n} \phi\left(w_i\right)$ tries to control the variance of the estimator. In addition, besides the key constraint developed from the balance condition in (2), we also allow other functions $u_\iota(X_i)$, $\iota = 1, \cdots, L$, to be balanced across the treated group, untreated group, and combined group. For example, taking $u_\iota\left(X_i\right) = X_i^\iota$ means to balance the ιth moment; see, for example, the papers by Imai and Ratkovic (2014) and Fan et al. (2023) for details. The algorithm to calculate the optimal distribution balancing weights is presented in Section 2.2.

Step 3: Let $\widehat{w}_{di} = 1(D_i = d)\widehat{w}_i$ for $d = 0, 1$ and $i = 1, \cdots, n$. Then, the counterfactual distribution functions $F_d(y)$ for $d = 0$ and 1 are estimated by

$$\widehat{F}_d(y) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}_{di} 1\left(Y_i \leq y\right) \tag{6}$$

for all $y \in \mathcal{Y}$.

Remark 1. It is worth to mention that Rothe (2010), Hsu et al. (2022), and Cai et al. (2022)

also considered the estimation of the counterfactual distributions $F_d(y)$ using $\widetilde{F}_d(y \mid x)$ in (4). Different from our method, they proposed estimating $F_d(y)$ by $\widetilde{F}_d(y) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{F}_d(y|X_i)$. Rothe (2010) demonstrated that the estimator $\widetilde{F}_d(y)$ achieves $\sqrt{n}$-consistency by using high-order kernels. However, the performance of $\widetilde{F}_d(y)$ greatly depends on the choice of bandwidth h_d in (4). For our method, $\widetilde{F}_d(y \mid x)$ is only used within the balance conditions, and the counterfactual distribution $F_d(y)$ is estimated by the weighted empirical CDF. The new estimator $\widehat{F}_d(y)$ is not very sensitive to the choice of bandwidth h_d in the first stage, which is supported by the results of Monte Carlo simulations reported in Section 4.

2.2 Implementation of Optimal Weights

The constrained optimization problem in (5) is a convex separable programming problem with linear constraints. Actually, Tseng and Bertsekas (1987) showed that its dual problem is an unconstrained convex maximization problem that can be solved by efficient and stable numerical algorithms. Therefore, we consider its dual problem in the following.

Define $\widetilde{\mathbf{U}}_d(x) = \left(\widetilde{\mathbf{F}}_d(\mathbf{q} \mid x)^T, \mathbf{u}(x)^T\right)^T$ with $\widetilde{\mathbf{F}}_d(\mathbf{q} \mid x) = \left(\widetilde{F}_d(q_1 \mid x), \cdots, \widetilde{F}_d(q_J \mid x)\right)^T$ and $\mathbf{u}(x) = (u_1(x), \cdots, u_L(x))^T$ for $d = 0$ and 1. The Lagrangian of the optimization problem in (5) can be written as

$$
\begin{aligned}
L_n(\boldsymbol{w}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1) =& \frac{1}{n}\sum_{i=1}^{n} \phi(w_i) + \sum_{j=1}^{J} \lambda_{0,j} \left[\frac{1}{n}\sum_{i=1}^{n} 1(D_i = 0) w_i \widetilde{F}_0(q_j \mid X_i) - \frac{1}{n}\sum_{i=1}^{n} \widetilde{F}_0(q_j \mid X_i)\right] \\
&+ \sum_{\iota=1}^{L} \lambda_{0,J+\iota} \left[\frac{1}{n}\sum_{i=1}^{n} 1(D_i = 0) w_i u_\iota(X_i) - \frac{1}{n}\sum_{i=1}^{n} u_\iota(X_i)\right] \\
&+ \sum_{j=1}^{J} \lambda_{1,j} \left[\frac{1}{n}\sum_{i=1}^{n} 1(D_i = 1) w_i \widetilde{F}_1(q_j \mid X_i) - \frac{1}{n}\sum_{i=1}^{n} \widetilde{F}_1(q_j \mid X_i)\right] \\
&+ \sum_{\iota=1}^{L} \lambda_{1,J+\iota} \left[\frac{1}{n}\sum_{i=1}^{n} 1(D_i = 1) w_i u_\iota(X_i) - \frac{1}{n}\sum_{i=1}^{n} u_\iota(X_i)\right] \\
=& \frac{1}{n}\sum_{i=1}^{n} \phi(w_i) + \frac{1}{n}\sum_{i=1}^{n}\sum_{d=0,1} 1(D_i = d) w_i \widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d - \frac{1}{n}\sum_{i=1}^{n}\sum_{d=0,1} \widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d,
\end{aligned}
\tag{7}
$$

where $\boldsymbol{\lambda}_0 = (\lambda_{0,1}, \cdots, \lambda_{0,J+L})^T$ and $\boldsymbol{\lambda}_1 = (\lambda_{1,1}, \cdots, \lambda_{1,J+L})^T$ are the Lagrange multipliers.

The first order condition $\partial L_n(\boldsymbol{w}, \boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1)/\partial w_i = 0$ yields

$$\phi'(w_i) = -\sum_{d=0,1} 1(D_i = d)\widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d,$$

where $\phi'(\cdot)$ is the first derivative of $\phi(\cdot)$. Let $(\phi')^{-1}(\cdot)$ be the inverse function of $\phi'(\cdot)$. Then,

$$w_i = (\phi')^{-1}\left(-\sum_{d=0,1} 1(D_i = d)\widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d\right). \tag{8}$$

For simplicity, define $\rho(t) = \phi\left\{(\phi')^{-1}(-t)\right\} + t(\phi')^{-1}(-t)$. Plugging (8) back into (7) eliminates the constraints, resulting in an unrestricted dual maximization problem given by

$$\begin{aligned}
\widetilde{G}_n(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1) =& \frac{1}{n}\sum_{i=1}^n \rho\left(\sum_{d=0,1} 1(D_i = d)\widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d\right) - \frac{1}{n}\sum_{i=1}^n \sum_{d=0,1} \widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d \\
=& \frac{1}{n}\sum_{i=1}^n \sum_{d=0,1} 1(D_i = d)\rho\left(\widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d\right) - \frac{1}{n}\sum_{i=1}^n \sum_{d=0,1} \widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d \\
:=& \widetilde{G}_{n,0}(\boldsymbol{\lambda}_0) + \widetilde{G}_{n,1}(\boldsymbol{\lambda}_1),
\end{aligned}$$

where

$$\widetilde{G}_{n,d}(\boldsymbol{\lambda}_d) = \frac{1}{n}\sum_{i=1}^n 1(D_i = d)\rho\left(\widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d\right) - \frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d. \tag{9}$$

It is clear that $\rho'(t) = (\phi')^{-1}(-t)$ and $\rho''(t) = -1/\phi''\left((\phi')^{-1}(-t)\right)$. Thus, both $\widetilde{G}_{n,0}(\boldsymbol{\lambda}_0)$ and $\widetilde{G}_{n,1}(\boldsymbol{\lambda}_1)$ are strictly concave due to the strict convexity of $\phi(\cdot)$. Therefore, the solution to the constrained optimization problem in (5) is

$$\widehat{w}_i = (\phi')^{-1}\left(-\sum_{d=0,1} 1(D_i = d)\widetilde{\mathbf{U}}_d(X_i)^T \widehat{\boldsymbol{\lambda}}_d\right) = \rho'\left(\sum_{d=0,1} 1(D_i = d)\widetilde{\mathbf{U}}_d(X_i)^T \widehat{\boldsymbol{\lambda}}_d\right), \tag{10}$$

where $\widehat{\boldsymbol{\lambda}}_0$ and $\widehat{\boldsymbol{\lambda}}_1$ are the unique maximizers of $\widetilde{G}_{n,0}(\boldsymbol{\lambda}_0)$ and $\widetilde{G}_{n,1}(\boldsymbol{\lambda}_1)$, respectively.

2.3 Asymptotic Properties

This subsection is devoted to investigating the asymptotic properties of the proposed counterfactual distribution estimators $\widehat{F}_0(y)$ and $\widehat{F}_1(y)$. For ease of presentation, we first introduce some notations. For $d = 0, 1$, define

$$G_d^*(\boldsymbol{\lambda}_d) = \mathbb{E}\left[1(D_i = d)\rho\left(\mathbf{U}_d(X)^T \boldsymbol{\lambda}_d\right)\right] - \mathbb{E}\left[\mathbf{U}_d(X)^T \boldsymbol{\lambda}_d\right],$$

where $\mathbf{U}_d(x) = \left(\mathbf{F}_d(\mathbf{q}\,|\,x)^T, \mathbf{u}(x)^T\right)^T$, $\mathbf{F}_d(\mathbf{q}\,|\,x) = (F(q_1\,|\,x), \cdots, F(q_J\,|\,x))^T$, and $\mathbf{u}(x) = (u_1(x), \cdots, u_L(x))^T$. Clearly, $G_d^*(\boldsymbol{\lambda}_d)$ is the probability limit of $\widetilde{G}_{n,d}(\boldsymbol{\lambda}_d)$ in (9). For the probability space $(\mathcal{X}, \mathcal{A}, P)$ and a Borel function g defined on the space, let $\|g\|_{P,t} = \left(\int |g|^t \, dP\right)^{1/t}$ denote the $L_t(P)$ norm of g for $1 \leq t < \infty$, and $\sup_{x\in\mathcal{X}} |g|$ denotes its L_∞ norm. Before we embark on establishing the asymptotic results, all regularity conditions for asymptotic analysis are gathered together in the following.

Assumption 2 (Sampling Process)**.** *The data* $\left\{(Y_i(0), Y_i(1), D_i, X_i)\right\}_{i=1}^n$*, are independent and identically distributed (i.i.d).*

Assumption 3 (Distribution of X)**.** *(i) The support* $\mathcal{X}$ *of* p*-dimensional covariate* X *is a Cartesian product of* p *compact intervals; (ii) The density function* $f_X(x)$ *is bounded away from 0 on* $\mathcal{X}$*; (iii) The density function* $f_X(x)$ *is twice continuously differentiable within the interior of* $\mathcal{X}$*.*

Assumption 4 (Distribution of $Y(d)$)**.** *(i)* $Y(d)$ *has a compact support* $\mathcal{Y}$ *for* $d = 0$ *and* 1*; (ii) The distribution functions of potential outcomes,* $F_0(y)$ *and* $F_1(y)$*, are continuous on* $\mathcal{Y}$*; (iii) The density functions of potential outcomes,* $f_0(y)$ *and* $f_1(y)$*, are bounded away from 0 and twice continuously differentiable within the interior of* $\mathcal{Y}$*.*

Assumption 5 (Conditional distributions)**.** *(i) The propensity score function* $\pi(x)$ *is* r*-times continuously differentiable within the interior of* $\mathcal{X}$ *for* $r > p/2$*; (ii) The conditional distributions* $F_0(y\,|\,x)$ *and* $F_1(y\,|\,x)$ *are* r*-times differentiable with respect to* x *on the interior of* $\mathcal{X}$ *and infinitely differentiable with respect to* y *within the interior of* $\mathcal{Y}$*. The absolute value of the derivative* $F_d^{(J)}(\xi_y\,|\,x) := \frac{\partial^J F_d(y\,|\,x)}{\partial y^J}$ *is bounded by a positive constant* C_0 *that does not depend on* J *for* $d = 0, 1$*; (iii)* $F_d(y|x) \in \mathcal{F}_d$*, which is a set of conditional CDF functions satisfying* $\log N_{[]}(\varepsilon, \mathcal{F}_d, L_2) \leq C_1(1/\varepsilon)^{p/r}$ *for a positive constant* C_1 *and* $d = 0, 1$*, where* $N_{[]}(\varepsilon, \mathcal{F}, L_2)$ *represents the bracketing number of* $\mathcal{F}$ *with respect to the* L_2 *norm by* ε*-brackets.*

Assumption 6 (Kernel function)**.** *The kernel function* $K(u)$ *is bounded and satisfies: (i)* $\int K(u)du = 1$*; (ii)* $K(u) = K(-u)$*; (iii)* $\int |u^2 K(u)|\, du < \infty$*; (iv)* $K(u) = 0$ *if* $|u| > 1$*; (v)* $K(u)$ *is twice continuously differentiable with respect to* u *on its support.*

Assumption 7 (Bandwidth)**.** *As the sample size* n *goes to infinity, the bandwidth* h_d *for* $d = 0, 1$ *satisfies (i)* $h_d \to 0$*; (ii)* $n^{(1/2-s)} h_d^p / \log n_d \to \infty$*; (iii)* $n^{(1/4+s/2)} h_d^2 \to 0$ *for some* $0 < s < 1/4$*.*

Assumption 8 (Dispersion measure)**.** *(i) $\rho(\cdot)$ is a twice continuously differentiable and strictly concave function defined on a bounded set; (ii) The first derivative $\rho'(\cdot) > 0$, the second derivative $\rho''(\cdot) < 0$, and both derivatives are bounded away from zero.*

Assumption 9 (Balance functions)**.** *(i) Let $M = J + L$. The number of balance conditions satisfies $J = O(\log n)$ and $M = O(n^s)$; (ii) There exists positive constant C_2, C_3, and C_4 such that $\sup_{x\in\mathcal{X}} \|\mathbf{U}_d(x)\|_2 \leq C_2 M^{1/2}$, $\mathbb{E}\left\{\mathbf{U}_d(X)^T \mathbf{U}_d(X)\right\} \leq C_3$, $\nu_{\min}\left(\mathbb{E}\left[\mathbf{U}_d(X)\mathbf{U}_d(X)^T\right]\right) \geq C_4$, and $\nu_{\min}\left(\frac{1}{n}\sum_{i=1}^n 1(D_i = d)\mathbf{U}_d(X_i)\mathbf{U}_d(X_i)^T\right) \geq C_4$ for $d = 0, 1$, where $\nu_{\min}(\cdot)$ denote the minimum eigenvalue; (iii) Let $m_0^*(x) = (\rho')^{-1}(1/(1-\pi(x)))$ and $m_1^*(x) = (\rho')^{-1}(1/\pi(x))$. Then $m_d^*(\cdot) \in \mathcal{M}_d$, which is a set of functions satisfying $\log N_{[]}\{\varepsilon, \mathcal{M}_d, L_2(P)\} \leq C_5(1/\varepsilon)^{1/\nu}$ for some constants $C_5 > 0$, $\nu > 1/2$ and $d = 0, 1$. (iv) There exist $r_\pi > 1/(2s)$, $\boldsymbol{\lambda}_0^\dagger \in \mathbb{R}^M$ and $\boldsymbol{\lambda}_1^\dagger \in \mathbb{R}^M$ such that $\sup_{x\in\mathcal{X}}\left|m_0^*(x) - \mathbf{U}_0(x)^T\boldsymbol{\lambda}_0^\dagger\right| = O\left(M^{-r_\pi}\right)$ and $\sup_{x\in\mathcal{X}}\left|m_1^*(x) - \mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^\dagger\right| = O\left(M^{-r_\pi}\right)$.*

Assumption 10 (Optimization)**.** *(i) $G_d^*(\boldsymbol{\lambda}_d)$ has a unique maximizer $\boldsymbol{\lambda}_d^*$ for $d = 0$ and 1. (ii) There exists a constant $C_6 > 0$ such that $\sum_{j=1}^J(\lambda_{d,j}^*)^2 \leq C_6 J$ and $\sum_{j=1}^M(\lambda_{d,j}^*)^2 \leq C_6 M$.*

Assumption 2 requires the sampling process to be i.i.d, which is standard in many microeconometric applications. Assumptions 3 and 4 restrict the continuity and smoothness of the distributions of the covariates and potential outcomes. Assumption 5 imposes smoothness conditions on the propensity score and conditional distribution functions, respectively. Assumptions 6 and 7 provide conditions for the kernel function and its bandwidth. Assumption 8 assumes the smoothness and concavity of $\rho(\cdot)$, which is a transformation of the measure of dispersion of the weights $\phi(\cdot)$. This makes it possible to translate the consistency of $\widehat{\boldsymbol{\lambda}}_d$ into the consistency of weights. Assumption 9 collects conditions about balance functions. Assumption 9(i) imposes conditions on the growth rate of the number of balance functions relative to the number of observations. Assumption 9(ii) restricts the magnitude of the balance functions, which is a standard technical assumption similar to that in Assumption 2 of Newey (1997) and Assumption E.1.6 of Fan et al. (2023). Assumption 9(iii) assumes that $(\rho')^{-1}(\cdot)$ is the link function for the inverse propensity model and the corresponding systematic component $m_d^*(x)$ belongs to a functional class $\mathcal{M}_d$, whose complexity is restricted in a manner similar to Assumption 2(v) of Wang and Zubizarreta (2020) and Assumption E.1.7 of Fan et al. (2023). Assumption 9(iv) presumes that $m_d^*(x)$ can be well approximated by linear combination of the balance functions $\mathbf{U}_d(x)$, which is a weaker condition compared to

Assumption 1(vi) of Wang and Zubizarreta (2020). Wang and Zubizarreta (2020) assumed that $m_d^*(x)$ can be well approximated by $\mathbf{U}_d(x)^T\boldsymbol{\lambda}_d^*$, where $\boldsymbol{\lambda}_d^*$ is the unique maximizer of $G_d^*(\boldsymbol{\lambda}_d)$. In contrast, we do not require $\boldsymbol{\lambda}_d^\dagger$ to be the same as $\boldsymbol{\lambda}_d^*$. Finally, Assumption 10 provides some standard regularity conditions for consistency of minimum risk estimators. It aligns with Assumption 1(i) and (ii) of Wang and Zubizarreta (2020) and Assumption 3.1 of Fan et al. (2023).

Under the above assumptions, we present the asymptotic properties of the estimated conditional CDF $\widetilde{F}_d(y \,|\, X)$, the estimated weights, and the estimated distribution function $\widehat{F}_d(\cdot)$, respectively. The detailed theoretical proofs are provided in Appendix. First, we start with the asymptotic properties of $\widetilde{F}_d(y \,|\, x)$. The following proposition given in Rothe (2010) provides an explicit uniform convergence rate for the first-step estimator based on the kernel method.

Proposition 1. *Under Assumptions 1-7, we have*

$$\sup_{y\in\mathcal{Y}}\sup_{x\in\mathcal{X}}\left|\widetilde{F}_d(y|x) - F_d(y|x)\right| = O_p\left(\left(\frac{\log n}{nh_d^p}\right)^{1/2} + h_d^2\right). \tag{11}$$

Next, we show consistency of the optimal distribution balancing weights obtained in the second step. To this end, define $\widehat{w}_d(x) = \rho'\left(\widetilde{\mathbf{U}}_d(x)^T\widehat{\boldsymbol{\lambda}}_d\right)$ for $d = 0$ and 1. Then, the optimal distribution balancing weight $\widehat{w}_i = 1(D_i = 0)\widehat{w}_0(X_i) + 1(D_i = 1)\widehat{w}_1(X_i)$ according to (10). We further define $w_d^*(x) = \rho'\left(\sum_{d=0,1} 1(D_i = d)\mathbf{U}_d(x)^T\boldsymbol{\lambda}_d^*\right)$. The following proposition shows that $\widehat{w}_d(x)$ converges to $w_d^*(x)$ and the inverse propensity score weight in both L_2 and L_∞ norms.

Proposition 2. *Under Assumptions 1-10, we have*

$$\begin{aligned}
\sup_{x\in\mathcal{X}}|\widehat{w}_d(x) - w_d^*(x)| &= O_p\left(MJn^{-(1/4+s/2)}\right), \\
\sup_{x\in\mathcal{X}}\left|\widehat{w}_d(x) - \pi(x)^{-d}(1-\pi(x))^{-(1-d)}\right| &= O_p\left(MJn^{-(1/4+s/2)} + M^{1/2-r_\pi/2}\right), \\
\|\widehat{w}_d(x) - w_d^*(x)\|_{P,2} &= O_p\left(\sqrt{M}Jn^{-(1/4+s/2)}\right),
\end{aligned}$$

and

$$\left\|\widehat{w}_d(x) - \pi(x)^{-d}(1-\pi(x))^{-(1-d)}\right\|_{P,2} = O_p\left(\sqrt{M}Jn^{-(1/4+s/2)} + M^{-r_\pi/2}\right)$$

for $d = 0$ and 1.

Based on Propositions 1 and 2, we can derive the asymptotic properties of $\widehat{F}_d(y)$. The following theorem shows the limiting performance of $\widehat{F}_d(y)$, with its detailed proof relegated to Appendix. To prove the theorem, we first show that $\sqrt{n}\left[\widehat{F}_d(y) - F_d(y)\right]$ is asymptotically linear with an influence function representation, similar to the Bahadur representation for sample quantile, that is,

$$\sqrt{n}\left[\widehat{F}_d(y) - F_d(y)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_d^F(y, \mathbf{Z}_i) + o_p(1),$$

where

$$\psi_d^F(y, \mathbf{Z}) = \frac{1\{D = d\}\left[1\{Y \leq y\} - F_d(y \mid X)\right]}{\pi(X)^d[1 - \pi(X)]^{1-d}} + \left[F_d\left(y \mid X\right) - F_d(y)\right] \tag{12}$$

with $\mathbf{Z} = (Y, D, X^T)^T$. Then, the result in Theorem 1 holds by the functional central limit theorem[1].

Theorem 1. *Let* $\mathbf{y} = (y_0, y_1)^T$, $\mathbf{F}(\mathbf{y}) = \left(F_0\left(y_0\right), F_1\left(y_1\right)\right)^T$ *and* $\widehat{\mathbf{F}}(\mathbf{y}) = \left(\widehat{F}_0\left(y_0\right), \widehat{F}_1\left(y_1\right)\right)^T$. *Under Assumptions 1-10, uniformly for* $\mathbf{y} \in \mathcal{Y} \times \mathcal{Y}$, *we have*

$$\sqrt{n}\left[\widehat{\mathbf{F}}(\mathbf{y}) - \mathbf{F}(\mathbf{y})\right] \;\Rightarrow\; \mathbb{F}(\mathbf{y}),$$

where " $\Rightarrow$ " *denotes the weak convergence,* $\mathbb{F}(\mathbf{y}) = \left(\mathbb{F}_0\left(y_0\right), \mathbb{F}_1\left(y_1\right)\right)^T$ *is a two-dimensional Gaussian process with zero mean and covariance function*

$$\Psi^F\left(\mathbf{y}, \mathbf{y}'\right) = \mathbb{E}\left[\psi^F\left(\mathbf{y}, \mathbf{Z}\right)\psi^F\left(\mathbf{y}', \mathbf{Z}\right)^T\right],$$

and the convergence takes place in $\ell^\infty(\mathcal{Y}) \times \ell^\infty(\mathcal{Y})$, *where* $\ell^\infty(\mathcal{Y})$ *is the set of bounded functions over* $\mathcal{Y}$. *Here,* $\psi^F\left(\mathbf{y}, \mathbf{Z}\right) = \left(\psi_0^F\left(y_0, \mathbf{Z}\right), \psi_1^F\left(y_1, \mathbf{Z}\right)\right)^T$ *with* $\psi_d^F(y, \mathbf{Z})$ *defined in* (12).

Theorem 1 shows that $\widehat{\mathbf{F}}(\mathbf{y}) - \mathbf{F}(\mathbf{y})$ converges weakly to a mean-zero Gaussian process at the usual parametric rate of $\sqrt{n}$ despite the use of nonparametric estimators in the first step.

2.4 Bootstrap Inference

Theorem 1 provides a theoretical foundation for conducting inference for the counterfactual distribution functions, but it needs a consistent estimation of the covariance func-

[1]For the detailed definitions of Donsker class, the weak convergence, and the functional central limit theorem, the reader is referred to the book by Billingsley (1999).

tion $\Psi^F(\mathbf{y}, \mathbf{y}')$, which is often a cumbersome task in practice. Here, we propose an easily implemented nonparametric Bootstrap scheme for inference. First, the Bootstrap data $\{(Y_i^b, D_i^b, X_i^b)\}_{i=1}^n$ are drawn with replacement from the observed sample $\{(Y_i, D_i, X_i)\}_{i=1}^n$. Then, the Bootstrap data are used to calculate the Bootstrap estimates of $F_0(y_0)$ and $F_1(y_1)$, denoted by $\widehat{\mathbf{F}}^b(\mathbf{y}) = \left(\widehat{F}_0^b(y_0), \widehat{F}_1^b(y_1)\right)^T$, using the method described in Section 2.1, and the above procedure is repeated many times, for example, B =1000 times. The Bootstrap estimates $\widehat{\mathbf{F}}^1(\mathbf{y}), \cdots, \widehat{\mathbf{F}}^B(\mathbf{y})$ can be used for inference of $\widehat{\mathbf{F}}(\mathbf{y})$. Theorem 2 gives a theoretical justification for using this Bootstrap procedure with its proof presented in Appendix.

Theorem 2. *Under Assumptions 1-10, uniformly for* $\mathbf{y} \in \mathcal{Y} \times \mathcal{Y}$*, we have:*

$$\sqrt{n}\left[\widehat{\mathbf{F}}^b(\mathbf{y}) - \widehat{\mathbf{F}}(\mathbf{y})\right] \;\Rightarrow\; \mathbb{F}(\mathbf{y})$$

in probability, conditional on the data[2]*, where* $\mathbb{F}(\mathbf{y})$ *is the Gaussian process defined in Theorem 1.*

2.5 Semiparametric Estimation of Conditional CDF

It is well known in the nonparametric statistics literature that when the covariate dimension is relatively large but still finite (p does not depend on n), it is not desirable to estimate the conditional CDF by kernel method as in (4), due to the so-called "curse of dimensionality". To circumvent this problem, some semiparametric estimators can be employed. One can adopt the approach suggested in Aït-Shahalia and Brant (2001) and Hall and Yao (2005), which involves using an index model. Specifically, it assumes that there exists a $p \times 1$ vector $\boldsymbol{\gamma}_d$ so that $F_d(y \,|\, x) = F_d(y \,|\, x^T\boldsymbol{\gamma}_d)$. One can first estimate $\boldsymbol{\gamma}_d$ to approximate $F_d(y \,|\, x)$ by $F_d(y \,|\, x^T\boldsymbol{\gamma}_d)$ under a least-squares criterion, then, use $x^T\boldsymbol{\gamma}_d$ as the smooth variable to estimate the conditional CDF[3]. Another approach is to estimate the conditional CDF using quantile regression as in Koenker and Bassett (1978). This idea was also used in Melly (2006), Chernozhukov et al. (2013), and Cai et al. (2022). We present this method below in detail.

Let $Q_d(\tau \,|\, x) = \inf\{y : F_d(y \,|\, x) \geq \tau\}$ denote the conditional quantile function of $Y(d)$ conditional on $X = x$ at the quantile level $\tau \in (0, 1)$. Assume that $Q_d(\tau \,|\, x) = x^T\boldsymbol{\beta}_d(\tau)$ for

[2]See Section 3.6 in van der Vaart and Wellner (1996) for a precise definition of conditional weak convergence in probability.

[3]For details, the reader is referred to the paper by Hall and Yao (2005).

each quantile level τ. The coefficients $\boldsymbol{\beta}_d(\tau)$ can be estimated by

$$\widehat{\boldsymbol{\beta}}_d(\tau) = \arg\min_{\boldsymbol{\beta}_d(\tau)} \sum_{i=1}^{n} 1(D_i = d)\, \rho_\tau \left(Y_i - \boldsymbol{\beta}_d(\tau)^T X_i\right), \tag{13}$$

where $\rho_\tau(v) = v[\tau - 1(v \le 0)]$ is the so-called check function. Since $F_d(y \,|\, x) = \int_0^1 1\left(Q_d(\tau \,|\, x) \le y\right) d\tau$, the conditional CDF can be estimated by

$$\widehat{F}_d(y \,|\, x) = \varepsilon + \int_{\varepsilon}^{1-\varepsilon} 1\left(x^T \widehat{\boldsymbol{\beta}}_d(\tau) \le y\right) d\tau \approx \varepsilon + \sum_{j=2}^{S} (\tau_j - \tau_{j-1})\, 1\left(x^T \widehat{\boldsymbol{\beta}}_d(\tau_j) \le y\right), \tag{14}$$

where the trimming by ε avoids estimation of tail quantiles, and $\widehat{\boldsymbol{\beta}}_d(\tau)$ is estimated by (13) on a equally spaced mesh $\varepsilon = \tau_1 < \cdots < \tau_S = 1 - \varepsilon$.

Remark 2. Based on Proposition 5 in Chernozhukov et al. (2010), the conditional CDF estimators obtained by (14) are $\sqrt{n}$-consistent when the linear conditional quantile models are correctly specified and the mesh width is $o(n^{-1/2})$. In this case, the convergence rate of $\widehat{F}_d(y \,|\, x)$ is faster than that of the kernel estimator $\widetilde{F}_d(y \,|\, x)$ as stated in Proposition 1. Consequently, the conclusions in Theorems 1 and 2 still hold if we use the estimator $\widehat{F}_d(y \,|\, x)$ to replace $\widetilde{F}_d(y \,|\, x)$ in the first stage.

Remark 3. In the above, we only consider the case that p is finite. In some applications, p might be allowed to depend on the sample size so that the model setting in this section becomes the case with either high-dimensional ($p \to \infty$ but $p/n \to 0$) or ultra-high dimensional ($p \gg n$) covariates. For such cases, one might follow the idea in Cai et al. (2024) for a mean model to do some extensions, which are not straightforward and can be warranted as future research topics.

3 Applications

In this section, the proposed estimation method is applied to some interesting application problems. Specifically, we consider making inference of the QTE and testing the stochastic dominance relationship between the distributions of potential outcomes.

3.1 Inferences for Quantile Treatment Effect

Given a quantile level $\tau \in (0,1)$, the QTE is defined as

$$\Delta(\tau) = Q_1(\tau) - Q_0(\tau),$$

where $Q_d(\tau) = \inf\{y : F_d(y) \geq \tau\}$ for $d = 0$ and 1. Using the estimations $\widehat{F}_0(\cdot)$ and $\widehat{F}_1(\cdot)$ in (6), the QTE can be estimated as

$$\widehat{\Delta}(\tau) = \widehat{Q}_1(\tau) - \widehat{Q}_0(\tau), \tag{15}$$

where $\widehat{Q}_d(\tau) = \inf\left\{y : \widehat{F}_d(y) \geq \tau\right\}$. In practice, if we are only interested in $\Delta(\tau)$ for τ in a short closed interval $[a,b]$, we do not need to select grid points $q_1 < \cdots < q_J$ across the entire range of the observable outcomes. Instead, we only need to find a closed interval $[l_d, u_d] \supset [Q_d(a), Q_d(b)]$ and take grid points within this interval. To determine the interval, we suggest using $\widetilde{Q}_d(\cdot)$ as a reference, where $\widetilde{Q}_d(\tau) = \inf\left\{y : \widetilde{F}_d(y) \geq \tau\right\}$ with $\widetilde{F}_d(y) = \frac{1}{n}\sum_{i=1}^n \widetilde{F}_d\left(y \,|\, X_i\right)$. The grid points can be distributed on the interval $\left[\widetilde{Q}_d(a-\epsilon), \widetilde{Q}_d(b+\epsilon)\right]$ for some small $\epsilon > 0$. For example, in our simulation study, when estimating $\Delta(\tau)$ for a fixed τ, we take 5 equally spaced grid points on the interval $\left[\widetilde{Q}_d(\tau-0.1), \widetilde{Q}_d(\tau+0.1)\right]$.

Under the conclusions given in Theorem 1, we can obtain asymptotic properties of the QTE estimator according to the Bahadur representation as follows.

Proposition 3. *Under Assumptions 1-10, uniformly for $\tau \in [a,b]$ with $0 < a < b < 1$, we have*

$$\sqrt{n}\left[\widehat{\Delta}(\tau) - \Delta(\tau)\right] \;\Rightarrow\; \mathbb{Q}(\tau),$$

where $\mathbb{Q}(\tau)$ is a Gaussian process with zero mean and covariance function $\Psi^Q(\tau_1,\tau_2) = \mathbb{E}\left[\psi^Q(\tau_1)\,\psi^Q(\tau_2)\right]$ with

$$\psi^Q(\tau) = \left[\psi_1^F\left(Q_1(\tau), \mathbf{Z}\right)/f_1\left(Q_1(\tau)\right) - \psi_0^F\left(Q_0(\tau), \mathbf{Z}\right)/f_0\left(Q_0(\tau)\right)\right],$$

where $\psi_d^F(\cdot)$ is given in (12)*. The convergence takes place in $\ell^\infty([a,b])$.*

Proposition 3 shows that $\widehat{\Delta}(\tau) - \Delta(\tau)$ converges weakly to a zero-mean Gaussian process at the usual convergence rate of $\sqrt{n}$. Denote the Bootstrap estimate of $\Delta(\tau)$ as $\widehat{\Delta}^b(\tau)$, the following proposition shows that the nonparametric bootstrap in Section 2.4 is valid in this case as well.

Proposition 4. *Under Assumptions 1-10, uniformly for $\tau \in [a,b]$ with $0 < a < b < 1$, we have:*

$$\sqrt{n}\left[\widehat{\Delta}^b(\tau) - \widehat{\Delta}(\tau)\right] \;\Rightarrow\; \mathbb{Q}(\tau),$$

conditional on the data and in probability, where $\mathbb{Q}(\cdot)$ is a Gaussian process defined in Proposition 3.

Based on Proposition 4, the level $1-\alpha$ simultaneously confidence band for $\widehat{\Delta}(\tau)$ on interval $[a,b]$ can be constructed as

$$\mathrm{CB}(1-\alpha) = \left\{\widehat{\Delta}(\tau) - \widehat{c}_{1-\alpha}\widehat{\sigma}(\tau), \widehat{\Delta}(\tau) + \widehat{c}_{1-\alpha}\widehat{\sigma}(\tau) : \tau \in [a,b]\right\},$$

where $\widehat{\sigma}^2(\tau)$ is the sample variance of the Bootstrap estimates $\widehat{\Delta}^1(\tau), \cdots, \widehat{\Delta}^B(\tau)$, and

$$\widehat{c}_{1-\alpha} = \min\left\{c : \frac{1}{B}\sum_{b=1}^{B} 1\left(\sup_{\tau}\left|\hat{\Delta}^b(\tau) - \widehat{\Delta}(\tau)\right| / \widehat{\sigma}(\tau) \le c\right) \ge 1-\alpha\right\}.$$

Such a confidence band can be used to test whether the treatment effects differ along different quantile levels. The testing problem can be formulated as

$$H_0 : \Delta(\tau) = \Delta \quad \text{for all } \tau \in [a,b] \;\text{ versus }\; H_1 : \Delta(\tau) \neq \Delta \quad \text{for some } \tau \in [a,b],$$

where Δ is a pre-specified constant. For such a testing problem, we can reject H_0 if the constant line $\Delta(\tau) \equiv \Delta$ for $\tau \in [a,b]$ is not contained in the confidence band $\mathrm{CB}(1-\alpha)$.

3.2 Testing Stochastic Dominance

Making inferences regarding stochastic dominance relationship plays an important role in social sciences, with a vast amount of literature in economics, including but not limited to, Anderson (1996), Barrett and Donald (2003), Linton et al. (2005), Donald and Hsu (2016), Whang (2019), Linton et al. (2023), and references therein. Different from the existing literature, our focus is on testing the stochastic dominance relationship between the counterfactual distributions, which are not derived from any observable populations. Interestingly, Rothe (2010), Maier (2011), and Donald and Hsu (2014) also considered such a test in a similar scenario, but they used different methods to estimate the counterfactual distributions. Indeed, Rothe (2010) used the estimation method outlined in Remark 1, while Maier (2011) and Donald and Hsu (2014) estimated the counterfactual distributions by inverse propensity score weights. In this paper, we use $\widehat{F}_0(\cdot)$ and $\widehat{F}_1(\cdot)$ obtained in Section

2.1 to test stochastic dominance.

We only discuss test for the first order stochastic dominance (SD1) between the potential outcomes $Y(0)$ and $Y(1)$. To test if $Y(1)$ SD1 $Y(0)$, the hypothesis is formulated as

$$H_0 : F_1(y) \leq F_0(y) \quad \text{for all } y \in \mathcal{Y} \quad \text{versus} \quad H_1 : F_1(y) > F_0(y) \quad \text{for some } y \in \mathcal{Y}. \tag{16}$$

A commonly used statistic for testing the first order stochastic dominance is the Kolmogorov-Smirnov (KS) statistic, which is given by

$$\widehat{\text{KS}} = \sqrt{n} \sup_{y \in \mathcal{Y}} \left[\widehat{F}_1(y) - \widehat{F}_0(y) \right] = \sqrt{n} \max_{y \in \{Y_1, \cdots, Y_n\}} \left[\widehat{F}_1(y) - \widehat{F}_0(y) \right].$$

The second equality follows from the fact that both $\widehat{F}_1(y)$ and $\widehat{F}_0(y)$ are step function and their values change only at the observed Y_i, $i = 1, \cdots, n$.

Note that we are testing a composite null hypothesis. For this case, it is challenging to find the limit null distribution since the limit null distribution depends on the underlying distributions, while there are infinitely many different combinations of $F_1(\cdot)$ and $F_0(\cdot)$ satisfying the null hypothesis. The typical way to solve this problem is to find the least favorable configuration (LFC)[4] to construct an asymptotically valid test procedure based on Bootstrapping the test statistic similar to that in Barrett and Donald (2003). It is easy to see that the LFC in this context corresponds to $F_1(y) = F_0(y)$ for all $y \in \mathcal{Y}$. Let $\widehat{F}_1^b$ and $\widehat{F}_0^b$ be the Bootstrap estimates of the potential outcomes' distributions based on the same Bootstrap scheme as in Section 2.4. Then, the Bootstrap p-value can be calculated as

$$\widehat{p} = B^{-1} \sum_{b=1}^{B} 1 \left(\widehat{\text{KS}}^b > \widehat{\text{KS}} \right),$$

where

$$\widehat{\text{KS}}^b = \sqrt{n} \max_{y \in \{Y_1, \ldots Y_n\}} \left\{ \left[\widehat{F}_1^b(y) - \widehat{F}_0^b(y) \right] - \left[\widehat{F}_1(y) - \widehat{F}_0(y) \right] \right\}.$$

Thus, we reject H_0 if $\widehat{p}$ is less than the significance level α. The following proposition delivers theoretical justification for this approach and its proof is shown in Appendix.

Proposition 5. *Suppose that Assumptions 1-10 hold. If we reject H_0 when $\widehat{p} < \alpha$ for $\alpha < 1/2$, then, (i) Under H_0 defined in* (16)*, $\lim_{n\to\infty} P(\widehat{p} < \alpha) \leq \alpha$, and (ii) Under a fixed*

[4]Under a composite null hypothesis, the least favorable case is the distribution for which the null holds, but which is most difficult to distinguish from any distribution in the alternative hypothesis. See Section 3 in Lehmann and Romano (2005).

alternative hypothesis defined in (16), $\lim_{n\to\infty} P(\widehat{p} < \alpha) = 1$.

Proposition 5 implies that the size of the proposed test is asymptotically no larger than the pre-specified significance level α, and the power of the proposed test is asymptotically approaching 1 under the alternative hypothesis.

4 Monte Carlo Simulation Study

In this section, a series of Monte Carlo simulations are conducted to evaluate the performances of the proposed QTE estimator and stochastic dominance test in finite samples.

4.1 Performance of Estimating QTE

We consider several different settings of the data generating process (DGP) and compare the performance of our QTE estimator with other estimators. For a comparison, we use "DIQ" to denote the inconsistent estimator that takes difference in quantiles of the outcome variables for the treatment and control groups. Let "Initial" denote the estimation method described in Rothe (2010) and Hsu et al. (2022), which estimates the QTE by the difference in quantiles of $\widetilde{F}_0(y)$ and $\widetilde{F}_1(y)$ as discussed in Remark 1. We also consider two QTE estimators derived using the inverse propensity score weights, denoted as "PLE-IPW" and "SLE-IPW". For "PLE-IPW", the propensity score is estimated by a logit estimator, while for "SLE-IPW", the propensity score is estimated by a series logit estimator, which includes X, X^2 and the cross terms in the propensity score model. Now, we consider the estimator $\widehat{\Delta}(\tau)$ in (15). To calculate $\widehat{\Delta}(\tau)$, we first need to solve the constrained optimization problem in (5). We choose $\phi(w_i) = w_i \log(w_i)$ as the measure of weight dispersion[5] in the objective function. Three different sets of constraints are considered: (1) only balance the first moment of covariate; that is, only keep the balance condition for $\mathbf{u}(x) = x$ in the constraint, but remove the balance condition for $\widetilde{\mathbf{F}}_d(\mathbf{q} \,|\, x)$. Such balance conditions are often used for estimating the ATE as in Hainmueller (2012); (2) only balance the estimated conditional distribution functions $\widetilde{\mathbf{F}}_d(\mathbf{q} \,|\, x)$; (3) balance both the estimated conditional distribution functions and the first moment of covariate. The estimators corresponding to these three different sets of

[5]We also run simulations by taking the stable balancing variance as in Zubizarreta (2015) as the dispersion measure. It does not significantly affect the estimation performance. Thus, we omit this part and the results are available upon request.

balance conditions are denoted as "Epy 1", "Epy 2", and "Epy 3", respectively. Note that only "Epy 2" and "Epy 3" use the balance conditions proposed in this paper.

We consider estimating the QTE for quantile levels 0.25, 0.50, and 0.75. For Examples 1 and 2, the Nadaraya-Watson estimator is used to estimate the conditional CDF $F_d(y \mid x)$. The Epanechnikov kernel is used and its bandwidth is set as $h_d = c \cdot n_d^{-1/5}$ for $c = 0.5$, 0.75 or 1.0. For Example 3, the dimension of X is set as $p = 8$, so that we use the quantile regression method presented in Section 2.5 to estimate the conditional CDF. For each example, we repeat the experiment 1000 times independently. The median of absolute deviation errors (MADE) and the root mean squared errors (RMSE) are reported for performance measurement. When the closed-form expression of the QTE is not available, we employ a simulation with a large sample size to obtain its true value.

Example 1: We consider the Skorohod representation[6] for the potential outcomes $Y(0)$ and $Y(1)$. Let the DGP be

$$Y = Y(1)D + Y(0)(1-D) \text{ with } Y(0) = 3X_1 + 0.4\sqrt{U_0}X_2 \text{ and } Y(1) = 4X_1 + 1.6\sqrt{U_1}X_2,$$

where $D \mid X \sim$ binomial$(\pi(X))$, U_0, U_1, X_1 and X_2 are independently drawn from the uniform $U(0,1)$ distribution, and the propensity score function is set as

$$\pi(X) = \exp\left(-1.5 + X_1 + X_2\right)\left\{1 + \exp\left(-1.5 + X_1 + X_2\right)\right\}^{-1}.$$

The estimation results are shown in Table 1. The "DIQ" estimator always performs the worst, which illustrates the importance of adjusting covariates. When the sample size increases from $n = 250$ to $n = 1000$, the MADEs and RMSEs of all other estimators decrease approximately by half, suggesting that the convergence rates are $\sqrt{n}$. Including the estimated conditional distribution functions in the balancing conditions, "Epy 2" and "Epy 3" consistently outperform other estimators. By adding moments into the balancing conditions, "Epy 3" performs slightly better than "Epy 2". Also, it is noted that the performance of "Initial" greatly depends on the bandwidth, but "Epy 2" and "Epy 3" are less affected, which is in line with our discussion in Remark 1.

Example 2: In this example, we use a nonlinear propensity score model

$$\pi(X) = \frac{\exp\left(-4 + \exp(X_1/2) + 1.5(X_1 + X_2)^2 + \sin(X_2)\right)}{1 + \exp\left(-4 + \exp(X_1/2) + 1.5(X_1 + X_2)^2 + \sin(X_2)\right)},$$

[6]For the definition, please see the book by Durret (2019).

Table 1: MADEs and RMSEs of estimating QTEs for Example 1

		$h_d = 0.5n^{-1/5}$			$h_d = 0.75n^{-1/5}$			$h_d = n^{-1/5}$		
		$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$
τ		MADE	MADE	MADE	MADE	MADE	MADE	MADE	MADE	MADE
	DIQ	0.4108	0.4115	0.4004	0.4108	0.4115	0.4004	0.4108	0.4115	0.4004
	Initial	0.0797	0.0539	0.0359	0.1017	0.0768	0.0590	0.1382	0.1086	0.0831
	PLE-IPW	0.1278	0.1046	0.1019	0.1278	0.1046	0.1019	0.1278	0.1046	0.1019
0.25	SLE-IPW	0.0781	0.0566	0.0360	0.0781	0.0566	0.0360	0.0781	0.0566	0.0360
	Epy 1	0.1027	0.0676	0.0481	0.1027	0.0676	0.0481	0.1027	0.0676	0.0481
	Epy 2	**0.0705**	**0.0478**	**0.0319**	**0.0752**	**0.0507**	**0.0368**	**0.0845**	**0.0549**	**0.0419**
	Epy 3	**0.0658**	**0.0462**	**0.0307**	**0.0625**	**0.0472**	**0.0336**	**0.0794**	**0.0533**	**0.0378**
	DIQ	0.4809	0.4870	0.4832	0.4809	0.4870	0.4832	0.4809	0.4870	0.4832
	Initial	0.0712	0.0518	0.0350	0.0947	0.0688	0.0498	0.1320	0.0949	0.0749
	PLE-IPW	0.1128	0.1063	0.0988	0.1128	0.1063	0.0988	0.1128	0.1063	0.0988
0.5	SLE-IPW	0.0814	0.0634	0.0419	0.0814	0.0634	0.0419	0.0814	0.0634	0.0419
	Epy 1	0.0858	0.0646	0.0438	0.0858	0.0646	0.0438	0.0858	0.0646	0.0438
	Epy 2	**0.0627**	**0.0461**	**0.0327**	**0.0670**	**0.0496**	**0.0305**	**0.0742**	**0.0517**	**0.0364**
	Epy 3	**0.0616**	**0.0440**	**0.0314**	**0.0649**	**0.0467**	**0.0292**	**0.0703**	**0.0491**	**0.0324**
	DIQ	0.3553	0.3514	0.3573	0.3553	0.3514	0.3573	0.3553	0.3514	0.3573
	Initial	0.0650	0.0437	0.0324	0.0834	0.0536	0.0369	0.1042	0.0706	0.0481
	PLE-IPW	0.1069	0.0852	0.0767	0.1069	0.0852	0.0767	0.1069	0.0852	0.0767
0.75	SLE-IPW	0.0666	0.0482	0.0346	0.0666	0.0482	0.0346	0.0666	0.0482	0.0346
	Epy 1	0.0952	0.0678	0.0464	0.0952	0.0678	0.0464	0.0952	0.0678	0.0464
	Epy 2	**0.0591**	**0.0423**	**0.0292**	**0.0655**	**0.0444**	**0.0307**	**0.0649**	**0.0479**	**0.0334**
	Epy 3	**0.0516**	**0.0419**	**0.0291**	**0.0644**	**0.0433**	**0.0302**	**0.0638**	**0.0472**	**0.0324**
		$h_d = 0.5n^{-1/5}$			$h_d = 0.75n^{-1/5}$			$h_d = n^{-1/5}$		
		$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$
		RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE
	DIQ	0.4729	0.4406	0.4173	0.4729	0.4406	0.4173	0.4729	0.4406	0.4173
	Initial	0.1214	0.0805	0.0541	0.1458	0.1017	0.0765	0.1881	0.1414	0.1031
	PLE-IPW	0.1775	0.1411	0.1192	0.1775	0.1411	0.1192	0.1775	0.1411	0.1192
0.25	SLE-IPW	0.1220	0.0858	0.0550	0.1220	0.0858	0.0550	0.1220	0.0858	0.0550
	Epy 1	0.1561	0.1026	0.0703	0.1561	0.1026	0.0703	0.1561	0.1026	0.0703
	Epy 2	**0.1069**	**0.0712**	**0.0471**	**0.1090**	**0.0728**	**0.0527**	**0.1183**	**0.0786**	**0.0542**
	Epy 3	**0.1055**	**0.0701**	**0.0459**	**0.1061**	**0.0692**	**0.0490**	**0.1120**	**0.0780**	**0.0531**
	DIQ	0.5309	0.5079	0.4925	0.5309	0.5079	0.4925	0.5309	0.5079	0.4925
	Initial	0.1063	0.0748	0.0514	0.1328	0.0968	0.0679	0.1714	0.1256	0.0934
	PLE-IPW	0.1633	0.1405	0.1195	0.1633	0.1405	0.1195	0.1633	0.1405	0.1195
0.5	SLE-IPW	0.1200	0.0917	0.0627	0.1200	0.0917	0.0627	0.1200	0.0917	0.0627
	Epy 1	0.1229	0.0933	0.0640	0.1229	0.0933	0.0640	0.1229	0.0933	0.0640
	Epy 2	**0.0968**	**0.0657**	**0.0479**	**0.0987**	**0.0727**	**0.0486**	**0.1102**	**0.0766**	**0.0488**
	Epy 3	**0.0915**	**0.0620**	**0.0434**	**0.0955**	**0.0697**	**0.0462**	**0.1052**	**0.0714**	**0.0477**
	DIQ	0.3977	0.3766	0.3687	0.3977	0.3766	0.3687	0.3977	0.3766	0.3687
	Initial	0.0989	0.0703	0.0470	0.1204	0.0806	0.0549	0.1466	0.1028	0.0694
	PLE-IPW	0.1565	0.1232	0.1024	0.1565	0.1232	0.1024	0.1565	0.1232	0.1024
0.75	SLE-IPW	0.1004	0.0749	0.0513	0.1004	0.0749	0.0513	0.1004	0.0749	0.0513
	Epy 1	0.1400	0.1027	0.0680	0.1400	0.1027	0.0680	0.1400	0.1027	0.0680
	Epy 2	**0.0924**	**0.0637**	**0.0426**	**0.0995**	**0.0667**	**0.0440**	**0.1066**	**0.0701**	**0.0486**
	Epy 3	**0.0929**	**0.0636**	**0.0417**	**0.0929**	**0.0653**	**0.0439**	**0.1050**	**0.0694**	**0.0474**

and all other settings are the same as in Example 1. Through simulation, one can observe that approximately 13.59% of the samples have a propensity score less than 0.1, and 6.95% have a propensity score greater than 0.9. Therefore, we anticipate that the IPW estimators may not perform well for this example. The estimation results are reported in Table 2. It can be seen that "PLE-IPW" has large MADEs and RMSEs due to misspecification of the propensity score model. "Epy 2" and "Epy 3" significantly outperform "PLE-IPW" and "SLE-IPW", especially when the sample size is small, which is in accordance with our anticipation. Other conclusions are similar to those as in Example 1.

Example 3: In this example, we consider the case of $p = 8$. Let the DGP be

$$Y(0) = \sum_{j=1}^{7} \boldsymbol{\beta}_j X_j + 0.4\sqrt{U_0}X_8 \text{ and } Y(1) = \sum_{j=1}^{7} \boldsymbol{\beta}_j X_j + 1.6\sqrt{U_1}X_8,$$

where $D \sim$ binomial$(\pi(X))$, U_0, U_1 and $X_1, \cdots, X_8$ are independently drawn from the uniform $U(0,1)$, $\boldsymbol{\beta} = (-2, 0.75, -1, 1.5, -2, 0.75, -1)$, and the true propensity score model is:

$$\pi(X) = \frac{\exp\left(-2 + \sum_{j=1}^{8} \boldsymbol{\gamma}_j X_j\right)}{1 + \exp\left(-2 + \sum_{j=1}^{8} \boldsymbol{\gamma}_j X_j\right)},$$

where $\boldsymbol{\gamma} = (2, -0.75, 1, -1.5, 2, -0.75, 1, -1.5)$. For this example, the kernel estimator for the conditional distribution functions is not applicable, so the quantile regression method presented in Section 2.5 is used. The simulation results are reported in Table 3. It can be seen that "Epy 2" performs better than "DIQ", "Initial", "PLE-IPW" and "Epy 1", and is close to "SLE-IPW". By adding the moments into the balancing conditions, "Epy 3" performs better than "Epy 2", and can outperform "SLE-IPW".

For all three examples, Table 4 reports the realized coverage rates (CR) of point-wise confidence intervals with nominal coverage level 90% for the proposed QTE estimators "Epy 2" and "Epy 3", using the Bootstrap method proposed in Section 3.1 with $B = 1000$. It can be seen that as the sample size increases to $n = 1000$, the proposed Bootstrap method can produce confidence intervals with good coverage probabilities.

4.2 Performance of Testing Stochastic Dominance

In this section, we use simulations to demonstrate the size and power of the stochastic dominance test proposed in Section 3.2. We estimate the counterfactual distributions using "Epy 2" and "Epy 3" as described in Section 4.1, respectively. Our proposed tests are

Table 2: MADEs and RMSEs of estimating QTEs for Example 2

		$h_d = 0.5n^{-1/5}$			$h_d = 0.75n^{-1/5}$			$h_d = n^{-1/5}$		
		$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$
τ		MADE	MADE	MADE	MADE	MADE	MADE	MADE	MADE	MADE
0.25	DIQ	1.1801	1.1702	0.9549	1.1801	1.1702	0.9549	1.1801	1.1702	0.9549
	Initial	0.1601	0.0933	0.0475	0.2041	0.1580	0.1173	0.3411	0.2730	0.1613
	PLE-IPW	0.3848	0.3834	0.2752	0.3848	0.3834	0.2752	0.3848	0.3834	0.2752
	SLE-IPW	0.1418	0.0916	0.0505	0.1418	0.0916	0.0505	0.1418	0.0916	0.0505
	Epy 1	0.1331	0.0992	0.0726	0.1331	0.0992	0.0726	0.1331	0.0992	0.0726
	Epy 2	**0.1114**	**0.0673**	**0.0401**	**0.1067**	**0.0801**	**0.0480**	**0.1120**	**0.0832**	**0.0484**
	Epy 3	**0.1174**	**0.0507**	**0.0383**	**0.1023**	**0.0658**	**0.0453**	**0.1016**	**0.0718**	**0.0428**
0.5	DIQ	1.1918	1.1910	1.0195	1.1918	1.1910	1.0195	1.1918	1.1910	1.0195
	Initial	0.1172	0.0629	0.0341	0.1909	0.1365	0.0824	0.3274	0.2527	0.1303
	PLE-IPW	0.3384	0.3358	0.2711	0.3384	0.3358	0.2711	0.3384	0.3358	0.2711
	SLE-IPW	0.1153	0.0924	0.0643	0.1153	0.0924	0.0643	0.1153	0.0924	0.0643
	Epy 1	0.1072	0.0787	0.0590	0.1072	0.0787	0.0590	0.1072	0.0787	0.0590
	Epy 2	**0.0784**	**0.0660**	**0.0519**	**0.0878**	**0.0703**	**0.0571**	**0.0951**	**0.0711**	**0.0654**
	Epy 3	**0.0786**	**0.0572**	**0.0474**	**0.0833**	**0.0570**	**0.0431**	**0.0948**	**0.0675**	**0.0536**
0.75	DIQ	0.9834	0.9939	0.8503	0.9834	0.9939	0.8503	0.9834	0.9939	0.8503
	Initial	0.1816	0.1366	0.0790	0.2560	0.2035	0.1656	0.3660	0.3011	0.1964
	PLE-IPW	0.3420	0.3214	0.2717	0.3420	0.3214	0.2717	0.3420	0.3214	0.2717
	SLE-IPW	0.1100	0.0879	0.0526	0.1100	0.0879	0.0526	0.1100	0.0879	0.0526
	Epy 1	0.1185	0.0942	0.0799	0.1185	0.0942	0.0799	0.1185	0.0942	0.0799
	Epy 2	**0.0915**	**0.0578**	**0.0489**	**0.0851**	**0.0547**	**0.0452**	**0.0877**	**0.0580**	**0.0455**
	Epy 3	**0.0852**	**0.0506**	**0.0471**	**0.0785**	**0.0450**	**0.0447**	**0.0842**	**0.0560**	**0.0427**
		$h_d = 0.5n^{-1/5}$			$h_d = 0.75n^{-1/5}$			$h_d = n^{-1/5}$		
		$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$	$n = 250$	$n = 500$	$n = 1000$
		RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE
0.25	DIQ	1.1932	1.1800	0.9578	1.1932	1.1800	0.9578	1.1932	1.1800	0.9578
	Initial	0.2222	0.1391	0.0690	0.2598	0.1938	0.1389	0.3806	0.2942	0.1746
	PLE-IPW	0.4181	0.4014	0.2867	0.4181	0.4014	0.2867	0.4181	0.4014	0.2867
	SLE-IPW	0.2386	0.1477	0.0779	0.2386	0.1477	0.0779	0.2386	0.1477	0.0779
	Epy 1	0.1983	0.1463	0.1057	0.1983	0.1463	0.1057	0.1983	0.1463	0.1057
	Epy 2	**0.1689**	**0.1065**	**0.0591**	**0.1590**	**0.1175**	**0.0729**	**0.1682**	**0.1237**	**0.0737**
	Epy 3	**0.1844**	**0.0979**	**0.0582**	**0.1538**	**0.1019**	**0.0683**	**0.1582**	**0.1035**	**0.0661**
0.5	DIQ	1.2011	1.1972	1.0226	1.2011	1.1972	1.0226	1.2011	1.1972	1.0226
	Initial	0.1646	0.0966	0.0521	0.2228	0.1578	0.1013	0.3508	0.2659	0.1434
	PLE-IPW	0.3870	0.3620	0.2825	0.3870	0.3620	0.2825	0.3870	0.3620	0.2825
	SLE-IPW	0.2154	0.1463	0.0933	0.2154	0.1463	0.0933	0.2154	0.1463	0.0933
	Epy 1	0.1631	0.1153	0.0867	0.1631	0.1153	0.0867	0.1631	0.1153	0.0867
	Epy 2	**0.1259**	**0.0938**	**0.0697**	**0.1290**	**0.1013**	**0.0774**	**0.1396**	**0.1023**	**0.0862**
	Epy 3	**0.1281**	**0.0842**	**0.0636**	**0.1232**	**0.0911**	**0.0709**	**0.1427**	**0.0994**	**0.0790**
0.75	DIQ	0.9948	0.9981	0.8554	0.9948	0.9981	0.8554	0.9948	0.9981	0.8554
	Initial	0.2153	0.1567	0.0922	0.2742	0.2163	0.1726	0.3812	0.3073	0.2024
	PLE-IPW	0.3800	0.3484	0.2846	0.3800	0.3484	0.2846	0.3800	0.3484	0.2846
	SLE-IPW	0.1791	0.1298	0.0620	0.1791	0.1298	0.0620	0.1791	0.1298	0.0620
	Epy 1	0.1701	0.1317	0.1062	0.1701	0.1317	0.1062	0.1701	0.1317	0.1062
	Epy 2	**0.1468**	**0.0878**	**0.0527**	**0.1280**	**0.0810**	**0.0519**	**0.1331**	**0.0886**	**0.0539**
	Epy 3	**0.1339**	**0.0839**	**0.0524**	**0.1219**	**0.0755**	**0.0427**	**0.1267**	**0.0792**	**0.0479**

Table 3: MADEs and RMSEs of estimating QTEs for Example 3

		$n = 250$		$n = 500$		$n = 1000$	
τ		MADE	RMSE	MADE	RMSE	MADE	RMSE
	DIQ	0.9499	0.9394	0.9401	0.8934	0.9509	0.9192
	Initial	0.1924	0.2593	0.1479	0.1768	0.1207	0.1376
	PLE-IPW	0.2150	0.3376	0.1811	0.2470	0.1726	0.2107
0.25	SLE-IPW	0.1143	0.1856	0.0665	0.1032	0.0458	0.0696
	Epy 1	0.1357	0.2066	0.0871	0.1288	0.0706	0.1002
	Epy 2	**0.1057**	**0.1607**	**0.0688**	**0.1006**	**0.0577**	**0.0790**
	Epy 3	**0.0987**	**0.1507**	**0.0581**	**0.0882**	**0.0399**	**0.0591**
	DIQ	0.9732	0.9763	0.9675	0.9571	0.9764	0.9643
	Initial	0.2285	0.2610	0.1676	0.1876	0.1415	0.1507
	PLE-IPW	0.1216	0.2139	0.0973	0.1450	0.0893	0.1187
0.5	SLE-IPW	0.1002	0.1512	0.0588	0.0898	0.0449	0.0640
	Epy 1	0.1001	0.1505	0.0646	0.0989	0.0483	0.0709
	Epy 2	**0.0837**	**0.1240**	**0.0523**	**0.0799**	**0.0395**	**0.0585**
	Epy 3	**0.0843**	**0.1230**	**0.0505**	**0.0746**	**0.0339**	**0.0502**
	DIQ	0.9937	1.0160	0.9762	0.9741	0.9787	0.9675
	Initial	0.1971	0.2288	0.1510	0.1689	0.1234	0.1329
	PLE-IPW	0.0922	0.1383	0.0620	0.0915	0.0422	0.0656
0.75	SLE-IPW	0.0784	0.1203	0.0495	0.0710	0.0363	0.0508
	Epy 1	0.1192	0.1684	0.0833	0.1157	0.0638	0.0903
	Epy 2	**0.0822**	**0.1241**	**0.0528**	**0.0773**	**0.0445**	**0.0617**
	Epy 3	**0.0797**	**0.1197**	**0.0485**	**0.0679**	**0.0296**	**0.0463**

Table 4: Realized coverage rates for Examples 1, 2 and 3 with the nominal coverage level is 90%

		$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
	n	Epy 2	Epy 3	Epy 2	Epy 3	Epy 2	Epy 3
	250	0.929	0.934	0.927	0.931	0.928	0.934
Example 1	500	0.908	0.911	0.909	0.914	0.910	0.919
	1000	0.897	0.898	0.900	0.901	0.899	0.901
	250	0.932	0.936	0.915	0.917	0.906	0.909
Example 2	500	0.917	0.921	0.911	0.916	0.903	0.905
	1000	0.901	0.903	0.900	0.901	0.897	0.898
	250	0.946	0.952	0.937	0.941	0.926	0.932
Example 3	500	0.934	0.941	0.924	0.927	0.916	0.925
	1000	0.906	0.910	0.907	0.909	0.905	0.911

compared with the "SLE-IPW" method, which estimates counterfactual distributions using the IPW approach, with the propensity score being estimated by a series logit estimator, and obtains the critical value using nonparametric Bootstrap. We also compare our tests with the method proposed by Donald and Hsu (2014), denoted as "SLE-IPW-m", which also estimates the propensity score by the series logit estimator but derives the critical value using the multiplier Bootstrap.

In the testing procedure, when use the method proposed in Section 2.1 to estimate $F_d(y)$, the Epanechnikov kernel is used and the bandwidth is set as $h_d = 0.5 \cdot n^{-1/(4+p)}$ for $d = 0$ and 1. We also set $J = 50$ and let $q_1 < q_2 < \cdots < q_{50}$ be 50 equally spaced points across the entire range of the observable outcomes. The number of Bootstrap repetitions is set to $B = 1000$. For each simulation, the rejected rates are estimated by conducting 1000 independent experiments. The results are reported in Table 5.

Example 4: We use the same data-generating process as in Example 1. In this setting, $F_1(y) < F_0(y)$ for all $y \in \mathcal{Y}$. It is the case that the null hypothesis $H_0 : Y(1)\ SD1\ Y(0)$ holds. This example is used to examine the test size of the proposed test. The top panel in Table 5 shows that all testing methods have rejection rates close to zero across all considered sample sizes.

Example 5: In this example, we consider the case that $F_1(y) = F_0(y)$ for all $y \in \mathcal{Y}$. Following Donald and Hsu (2014), let the DGP be

$$X = 0.3 + 0.4U_x, \quad Y(0) = 1\left(U_0 \leq X\right) U_0^2/X + 1\left(U_0 > X\right) U_0,$$
$$Y(1) = 1\left(U_1 \leq 1 - X\right) U_1^2/(1-X)^2 + 1\left(U_1 > 1 - X\right) U_1,$$

where $D = 1\left(U < X\right)$ and U_x, U_0, U_1 and U are independent uniform distributions over $[0, 1]$. The middle panel in Table 5 demonstrates that both "SLE-IPW" and "SLE-IPW-m" are conservative when the sample size is small, but their empirical sizes get closer to the nominal level as the sample size increases. Compared to "SLE-IPW" and "SLE-IPW-m", our testing methods exhibit slightly higher empirical sizes that are closer to the nominal levels at the same sample size. As the sample size increases, the empirical sizes of our methods also converge to the nominal level.

Example 6: To examine the power of our proposed tests, we consider the case such that the null hypothesis does not hold. Following Donald and Hsu (2014), let the potential outcomes

be generated by

$$Y(0) = U_0, \quad Y(1) = 1\left(U_1 \leq X\right) U_1^2/X + 1\left(U_1 > X\right) U_1.$$

In this example, it can be shown that $F_1(y) > F_0(y)$ for all $0 < y < 0.7$ and $F_1(y) = F_0(y)$ for all $y \geq 0.7$. The bottom panel in Table 5 illustrates that the powers for all test methods increase with the increase of sample size. When $n = 1000$, the powers of our proposed test are exactly 1, which is consistent with the results in Proposition 5. Compared to "SLE-IPW" and "SLE-IPW-m", the powers of our test are larger under the same sample size.

Table 5: Rejection rates of testing stochastic dominance for Examples 4 (the top panel), 5 (the middle panel) and 6 (the bottom panel)

		$\alpha = 0.05$				$\alpha = 0.10$			
	n	SLE-IPW	SLE-IPW-m	**Epy 2**	**Epy 3**	SLE-IPW	SLE-IPW-m	**Epy 2**	**Epy 3**
	250	0.000	0.000	**0.000**	**0.000**	0.000	0.000	**0.000**	**0.000**
4	500	0.000	0.000	**0.000**	**0.000**	0.000	0.000	**0.000**	**0.000**
	1000	0.000	0.000	**0.000**	**0.000**	0.000	0.000	**0.000**	**0.000**
	250	0.041	0.032	**0.068**	**0.060**	0.078	0.080	**0.110**	**0.112**
5	500	0.042	0.044	**0.058**	**0.052**	0.077	0.096	**0.106**	**0.102**
	1000	0.052	0.054	**0.052**	**0.050**	0.108	0.102	**0.103**	**0.101**
	250	0.502	0.540	**0.556**	**0.554**	0.674	0.696	**0.724**	**0.724**
6	500	0.838	0.830	**0.884**	**0.840**	0.938	0.936	**0.952**	**0.952**
	1000	0.997	0.996	**1.000**	**1.000**	0.999	1.000	**1.000**	**1.000**

5 An Empirical Example

In this section, we apply the proposed methods to analyze the distributional effect of maternal smoking during pregnancy on infant birth weight. Low birth weight causes a range of subsequent adverse health problems and increased health care costs and is also linked to later educational attainment and labor market outcomes as addressed in Almond et al. (2005), Black et al. (2007) and references therein. Therefore, this issue should be of concern to policy makers. Smoking is generally recognized as one of the main modifiable risk factors for low birth weight, and many studies have attempted to estimate its causal effect as investigated by Abrevaya (2006), Abrevaya et al. (2015), and Tang et al. (2021).

5.1 Data Description

We use a subsample of the dataset published by the North Carolina State Center Health Services. Abrevaya et al. (2015) and Tang et al. (2021) gave a more extensive study to the full data set. The subsample we are using was recorded in 2002, which includes 26, 739 first-time white mothers who gave birth to a live single baby, 3960 (14.81%) of whom smoked during pregnancy. For each mother, we record whether she smoked, the infant's birth weight (in grams), and other 14 variables that might confound the relationship between birth weight and the mother's smoking decision. These variables include the mother's age, education, the month of first prenatal visit, the number of prenatal visits, weight gain during the pregnancy (in pounds), and indicators for the baby's gender, the mother's marital status, whether or not the father's age is missing, gestational diabetes, hypertension, amniocentesis, ultrasound exams, previous (terminated) pregnancies, and alcohol use. Table 6 presents some descriptive statistics for this dataset. Figure 1 shows the kernel density plots of birth

Table 6: Descriptive Statistics

	Mean	Sd	Max	75%	Median	25%	Min
Birth weight	3334.45	582.94	5840.01	3713.79	3373.59	3033.40	198.45
Smoking	0.15	0.36	-	-	-	-	-
Mother's age	25.72	5.86	46	30	26	21	12
Mother's weight gain	34.51	13.27	98	42	34	25	1
Mother's education	13.68	2.41	17	16	14	12	2
1st Prenatal	2.07	1.07	9	2	2	1	0
# Prenatal	13.17	3.43	49	15	13	11	0
Male baby	0.52	0.50	1	-	-	-	0
Mother's marital status	0.73	0.45	1	-	-	-	0
Miss father's age	0.10	0.30	1	-	-	-	0
Diabetes	0.02	0.16	1	-	-	-	0
Hypertension	0.08	0.27	1	-	-	-	0
Amniocentesis	0.01	0.11	1	-	-	-	0
Ultra sound exams	0.79	0.41	1	-	-	-	0
# Terminated	0.25	0.62	12	0	0	0	0
Alcohol use	0.01	0.08	1	-	-	-	0

NOTE: Birth weight is measured in grams, mother's weight gain is measured in pounds, mother's education is measured in years, and "# Prenatal" denotes the number of prenatal visits.

weight for smoking and non-smoking mothers. It can be seen from Figure 1 that the two distributions of the infant weight for smoking and non-smoking mothers are different. Table

7 gives the results for the symmetry test. Both the density plots and symmetry tests indicate that birth weight distributions are skewed to the left with a thick tail. Therefore, in addition to the average treatment effect, the distributional effect of maternal smoking on infant birth weight can provide a better understanding on how smoking has an effect on the infant weight, in particular, on low infant weights.

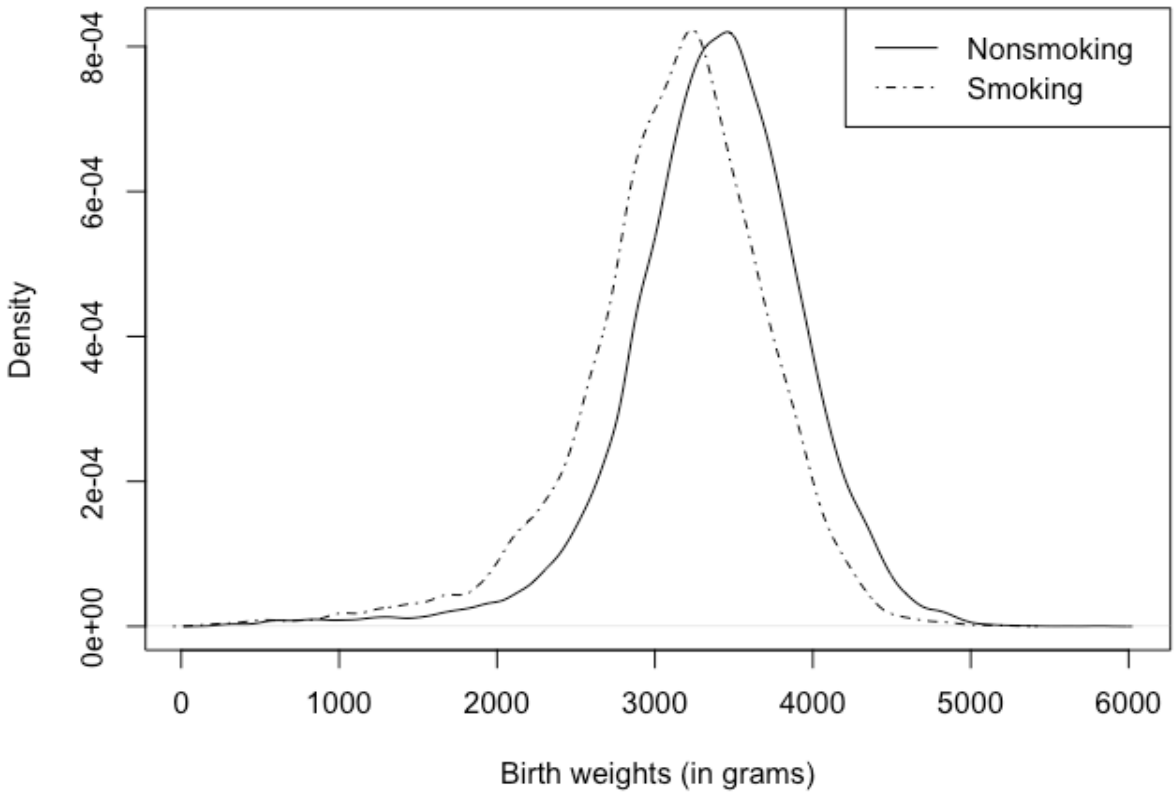

Figure 1: The density plots of infant birth weights for nonsmoking mother (solid line) and smoking mother (dot-dashed line).

Table 7: Symmetry Test Results

	Non-smoking	Smoking
Average birth weight	3372.34	3116.53
Skewness	-0.88	-0.85
Kurtosis	5.96	5.14
Symmetry test (p-value)	0.00	0.00
Number of observations	22779	3960

5.2 Empirical Results

To assess the performance of the proposed estimator, we consider three different scenarios to estimate the quantile treatment effects. In the first two scenarios, we generate the pseudo population by mimicking the original data to know the true QTE and compare the performance of our method with other methods. In the last scenario, we use all original data to revisit the effect of maternal smoking on the infant's birth weight.

Scenario 1: We consider three covariates, the same as those in Rothe (2010), which are the mother's age, weight gain during the pregnancy (in pounds), and whether the mother is married. To mimic the original data, we keep the covariates X and the treatment variable D, and generate potential outcomes by $Y(0) = 2911.19 + 4.16X_1 + 9.92X_2 + 64.23X_3 + \varepsilon_0$, and $Y(1) = 2758.78 + 2.31X_1 + 10.32X_2 + 143.96X_3 + \varepsilon_1$, where ε_0 and ε_1 are independently drawn from $N(0, 100)$, the coefficients are derived from the linear regression based on the original data. Then, the observable outcomes are obtained by letting $Y = DY(1) + (1 - D)Y(0)$ and thus, we can get a population data $\{Y_i, Y_i(1), Y_i(0), X_i, D_i\}_{i=1}^N$ with $N = 26,739$. The true quantile treatment effects are plotted in Figure 2 (the left panel). We draw n samples from this population independently and consider $n = 1000, 2000, 4000$. The kernel method is used to estimate conditional CDF, and the bandwidths are set as $h_d = 2n_d^{-1/7}$ for $d = 0$ and 1. Same as in Section 4.1, we compare our estimators, "Epy 2" and "Epy 3", against "DIQ", "Initial", "PLE-IPW", "SLE-IPW" and "Epy 1". The estimation results are reported in Table 8, from which, it can be seen that our estimators always outperform the others.

Scenario 2: Following Abrevaya et al. (2015), we consider 13 covariates, which are the mother's age, education, month of first prenatal visit, number of prenatal visits, and indicators for the baby's gender, the mother's marital status, whether or not the father's age is missing, gestational diabetes, hypertension, amniocentesis, ultrasound exams, previous (terminated) pregnancies, and alcohol use. Again, we keep the covariates X and the treatment variable D. The potential outcomes are generated by $Y(0) = 2705.85 + X^T\boldsymbol{\beta}_0 + \varepsilon_0$ and $Y(1) = 2245.04 + X^T\boldsymbol{\beta}_1 + \varepsilon_1$, where ε_0 and ε_1 are independently drawn from $N(0, 100)$, $\boldsymbol{\beta}_0^T = (-2.17, 12.12, 35.27, 30.89, 125.91, 37.77, -41.29, -24.63, -276.69, -109.09, 12.33, -22.57, 101.37)$, and $\boldsymbol{\beta}_1^T = (-7.81, 32.72, 42.26, 40.03, 134.51, 35.77, 1.90, 72.01, -166.84, -164.85, -9.27, -19.70, 120.76)$. The true quantile treatment effects are plotted in Figure 2 (the right figure). We estimate conditional CDF by quantile regression. The estimation results are shown in Table 9. In this scenario, our estimators still consistently outperform

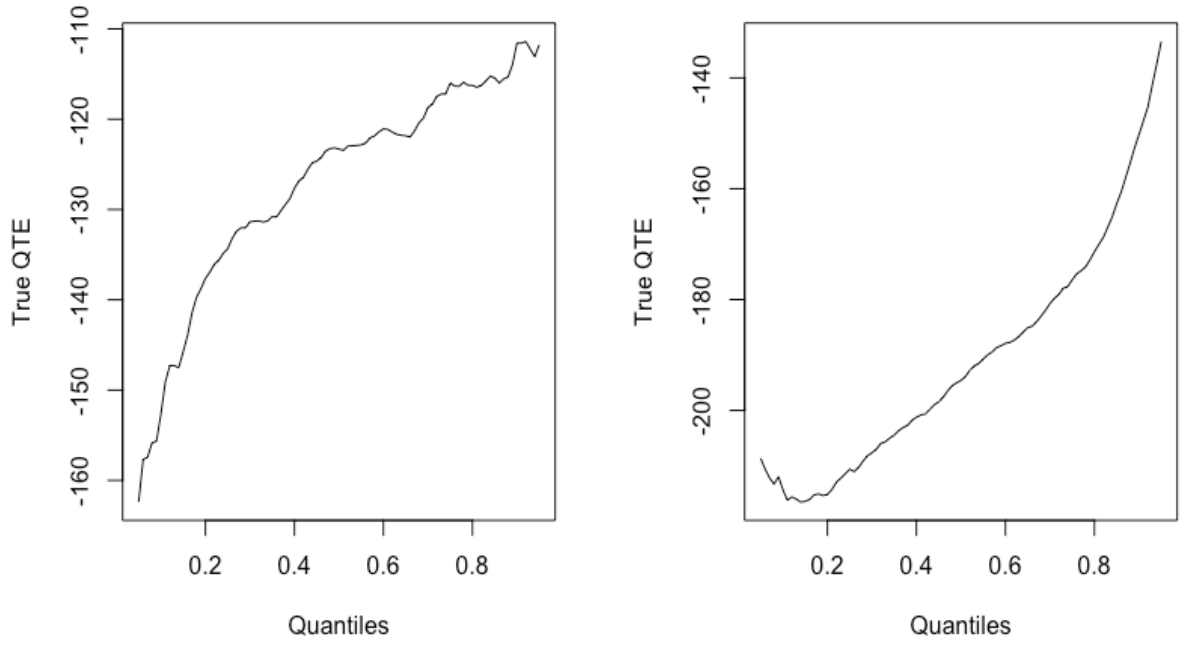

Figure 2: True quantile treatment effects in Scenario 1 (left panel) and in Scenario 2 (right panel).

the other estimators.

Scenario 3: Finally, we use the original data and consider all 14 covariates. We consider $\tau \in \{0.1, 0.2, 0.3, \cdots, 0.9\}$ and estimate the conditional CDF by quantile regression. The quantile effect of maternal smoking on the baby's birth weight is estimated by our method. Figure 3 shows the estimation of the average treatment effect[7] (-209.73, dot-dashed line) and quantile treatment effect (solid line) together with simultaneously 90% confidence bands (light-gray area) via the Bootstrap with $B = 1000$ replications. The graph shows that maternal smoking was associated with lower birth weight at all the quantiles considered. Compared with the average treatment effect, the quantile treatment effects estimates suggest that low quantiles are significantly affected by maternal smoking particularly. This becomes evident when we observe that the confidence bands for the quantile treatment effects at the lower percentiles do not cover the average treatment effect.

In this scenario, we also use the method proposed in Section 3.2 to test if the birth

[7]The average treatment effect is estimated by parametric logistic model based on the aforementioned covariates.

Table 8: Estimation Results for Scenario 1

		$n = 1000$		$n = 2000$		$n = 4000$	
τ		MADE	RMSE	MADE	RMSE	MADE	RMSE
	DIQ	71.1362	72.7177	72.5827	72.5084	72.7478	72.6937
	Initial	9.3515	13.4836	6.6527	9.3800	4.8205	6.7126
	PLE-IPW	19.9127	25.0024	21.8630	23.5871	21.9972	22.7151
0.25	SLE-IPW	13.0841	20.2038	10.4075	15.6709	11.7496	13.6737
	Epy 1	20.7864	25.0819	21.0398	23.5260	22.3840	22.9411
	Epy 2	**8.6908**	**13.5351**	**5.3497**	**8.1676**	**3.2366**	**5.1271**
	Epy 3	**8.2059**	**12.6224**	**4.8978**	**8.0106**	**3.0685**	**5.0920**
	DIQ	56.6225	58.4369	55.9834	57.2361	57.1878	57.1799
	Initial	7.4049	11.2639	5.2672	7.5013	3.2460	4.5456
	PLE-IPW	11.6670	17.5323	8.3084	11.3372	5.6194	8.0114
0.5	SLE-IPW	11.4304	20.3233	8.2674	12.0240	5.1498	7.7315
	Epy 1	10.1178	14.9034	7.3763	10.1802	4.9897	7.1223
	Epy 2	**7.1303**	**11.2102**	**4.5268**	**6.9034**	**3.1676**	**4.2732**
	Epy 3	**6.9942**	**10.9769**	**4.3322**	**6.5027**	**3.0836**	**4.1245**
	DIQ	30.3957	35.9620	28.8384	32.8794	28.4996	30.3816
	Initial	8.2750	12.4776	4.8916	7.5742	3.3907	4.9336
	PLE-IPW	15.4748	23.5728	11.7182	18.3076	12.3547	16.3067
0.75	SLE-IPW	12.6980	24.0745	8.7673	15.2646	7.9955	12.1334
	Epy 1	12.1515	19.0018	10.4945	15.6024	10.4828	13.8020
	Epy 2	**7.8813**	**11.9347**	**4.5421**	**6.9916**	**2.8515**	**4.4476**
	Epy 3	**7.5608**	**11.5842**	**4.3293**	**6.7073**	**2.5328**	**4.2188**

NOTE: The true coefficients are -134.61, -123.65, -116.06

weight of infants not exposed to maternal smoking first-order stochastically dominates that of infants exposed to it; that is to test $H_0 : F_0(y) \leq F_1(y)$. The resulting p-value of 0.79 via the Bootstrap with $B = 1000$ replication indicates the presence of this first-order dominance relationship.

6 Conclusion

This paper introduces a new method for estimating counterfactual distribution functions based on distributional balancing to avoid the shortcomings of inverse propensity weights. This method firstly estimates the conditional CDF for the treated and untreated groups respectively, and then finds the weights of minimum dispersion that exactly balance the estimated conditional CDF among the treated, the untreated, and the combined group.

Table 9: Estimation Results for Scenario 2

		$n = 1000$		$n = 2000$		$n = 4000$	
τ		MADE	RMSE	MADE	RMSE	MADE	RMSE
	DIQ	68.2421	70.4947	66.4765	68.1778	67.8673	68.1378
	Initial	5.0503	8.3076	4.1248	5.1308	2.7201	4.4752
	PLE-IPW	15.7962	24.5687	15.6598	19.4673	16.8518	18.5306
0.25	SLE-IPW	18.0379	37.4340	13.0730	27.5936	9.5234	15.3803
	Epy1	11.9371	19.2539	9.0711	13.4618	6.1905	9.2917
	Epy2	**3.3889**	**5.1491**	**2.6188**	**3.5079**	**1.7616**	**2.4892**
	Epy3	**2.9104**	**5.0456**	**2.4575**	**3.3540**	**1.5936**	**2.2394**
	DIQ	71.2544	72.5369	68.7540	70.5491	69.8058	70.3107
	Initial	3.3848	5.5056	2.6730	2.4991	1.2304	2.7589
	PLE-IPW	18.0137	24.2818	15.7130	18.9276	14.7804	16.9366
0.5	SLE-IPW	25.4774	44.9860	15.9828	28.2061	9.6983	14.9023
	Epy1	12.5762	17.8912	8.6639	13.0944	6.8314	9.2330
	Epy2	**3.3407**	**5.0234**	**2.6626**	**3.5630**	**1.5336**	**2.2149**
	Epy3	**3.2068**	**4.6921**	**2.3892**	**3.1718**	**1.4914**	**2.1994**
	DIQ	67.5348	70.8074	68.1728	69.1696	67.4725	67.6378
	Initial	3.0072	5.3259	2.4609	4.1770	1.9968	2.5586
	PLE-IPW	20.3868	28.6860	16.0568	21.9713	14.7413	17.7386
0.75	SLE-IPW	27.9042	48.0812	19.6444	31.8026	12.1789	17.3446
	Epy1	14.8753	22.7379	12.2730	15.5138	7.9605	10.7027
	Epy2	**3.4066**	**5.1225**	**2.5106**	**4.0738**	**1.9204**	**2.8216**
	Epy3	**3.2803**	**4.9420**	**2.4368**	**3.8983**	**1.5990**	**2.4658**

NOTE: The true coefficients are -201.36, -182.85, -165.36

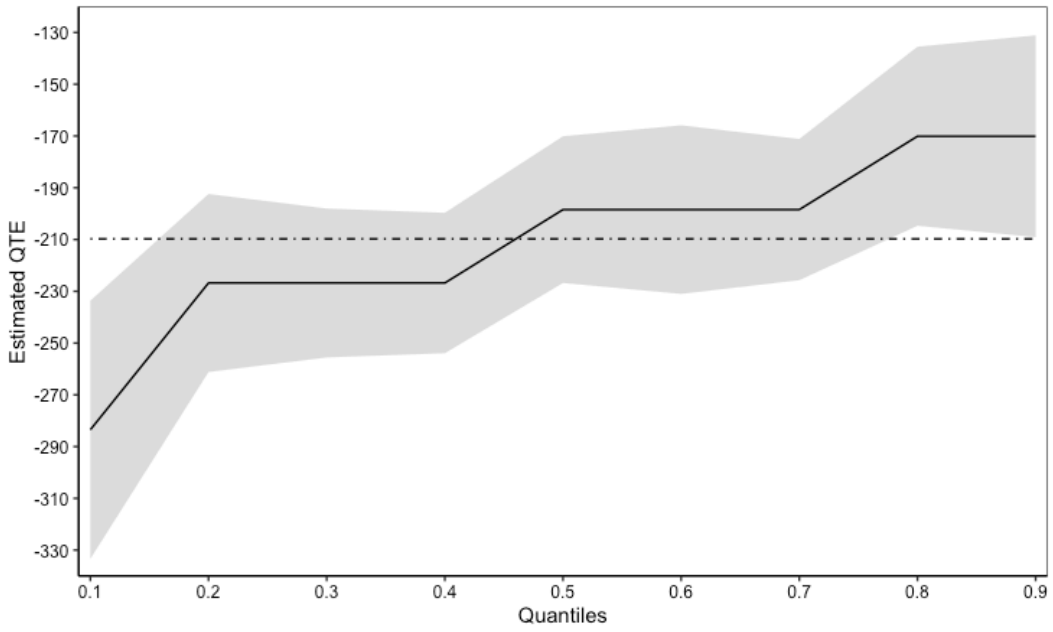

Figure 3: The average treatment effect (dot-dashed line) and quantile treatment effects (solid line) of maternal smoking on infant birth weight, with simultaneously 90% confidence bands (light-gray area) based on 1000 Bootstrap replications.

These weights are used to estimate the counterfactual distribution functions. We show that our estimator converges weakly to a Gaussian process with zero mean at the usual parametric rate of $\sqrt{n}$ and a properly designed Bootstrap method can be used to obtain confidence intervals and conduct inference, together with its theoretical justification. With the estimates of counterfactual distribution functions, we also provide the methods and theories to estimate the quantile treatment effects and test the stochastic dominance relationship between the potential outcome distributions. Monte Carlo simulations demonstrate that our estimator performs better than the inverse propensity weighting estimator in many scenarios. The empirical study revisits the effect of maternal smoking on infant birth weights.

As mentioned earlier, some extensions might be interesting, which can be warranted as future research topics. For example, when covariate X_i is high-dimensional (either $p = p_n \to \infty$ but $p_n/n \to 0$ or $p_n \gg n$), it is worth to investigate the asymptotic theory for such cases. Additionally, note that this paper considers only iid data and it should be extended to the time series context.

Disclosure Statement

We claim that this work is original and has not been published elsewhere nor is it currently under consideration for publication elsewhere and, also, we declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Also, the authors declare that they do not use any generative AI and AI-assisted technologies in the writing process.

References

Abrevaya, J. (2006). Estimating the effect of smoking on birth outcomes using a matched panel data approach. *Journal of Applied Econometrics 21*(4), 489–519.

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics 33*(4), 485–505.

Aït-Shahalia, Y. and M. W. Brant (2001). Variable selection for portfolio choice. *Journal of Finance 56*(4), 1297–1351.

Almond, D., K. Y. Chay, and D. S. Lee (2005). The costs of low birth weight. *Quarterly Journal of Economics 120*(3), 1031–1083.

Anderson, G. (1996). Nonparametric tests of stochastic dominance in income distributions. *Econometrica 65*(4), 1183–1193.

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B 80*(4), 597–623.

Barrett, G. F. and S. G. Donald (2003). Consistent tests for stochastic dominance. *Econometrica 71*(1), 71–104.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley, New Year.

Black, S. E., P. J. Devereux, and K. G. Salvanes (2007). From the cradle to the labor market? the effect of birth weight on adult outcomes. *Quarterly Journal of Economics 122*(1), 409–439.

Cai, Z., Y. Fang, M. Lin, and Z. Wu (2024). A quasi synthetic control method for nonlinear models with high-dimensional covariates. *Statistica Sinica*.

Cai, Z., Y. Fang, M. Lin, and M. Zhan (2022). Estimating quantile treatment effects for panel data. *Working Paper*, Department of Economics, University of Kansas.

Chan, K. C. G., S. C. P. Yam, and Z. Zhang (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society, Series B 78*(3), 673.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*(3), 1093–1125.

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

Donald, S. G. and Y.-C. Hsu (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics 178*, 383–397.

Donald, S. G. and Y.-C. Hsu (2016). Improving the power of tests of stochastic dominance. *Econometric Reviews 35*(4), 553–585.

Durret, R. (2019). *Probability: Theory and Examples, 5th Edition*. Cambridge University Rress, New York.

Fan, J., K. Imai, I. Lee, H. Liu, Y. Ning, and X. Yang (2023). Optimal covariate balancing conditions in propensity score estimation. *Journal of Business & Economic Statistics 41*(1), 97–110.

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

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 315–331.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis 20*(1), 25–46.

Hall, P. and Q. Yao (2005). Approximating conditional distribution functions using dimension reduction. *Annals of Statistics 33*(3), 1404–1421.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies 64*(4), 605–654.

Hirano, K., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

Hsu, Y.-C., T.-C. Lai, and R. P. Lieli (2022). Counterfactual treatment effects: Estimation and inference. *Journal of Business & Economic Statistics 40*(1), 240–255.

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B 76*(1), 243–263.

Josey, K. P., E. Juarez-Colunga, F. Yang, and D. Ghosh (2021). A framework for covariate balance using Bregman distances. *Scandinavian Journal of Statistics 48*(3), 790–816.

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

LaLonde, R. J. (1995). The promise of public sector-sponsored training programs. *Journal of Economic Perspectives 9*(2), 149–168.

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*, Volume 3. Springer.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association 113*(521), 390–400.

Linton, O., E. Maasoumi, and Y.-J. Whang (2005). Consistent testing for stochastic dominance under general sampling schemes. *Review of Economic Studies 72*(3), 735–765.

Linton, O., M. H. Seo, and Y.-J. Whang (2023). Testing stochastic dominance with many conditioning variables. *Journal of Econometrics 235*(2), 507–527.

Maier, M. (2011). Tests for distributional treatment effects under unconfoundedness. *Economics Letters 110*(1), 49–51.

Melly, B. (2006). Estimation of counterfactual distributions using quantile regression. *Working Paper,* Swiss Institute for International Economics and Applied Economic Research (SIAW), University of St. Gallen.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Ning, Y., P. Sida, and K. Imai (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika 107*(3), 533–554.

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

Rothe, C. (2010). Nonparametric estimation of distributional policy effects. *Journal of Econometrics 155*(1), 56–70.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

Süli, E. and D. F. Mayers (2003). *An Introduction to Numerical Analysis*. Cambridge University press.

Tang, S., Z. Cai, Y. Fang, and M. Lin (2021). A new quantile treatment effect model for studying smoking effect on birth weight during mother's pregnancy. *Journal of Management Science and Engineering 6*(3), 336–343.

Tropp, J. A. et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning 8*(1-2), 1–230.

Tseng, P. and D. P. Bertsekas (1987). Relaxation methods for problems with strictly convex separable costs and linear constraints. *Mathematical Programming 38*(3), 303–321.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer.

Wang, Y. and J. R. Zubizarreta (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika 107*(1), 93–105.

Whang, Y.-J. (2019). *Econometric Analysis of Stochastic Dominance: Concepts, Methods, Tools, and Applications*. Cambridge University Press.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Annals of Statistics 47*(2), 965–993.

Zhao, Q. and D. Percival (2017). Entropy balancing is doubly robust. *Journal of Causal Inference 5*(1), 1–19.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*(511), 910–922.

Appendix to "Estimating Counterfactual distribution functions via Optimal Distribution Balancing with Applications"

Zongwu Cai[a], Ying Fang[b,c], Ming Lin[b,c], and Yaqian Wu[d]

[a]Department of Economics, University of Kansas, Lawrence, KS 66045, USA

[b]Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian 361005, China

[c]Department of Statistics and Data Sciences, Xiamen University, Xiamen, Fujian 361005, China

[d]School of Economics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

A Proofs of Propositions 1 and 2

Proof of Proposition 1: The result is given by Proposition 2 in Rothe (2010). □

Next, we will provide some lemmas for proving Proposition 2. Let $\|\cdot\|_2$ denote the spectral norm of a matrix or the L_2 norm of a vector.

Lemma 1. (Bernstein's Inequality.) *Let $\{\mathbf{A}_i\}$ be a sequence of independent random matrices with dimensions $d_1 \times d_2$. Suppose that $\mathbb{E}[\mathbf{A}_i] = 0$ and $\|\mathbf{A}_i\|_2 \leq R_n$ almost surely. Define*

$$\sigma_n^2 = \max\left\{\left\|\sum_{i=1}^n \mathbb{E}\left(\mathbf{A}_i\mathbf{A}_i^T\right)\right\|_2, \quad \left\|\sum_{i=1}^n \mathbb{E}\left(\mathbf{A}_i^T\mathbf{A}_i\right)\right\|_2\right\}.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n \mathbf{A}_i\right\|_2 \geq t\right) \leq (d_1+d_2)\exp\left(-\frac{t^2/2}{\sigma_n^2 + R_n t/3}\right).$$

Proof of Lemma 1: The result is given by Tropp et al. (2015). □

Lemma 2. *Under Assumptions 1-7 and 9, one has*

$$\nu_{\min}\left(\frac{1}{n}\sum_{i=1}^n 1(D_i = d)\widetilde{\mathbf{U}}_d(X_i)\widetilde{\mathbf{U}}_d(X_i)^T\right) \geq b_1$$

in probability for some constant $b_1 > 0$. Here $d = 0$ or 1.

Proof of Lemma 2: First, we have

$$
\begin{aligned}
&\nu_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\widetilde{\mathbf{U}}_d(X_i)\widetilde{\mathbf{U}}_d(X_i)^T\right)\\
\geq&\nu_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\mathbf{U}_d(X_i)\mathbf{U}_d(X_i)^T\right)\\
&+\nu_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\widetilde{\mathbf{U}}_d(X_i)\widetilde{\mathbf{U}}_d(X_i)^T-\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\mathbf{U}_d(X_i)\mathbf{U}_d(X_i)^T\right)\\
=&\nu_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\mathbf{U}_d(X_i)\mathbf{U}_d(X_i)^T\right)\\
&-\nu_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\mathbf{U}_d(X_i)\mathbf{U}_d(X_i)^T-\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\widetilde{\mathbf{U}}_d(X_i)\widetilde{\mathbf{U}}_d(X_i)^T\right)\\
\geq&C_4-\left\|\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\mathbf{U}_d(X_i)\mathbf{U}_d(X_i)^T-\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\widetilde{\mathbf{U}}_d(X_i)\widetilde{\mathbf{U}}_d(X_i)^T\right\|_2, \qquad \text{(A.1)}
\end{aligned}
$$

where the first inequality follows from Weyl's inequality, and the second inequality holds since $\nu_{\min}\left(\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\mathbf{U}_d(X_i)\mathbf{U}_1(X_i)^T\right)\geq C_4$ by Assumption 9 (ii), and for any real symmetric matrix $\mathbf{A}$, $\nu_{\max}(\mathbf{A})\leq\sqrt{\nu_{\max}\left(\mathbf{A}^T\mathbf{A}\right)}=\|\mathbf{A}\|_2$.

Next, we focus on the second term in the above equation. We have

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\mathbf{U}_d(X_i)\mathbf{U}_d(X_i)^T-\frac{1}{n}\sum_{i=1}^{n}1(D_i=d)\widetilde{\mathbf{U}}_d(X_i)\widetilde{\mathbf{U}}_d(X_i)^T\right\|_2\\
\leq&\sup_i\left\|\mathbf{U}_d(X_i)\mathbf{U}_d(X_i)^T-\widetilde{\mathbf{U}}_d(X_i)\widetilde{\mathbf{U}}_d(X_i)^T\right\|_2\\
\leq&\sup_i\left\|\left[\widetilde{\mathbf{U}}_d(X_i)-\mathbf{U}_d(X_i)\right]\left[\widetilde{\mathbf{U}}_d(X_i)-\mathbf{U}_d(X_i)\right]^T\right\|_2+\sup_i\left\|\left[\widetilde{\mathbf{U}}_d(X_i)-\mathbf{U}_d(X_i)\right]\mathbf{U}_d(X_i)^T\right\|_2\\
&+\sup_i\left\|\mathbf{U}_d(X_i)\left[\widetilde{\mathbf{U}}_d(X_i)-\mathbf{U}_d(X_i)\right]^T\right\|_2\\
=&\sup_i\left\|\widetilde{\mathbf{U}}_d(X_i)-\mathbf{U}_d(X_i)\right\|_2^2+2\sup_i\left(\left\|\widetilde{\mathbf{U}}_d(X_i)-\mathbf{U}_d(X_i)\right\|_2\|\mathbf{U}_d(X_i)\|_2\right)\\
\leq&J\sup_i\sup_j\left|\widetilde{F}(q_j\,|\,X_i)-F(q_j\,|\,X_i)\right|^2+2\sqrt{J}\sup_i\sup_j\left|\widetilde{F}(q_j\,|\,X_i)-F(q_j\,|\,X_i)\right|\sup_i\|\mathbf{U}_d(X_i)\|_2\\
=&O_p\left(J\left(\left(\frac{\log n}{nh_d^p}\right)^{1/2}+h_d^2\right)^2\right)+O_p\left(J^{1/2}M^{1/2}\left(\left(\frac{\log n}{nh_d^p}\right)^{1/2}+h_d^2\right)\right)\\
=&o_p(1). \qquad \text{(A.2)}
\end{aligned}
$$

The last equality holds according to Assumptions 7 and 9. Finally, combining (A.1) and

(A.2) leads to the result in Lemma 2. □

Lemma 3. *For $d = 0$ and 1, let $\widetilde{G}'_{n,d}(\boldsymbol{\lambda}_d^*)$ denote the first derivative of $\widetilde{G}_{n,d}(\boldsymbol{\lambda}_d)$ at $\boldsymbol{\lambda}_d^* = \arg\max_{\boldsymbol{\lambda}_d} G_d^*(\boldsymbol{\lambda}_d)$. Under Assumptions 1-10, we have*

$$\left\|\widetilde{G}'_{n,d}(\boldsymbol{\lambda}_d^*)\right\|_2 = O_p\left(\sqrt{M}Jn^{-(1/4+s/2)}\right).$$

Proof of Lemma 3: Recall that

$$G_d^*(\boldsymbol{\lambda}_d) = \mathbb{E}\left[1(D_i = d)\rho\left(\mathbf{U}_d(X)^T\boldsymbol{\lambda}_d\right)\right] - \mathbb{E}\left[\mathbf{U}_d(X)^T\boldsymbol{\lambda}_d\right],$$

and

$$\widetilde{G}_{n,d}(\boldsymbol{\lambda}_d) = \frac{1}{n}\sum_{i=1}^n 1(D_i = d)\rho\left(\widetilde{\mathbf{U}}_d(X_i)^T\boldsymbol{\lambda}_d\right) - \frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{U}}_d(X_i)^T\boldsymbol{\lambda}_d.$$

By the mean value theorem, we have

$$\begin{aligned}
\widetilde{G}'_{n,d}(\boldsymbol{\lambda}_d^*) =& \frac{1}{n}\sum_i 1(D_i = d)\rho'\left(\widetilde{\mathbf{U}}_d(X_i)^T\boldsymbol{\lambda}_d^*\right)\widetilde{\mathbf{U}}_d(X_i) - \frac{1}{n}\sum_i \widetilde{\mathbf{U}}_d(X_i) \\
=& \frac{1}{n}\sum_i \left\{1(D_i = d)\left[\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) + \rho''\left(\xi_i\right)\left[\widetilde{\mathbf{U}}_d(X_i) - \mathbf{U}_d(X_i)\right]^T\boldsymbol{\lambda}_d^*\right]\right. \\
& \left.\times\left[\widetilde{\mathbf{U}}_d(X_i) - \mathbf{U}_d(X_i) + \mathbf{U}_d(X_i)\right]\right\} - \frac{1}{n}\sum_i\left[\widetilde{\mathbf{U}}_d(X_i) - \mathbf{U}_d(X_i) + \mathbf{U}_d(X_i)\right] \\
=& \frac{1}{n}\sum_i\left[1(D_i = d)\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) - 1\right]\mathbf{U}_d(X_i) \\
& + \frac{1}{n}\sum_i\left[1(D_i = d)\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) - 1\right]\left[\widetilde{\mathbf{U}}_d(X_i) - \mathbf{U}_d(X_i)\right] \\
& + \frac{1}{n}\sum_i 1(D_i = d)\rho''\left(\xi_i\right)\left[\widetilde{\mathbf{U}}_d(X_i) - \mathbf{U}_d(X_i)\right]^T\boldsymbol{\lambda}_d^*\left[\widetilde{\mathbf{U}}_d(X_i) - \mathbf{U}_d(X_i)\right] \\
& + \frac{1}{n}\sum_i 1(D_i = d)\rho''\left(\xi_i\right)\left[\widetilde{\mathbf{U}}_d(X_i) - \mathbf{U}_d(X_i)\right]^T\boldsymbol{\lambda}_d^*\mathbf{U}_d(X_i) \\
:=& I_1 + I_2 + I_3 + I_4,
\end{aligned} \tag{A.3}$$

where ξ_i is a point between $\widetilde{\mathbf{U}}_d(X_i)^T\boldsymbol{\lambda}_d^*$ and $\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*$. In the following, we will consider the orders of the terms I_1 to I_4.

We first consider the term I_1. Define

$$\mathbf{A}_i = \frac{1}{n}\left[1(D_i = d)\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) - 1\right]\mathbf{U}_d(X_i),$$

which is an $M \times 1$ vector. Note that $\boldsymbol{\lambda}_d^*$ is the unique maximizer of $G_d^*(\boldsymbol{\lambda}_d)$, which implies

$$(G_d^*)'(\boldsymbol{\lambda}_d^*) = \mathbb{E}\left\{\left[1(D=d)\rho'\left(\mathbf{U}_d(X)^T\boldsymbol{\lambda}_d^*\right) - 1\right]\mathbf{U}_d(X)\right\} = 0. \tag{A.4}$$

Hence, $\mathbb{E}[\mathbf{A}_i] = 0$. We also have

$$\begin{aligned}\|\mathbf{A}_i\|_2 =& \left\|\frac{1}{n}\left[1(D_i=d)\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) - 1\right]\mathbf{U}_d(X_i)\right\|_2 \\ \leq& \frac{1}{n}\sup_i\left|1(D_i=d)\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) - 1\right|\sup_{x\in\mathcal{X}}\|\mathbf{U}_d(x)\|_2 \leq \frac{b_2\sqrt{M}}{n},\end{aligned} \tag{A.5}$$

$$\begin{aligned}\left\|\sum_{i=1}^n \mathbb{E}\left(\mathbf{A}_i^T\mathbf{A}_i\right)\right\|_2 =& \left\|\sum_{i=1}^n \mathbb{E}\left\{\frac{1}{n^2}\left[1(D_i=d)\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) - 1\right]^2\mathbf{U}_d(X_i)^T\mathbf{U}_d(X_i)\right\}\right\|_2 \\ \leq& \frac{1}{n}\sup_i\left[1(D_i=d)\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) - 1\right]^2\mathbb{E}\left[\mathbf{U}_d(X)^T\mathbf{U}_d(X)\right] \leq \frac{b_3}{n},\end{aligned} \tag{A.6}$$

and

$$\begin{aligned}\left\|\sum_{i=1}^n \mathbb{E}\left(\mathbf{A}_i\mathbf{A}_i^T\right)\right\|_2 =& \left\|\sum_{i=1}^n \mathbb{E}\left\{\frac{1}{n^2}\left[1(D_i=d)\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) - 1\right]^2\mathbf{U}_d(X_i)\mathbf{U}_d(X_i)^T\right\}\right\|_2 \\ \leq& \frac{1}{n}\sup_i\left[1(D_i=d)\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) - 1\right]^2\left\|\mathbb{E}\left(\mathbf{U}_d(X)\mathbf{U}_d(X)^T\right)\right\|_2, \\ \leq& \frac{1}{n}\sup_i\left[1(D_i=d)\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right) - 1\right]^2\mathbb{E}\left(\mathbf{U}_d(X)^T\mathbf{U}_d(X)\right) \leq \frac{b_3}{n},\end{aligned} \tag{A.7}$$

for some constants $b_2 > 0$ and $b_3 > 0$. By Lemma 1, combining (A.4), (A.5), (A.6) and (A.7) leads to

$$\begin{aligned}\mathbb{P}\{\|I_1\|_2 \geq t\} = \mathbb{P}\left\{\left\|\sum_{i=1}^n \mathbf{A}_i\right\|_2 \geq t\right\} &\leq (M+1)\exp\left(-\frac{n\,t^2/2}{b_3 + b_2\sqrt{M}\cdot t/3}\right) \\ &= \exp\left(\log(M+1) - \frac{n\,t^2/2}{b_3 + b_2\sqrt{M}\cdot t/3}\right),\end{aligned}$$

which tends to zero when $t = c\sqrt{M}\log n/n$ for any $c > 0$ and n goes to infinity. Therefore, we have

$$\|I_1\|_2 = O_p(\sqrt{M}\log n/\sqrt{n}). \tag{A.8}$$

For the term I_2, one can obtain that

$$\begin{aligned}
\|I_2\|_2 =& \left\| \frac{1}{n} \sum_i \left[1(D_i = d) \rho' \left(\mathbf{U}_d (X_i)^T \boldsymbol{\lambda}_d^* \right) - 1 \right] \left[\widetilde{\mathbf{U}}_d (X_i) - \mathbf{U}_d (X_i) \right] \right\|_2 \\
=& \left\| \frac{1}{n} \sum_i \left[1(D_i = d) \rho' \left(\mathbf{U}_d (X_i)^T \boldsymbol{\lambda}_d^* \right) - 1 \right] \left[\widetilde{\mathbf{F}}_1(\mathbf{q}|X_i) - \mathbf{F}_1(\mathbf{q}|X_i) \right] \right\|_2 \\
\leq& \sqrt{J} \sup_i \left| 1(D_i = d) \rho' \left(\mathbf{U}_d (X_i)^T \boldsymbol{\lambda}_d^* \right) - 1 \right| \cdot \sup_i \sup_j \left| \widetilde{F}_1(q_j|X_i) - F_1(q_j|X_i) \right| \\
=& O_p \left(\sqrt{J} \left(\left(\frac{\log n}{n h_d^p} \right)^{1/2} + h_d^2 \right) \right) = O_p \left(J^{1/2} n^{-(1/4+s/2)} \right), \qquad \text{(A.9)}
\end{aligned}$$

by Proposition 1, Assumption 7, and the boundedness of $\rho'(\cdot)$ in Assumption 8.

Next, for the term I_3, we have

$$\begin{aligned}
\|I_3\|_2 =& \left\| \frac{1}{n} \sum_i 1(D_i = d) \rho'' (\xi_i) \left[\widetilde{\mathbf{U}}_d (X_i) - \mathbf{U}_d (X_i) \right]^T \boldsymbol{\lambda}_d^* \left[\widetilde{\mathbf{U}}_d (X_i) - \mathbf{U}_d (X_i) \right] \right\|_2 \\
=& \left\| \frac{1}{n} \sum_i 1(D_i = d) \rho'' (\xi_i) \left[\widetilde{\mathbf{F}}_1(\mathbf{q}|X_i) - \mathbf{F}_1(\mathbf{q}|X_i) \right]^T \boldsymbol{\lambda}_{d,J}^* \left[\widetilde{\mathbf{F}}_1(\mathbf{q}|X_i) - \mathbf{F}_1(\mathbf{q}|X_i) \right] \right\|_2 \\
\leq& \sup_i |D_i \rho'' (\xi_i)| \cdot \left\| \boldsymbol{\lambda}_{d,J}^* \right\|_2 \cdot \sup_i \left\| \widetilde{\mathbf{F}}_1(\mathbf{q}|X_i) - \mathbf{F}_1(\mathbf{q}|X_i) \right\|_2^2 \\
\leq& J \sup_i |D_i \rho'' (\xi_i)| \cdot \left\| \boldsymbol{\lambda}_{d,J}^* \right\|_2 \cdot \sup_i \sup_j \left| \widetilde{F}_1(q_j|X_i) - F_1(q_j|X_i) \right|^2 \\
=& O_p \left(J^{3/2} \left(\left(\frac{\log n}{n h_d^p} \right)^{1/2} + h_d^2 \right)^2 \right) = O_p \left(J^{3/2} n^{-(1/2+s)} \right), \qquad \text{(A.10)}
\end{aligned}$$

where $\boldsymbol{\lambda}_{d,J}^* = (\lambda_{d,1}^*, \cdots, \lambda_{d,J}^*)^T$, the third equality holds by the result in Proposition 1, the boundedness of $\rho''(\cdot)$ in Assumption 8, and the restriction about $\boldsymbol{\lambda}_d^*$ in Assumption 10, and the last equality holds since $\left(\frac{\log n}{n h_d^p} \right)^{1/2} + h_d^2 = o\left(n^{-(1/4+s/2)} \right)$ under Assumption 7.

Similarly, for the last term I_4, we have

$$\begin{aligned}
\|I_4\|_2 =& \left\| \frac{1}{n} \sum_i 1(D_i = d) \rho''(\xi_i) \left[\widetilde{\mathbf{U}}_d(X_i) - \mathbf{U}_d(X_i) \right]^T \boldsymbol{\lambda}_d^* \mathbf{U}_d(X_i) \right\|_2 \\
=& \left\| \frac{1}{n} \sum_i 1(D_i = d) \rho''(\xi_i) \left[\widetilde{\mathbf{F}}_1(\mathbf{q}|X_i) - \mathbf{F}_1(\mathbf{q}|X_i) \right]^T \boldsymbol{\lambda}_{d,J}^* \mathbf{U}_d(X_i) \right\|_2 \\
\leq & \sup_i |1(D_i = d) \rho''(\xi_i)| \cdot \left\| \boldsymbol{\lambda}_{d,J}^* \right\|_2 \cdot \sup_i \left\| \widetilde{\mathbf{F}}_1(\mathbf{q}|X_i) - \mathbf{F}_1(\mathbf{q}|X_i) \right\| \cdot \sup_{x \in \mathcal{X}} \|\mathbf{U}_d(x)\|_2 \\
\leq & \sqrt{J} \sup_i |1(D_i = d) \rho''(\xi_i)| \cdot \left\| \boldsymbol{\lambda}_{d,J}^* \right\|_2 \cdot \sup_i \sup_j \left| \widetilde{F}_1(q_j|X_i) - F_1(q_j|X_i) \right| \cdot \sup_{x \in \mathcal{X}} \|\mathbf{U}_d(x)\|_2 \\
=& O_p \left(J M^{1/2} \left(\left(\frac{\log n}{n h_d^p} \right)^{1/2} + h_d^2 \right) \right) = O_p \left(\sqrt{M} J n^{-(1/4+s/2)} \right). \qquad \text{(A.11)}
\end{aligned}$$

Combining (A.3) and (A.8) - (A.11) leads to

$$\left\| \widetilde{G}'_{n,d}(\boldsymbol{\lambda}_d^*) \right\|_2 \leq \|I_1\|_2 + \|I_2\|_2 + \|I_3\|_2 + \|I_4\|_2 = O_p \left(\sqrt{M} J n^{-(1/4+s/2)} \right).$$

This completes the proof of Lemma 3. □

Lemma 4. *Let $\widehat{\boldsymbol{\lambda}}_d$ be unique maximizer of $\widetilde{G}_{n,d}(\boldsymbol{\lambda}_d)$. Under Assumptions 1-10, we have*

$$\left\| \widehat{\boldsymbol{\lambda}}_d - \boldsymbol{\lambda}_d^* \right\|_2 = O_p \left(\sqrt{M} J n^{-(1/4+s/2)} \right).$$

Proof of Lemma 4: Recall that

$$\widetilde{G}_{n,d}(\boldsymbol{\lambda}_d) = \frac{1}{n} \sum_{i=1}^n 1(D_i = d) \rho \left(\widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d \right) - \frac{1}{n} \sum_{i=1}^n \widetilde{\mathbf{U}}_d(X_i)^T \boldsymbol{\lambda}_d,$$

which is a strictly concave function. To prove $\left\| \widehat{\boldsymbol{\lambda}}_d - \boldsymbol{\lambda}_d^* \right\|_2 = O_p \left(\sqrt{M} J n^{-(1/4+s/2)} \right)$, it is sufficient to show that

$$\mathbb{P}\{ \widetilde{G}_{n,d}(\boldsymbol{\lambda}_d) - \widetilde{G}_{n,d}(\boldsymbol{\lambda}_d^*) < 0, \forall \boldsymbol{\lambda}_d \in \bar{\mathcal{C}} \} \to 1, \text{ as } n \to \infty, \qquad \text{(A.12)}$$

where $\bar{\mathcal{C}} = \left\{ \boldsymbol{\lambda}_d \in \mathbb{R}^M : \|\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*\|_2 = b_4 \left(\sqrt{M} J n_1^{-(1/4+s/2)} \right) \right\}$ for some constant $b_4 > 0$.

It is clear that for any $\boldsymbol{\lambda}_d \in \bar{\mathcal{C}}$, there exists a $\bar{\boldsymbol{\lambda}}_d$ between $\boldsymbol{\lambda}_d$ and $\boldsymbol{\lambda}_d^*$ such that

$$\widetilde{G}_{n,d}(\boldsymbol{\lambda}_d) - \widetilde{G}_{n,d}(\boldsymbol{\lambda}_d^*) = (\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*)^T \widetilde{G}'_{n,d}(\boldsymbol{\lambda}_d^*) + \frac{1}{2} (\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*)^T \widetilde{G}''_{n,d}(\bar{\boldsymbol{\lambda}}_d) (\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*). \qquad \text{(A.13)}$$

For the first term, by Cauchy-Schwarz inequality and the results in Lemma 3, we have

$$(\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*)^T \widetilde{G}'_{n,d}(\boldsymbol{\lambda}_d^*) \leq \|\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*\|_2 \left\|\widetilde{G}'_{n,d}(\boldsymbol{\lambda}_d^*)\right\|_2 \leq b_5 \|\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*\|_2 \sqrt{M} J n_1^{-(1/4+s/2)} \quad \text{(A.14)}$$

for some constant $b_5 > 0$. For the second term, by the results in Lemmas 2 and 3, we have

$$\begin{aligned}
&(\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*)^T \widetilde{G}''_{n,d}(\bar{\boldsymbol{\lambda}}_d)(\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*) = \frac{1}{n}\sum_{i=1}^n 1(D_i = d)\rho''\left(\widetilde{\mathbf{U}}_d(X_i)^T \bar{\boldsymbol{\lambda}}_d\right)(\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*)^T \widetilde{\mathbf{U}}_d(X_i)\widetilde{\mathbf{U}}_d(X_i)^T(\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*) \\
\leq& - b_6 (\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*)^T \left[\frac{1}{n}\sum_{i=1}^n 1(D_i = d)\widetilde{\mathbf{U}}_d(X_i)\widetilde{\mathbf{U}}_d(X_i)^T\right](\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*) \\
\leq& - b_6 \|\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*\|_2^2 \nu_{\min}\left[\frac{1}{n}\sum_{i=1}^n 1(D_i = d)\widetilde{\mathbf{U}}_d(X_i)\widetilde{\mathbf{U}}_d(X_i)^T\right] \leq -b_1 b_6 \|\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*\|_2^2, \qquad \text{(A.15)}
\end{aligned}$$

where $-b_6 = \sup_i \rho''\left(\widetilde{\mathbf{U}}_d(X_i)^T \bar{\boldsymbol{\lambda}}_d\right)$, $0 < b_6 < \infty$ since the negativity and boundedness of $\rho''(\cdot)$ in Assumption 8.

Combining (A.13), (A.14) and (A.15) yields

$$\begin{aligned}
\widetilde{G}_{n,d}(\boldsymbol{\lambda}_d) - \widetilde{G}_{n,d}(\boldsymbol{\lambda}_d^*) \leq& b_5 \|\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*\|_2 \sqrt{M} J n_1^{-(1/4+s/2)} - b_1 b_6 \|\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*\|_2^2 \\
=& b_4 b_5 \left(\sqrt{M} J n_1^{-(1/4+s/2)}\right)^2 - b_1 b_4^2 b_6 \left(\sqrt{M} J n_1^{-(1/4+s/2)}\right)^2 \\
=& b_4 (b_5 - b_1 b_4 b_6)\left(\sqrt{M} J n_1^{-(1/4+s/2)}\right)^2 < 0
\end{aligned}$$

for $\|\boldsymbol{\lambda}_d - \boldsymbol{\lambda}_d^*\|_2 = b_4\left(\sqrt{M} J n_1^{-(1/4+s/2)}\right)$ with large enough constant $b_4 > 0$. Thus, (A.12) is proved. We complete the proof of Lemma 4. □

Lemma 5. *Define $w_d^*(x) = \rho'\left(\mathbf{U}_d(X_i)^T \boldsymbol{\lambda}_d^*\right)$. Under Assumptions 1-5 and 8-10, we have*

$$\sup_{x\in\mathcal{X}}\left|\frac{1}{\pi(x)} - w_1^*(x)\right| = O_p\left(M^{1/2-r_\pi/2}\right) \quad \text{and} \quad \sup_{x\in\mathcal{X}}\left|\frac{1}{1-\pi(x)} - w_0^*(x)\right| = O_p\left(M^{1/2-r_\pi/2}\right). \qquad \text{(A.16)}$$

Proof of Lemma 5: Without loss of generality, we focus on proving $\sup_{x\in\mathcal{X}}\left|\frac{1}{\pi(x)} - w_1^*(x)\right| = O_p\left(M^{1/2-r_\pi/2}\right)$. We have

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|\frac{1}{\pi(x)} - w_1^*(x)\right| = \sup_{x\in\mathcal{X}}\left|\frac{1}{\pi(x)} - \rho'\left(\mathbf{U}_1(x)^T \boldsymbol{\lambda}_1^*\right)\right| \\
\leq& \sup_{x\in\mathcal{X}}\left|\frac{1}{\pi(x)} - \rho'\left(\mathbf{U}_1(x)^T \boldsymbol{\lambda}_1^\dagger\right)\right| + \sup_{x\in\mathcal{X}}\left|\rho'\left(\mathbf{U}_1(x)^T \boldsymbol{\lambda}_1^\dagger\right) - \rho'\left(\mathbf{U}_1(x)^T \boldsymbol{\lambda}_1^*\right)\right|. \qquad \text{(A.17)}
\end{aligned}$$

where $\boldsymbol{\lambda}_1^\dagger$ is an $M \times 1$ vector satisfying the condition given in Assumption 9.

As defined in Assumption 9, $m_1^*(x) = -\phi'\big(1/\pi(x)\big) = (\rho')^{-1}\big(1/\pi(x)\big)$. Under Assumption 9, the first term in (A.17) satisfies

$$\begin{aligned}\sup_{x\in\mathcal{X}}\left|\frac{1}{\pi(x)} - \rho'\left(\mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^\dagger\right)\right| =& \sup_{x\in\mathcal{X}}\left|\rho'\left(m_1^*(x)\right) - \rho'\left(\mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^\dagger\right)\right| \\ \lesssim& \sup_{x\in\mathcal{X}}\left|m_1^*(x) - \mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^\dagger\right| \le b_7 M^{-r_\pi} \end{aligned} \tag{A.18}$$

for some constant $b_7 > 0$. Here "$\lesssim$" denotes "less than or equal to" up to a universal constant.

For the second term in (A.17), we have

$$\begin{aligned}&\sup_{x\in\mathcal{X}}\left|\rho'\left(\mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^\dagger\right) - \rho'\left(\mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^*\right)\right|_2 \lesssim \sup_{x\in\mathcal{X}}\left|\mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^\dagger - \mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^*\right| \\ \le& \sup_{x\in\mathcal{X}}\|\mathbf{U}_1(x)\|_2\left\|\boldsymbol{\lambda}_1^\dagger - \boldsymbol{\lambda}_1^*\right\|_2 \le C_3\sqrt{M}\left\|\boldsymbol{\lambda}_1^\dagger - \boldsymbol{\lambda}_1^*\right\|_2, \end{aligned} \tag{A.19}$$

according to Assumption 9 (ii). Now we focus on $\left\|\boldsymbol{\lambda}_1^\dagger - \boldsymbol{\lambda}_1^*\right\|_2$. Recall that $\boldsymbol{\lambda}_1^*$ is the unique maximizer of

$$G_1^*(\boldsymbol{\lambda}_1) = \mathbb{E}\left[D\rho\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1\right) - \mathbf{U}_1(X)^T\boldsymbol{\lambda}_1\right] = \mathbb{E}\left[\pi(X)\rho\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1\right) - \mathbf{U}_1(X)^T\boldsymbol{\lambda}_1\right].$$

Consider a set $\mathcal{C}^\dagger = \left\{\boldsymbol{\lambda}_1 \in \mathbb{R}^M : \left\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger\right\|_2 \le b_8 M^{-r_\pi/2}\right\}$ for some constant $b_8 > 0$. To show that $\boldsymbol{\lambda}_1^* \in \mathcal{C}^\dagger$, it suffices to show that

$$\mathbb{P}\{G_1^*(\boldsymbol{\lambda}_1) - G_1^*(\boldsymbol{\lambda}_1^\dagger) < 0, \forall \boldsymbol{\lambda}_1 \in \bar{\mathcal{C}^\dagger}\} \to 1, \text{ as } n \to \infty, \tag{A.20}$$

where $\bar{\mathcal{C}^\dagger} = \left\{\boldsymbol{\lambda}_1 \in \mathbb{R}^M : \left\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger\right\|_2 = b_8 M^{-r_\pi/2}\right\}$.

To this end, define

$$G_1^\dagger(\boldsymbol{\lambda}_1) = \mathbb{E}\left[\frac{1}{\rho'\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1^\dagger\right)}\rho\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1\right) - \mathbf{U}_1(X)^T\boldsymbol{\lambda}_1\right].$$

Notice that $G_1^\dagger(\cdot)$ is a concave function and

$$(G_1^\dagger)'(\boldsymbol{\lambda}_1^\dagger) = \mathbb{E}\left[\frac{\rho'\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1^\dagger\right)}{\rho'\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1^\dagger\right)}\mathbf{U}_1(X) - \mathbf{U}_1(X)\right] = 0. \tag{A.21}$$

Thus, $\boldsymbol{\lambda}_1^\dagger$ is the unique maximizer of $G_1^\dagger(\boldsymbol{\lambda}_1)$. Moreover, for any $\boldsymbol{\lambda}_1 \in \bar{\mathcal{C}^\dagger}$, there exists a $\check{\boldsymbol{\lambda}}_1$

between $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_1^\dagger$ such that

$$\begin{aligned}
G_1^\dagger(\boldsymbol{\lambda}_1) - G_1^\dagger(\boldsymbol{\lambda}_1^\dagger) =& (G_1^\dagger)'(\boldsymbol{\lambda}_1^\dagger)^T(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger) + \frac{1}{2}(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger)^T(G_1^\dagger)''(\check{\boldsymbol{\lambda}}_1)(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger) \\
=& 0 + \frac{1}{2}\mathbb{E}\left[\frac{\rho''\left(\mathbf{U}_1(X)^T\check{\boldsymbol{\lambda}}_1\right)}{\rho'\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1^\dagger\right)}(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger)^T\mathbf{U}_1(X)\mathbf{U}_1(X)^T(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger)\right] \\
\leq& -b_9(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger)^T\mathbb{E}\left[\mathbf{U}_1(X)\mathbf{U}_1(X)^T\right](\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger) \\
\leq& -b_9\left\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger\right\|_2^2 \nu_{\min}\left(\mathbb{E}\left[\mathbf{U}_1(X)\mathbf{U}_1(X)^T\right]\right) \\
\leq& -b_9C_4\left\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger\right\|_2^2, \qquad \text{(A.22)}
\end{aligned}$$

for some constant $b_9 > 0$. We also have, for any $\boldsymbol{\lambda}_1 \in \bar{\mathcal{C}}^\dagger$,

$$\begin{aligned}
&\left|G_1^\dagger(\boldsymbol{\lambda}_1) - G_1^*(\boldsymbol{\lambda}_1)\right| = \left|\mathbb{E}\left[\left(\frac{1}{\rho'\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1^\dagger\right)} - \pi(X)\right)\rho\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1\right)\right]\right| \\
=& \left|\mathbb{E}\left[\frac{\pi(X)}{\rho'\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1^\dagger\right)}\left(\frac{1}{\pi(X)} - \rho'\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1^\dagger\right)\right)\rho\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1\right)\right]\right| \\
\leq& \mathbb{E}\left[\left|\frac{\pi(X)}{\rho'\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1^\dagger\right)}\right|\left|\frac{1}{\pi(X)} - \rho'\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1^\dagger\right)\right|\left|\rho\left(\mathbf{U}_1(X)^T\boldsymbol{\lambda}_1\right)\right|\right] \\
\leq& \sup_{x\in\mathcal{X}}\left|\frac{\pi(x)}{\rho'\left(\mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^\dagger\right)}\right| \cdot \sup_{x\in\mathcal{X}}\left|\frac{1}{\pi(x)} - \rho'\left(\mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^\dagger\right)\right| \cdot \sup_{x\in\mathcal{X}}\left|\rho\left(\mathbf{U}_1(x)^T\boldsymbol{\lambda}_1\right)\right| \\
\leq& b_{10}M^{-r_\pi}
\end{aligned}$$

for some constant $b_{10} > 0$ according to (A.18) and the boundedness of $\pi(\cdot)$, $\rho'(\cdot)$ and $\rho(\cdot)$. Hence, for any $\boldsymbol{\lambda}_1 \in \mathcal{C}^\dagger$,

$$G_1^*(\boldsymbol{\lambda}_1) - G_1^*(\boldsymbol{\lambda}_1^\dagger) \leq G_1^\dagger(\boldsymbol{\lambda}_1) - G_1^\dagger(\boldsymbol{\lambda}_1^\dagger) + 2b_{10}M^{-r_\pi}. \qquad \text{(A.23)}$$

Plugging (A.22) into (A.23), we have

$$G_1^*(\boldsymbol{\lambda}_1) - G_1^*(\boldsymbol{\lambda}_1^\dagger) \leq -b_9C_4\left\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1^\dagger\right\|_2^2 + 2b_{10}M^{-r_\pi} = (2b_{10} - b_9C_4b_8^2)M^{-r_\pi} < 0$$

for a large enough $b_8 > 0$. Thus, (A.20) is proved, which implies

$$\left\|\boldsymbol{\lambda}_1^* - \boldsymbol{\lambda}_1^\dagger\right\|_2 \leq b_8M^{-r_\pi/2}. \qquad \text{(A.24)}$$

Finally, combining (A.17), (A.18), (A.19) and (A.24) yields that

$$\sup_{x\in\mathcal{X}}\left|\frac{1}{\pi(x)}-\rho'\left(\mathbf{U}_1(x)^T\boldsymbol{\lambda}_1^*\right)\right|\leq b_7M^{-r_\pi}+C_3b_8\sqrt{M}M^{-r_\pi/2}=O_p\left(M^{1/2-r_\pi/2}\right).$$

This completes the proof of Lemma 5. □

Lemma 6. *For $w_d^*(x)=\rho'\left(\mathbf{U}_d(X_i)^T\boldsymbol{\lambda}_d^*\right)$, one has*

$$\left\|\frac{1}{\pi(x)}-w_1^*(x)\right\|_{P,2}=O_p\left(M^{-r_\pi/2}\right)\ \text{ and }\ \left\|\frac{1}{1-\pi(x)}-w_0^*(x)\right\|_{P,2}=O_p\left(M^{-r_\pi/2}\right). \quad (A.25)$$

under Assumptions 1-5 and 8-10.

Proof of Lemma 6: Without loss of generality, we focus on proving $\left\|\frac{1}{\pi(x)}-w_1^*(x)\right\|_{P,2}=O_p\left(M^{-r_\pi/2}\right)$. We have

$$\begin{aligned}
&\left\|\frac{1}{\pi(x)}-w_1^*(x)\right\|_{P,2}=\left\|\rho'\left(m_1^*(x)\right)-\rho'\left(\mathbf{U}_d(x)^T\boldsymbol{\lambda}^*\right)\right\|_{P,2}\\
\leq&\left\|\rho'\left(m_1^*(x)\right)-\rho'\left(\mathbf{U}_d(x)^T\boldsymbol{\lambda}^\dagger\right)\right\|_{P,2}+\left\|\rho'\left(\mathbf{U}_d(x)^T\boldsymbol{\lambda}^\dagger\right)-\rho'\left(\mathbf{U}_d(x)^T\boldsymbol{\lambda}^*\right)\right\|_{P,2}\\
\lesssim&\left\|m_1^*(x)-\mathbf{U}_d(x)^T\boldsymbol{\lambda}^\dagger\right\|_{P,2}+\left\|\mathbf{U}_d(x)^T\boldsymbol{\lambda}^\dagger-\mathbf{U}_d(x)^T\boldsymbol{\lambda}^*\right\|_{P,2}\\
\leq&\sup_{x\in\mathcal{X}}\left|m_1^*(x)-\mathbf{U}_d(x)^T\boldsymbol{\lambda}^\dagger\right|+\left\|\mathbf{U}_d(x)\right\|_{P,2}\left\|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}^\dagger\right\|_2\\
=&O_p\left(M^{-r_\pi}\right)+O_p\left(M^{-r_\pi/2}\right)=O_p\left(M^{-r_\pi/2}\right)
\end{aligned}$$

according to Assumption 9 and (A.24). This completes the proof of Lemma 6. □

Proof of Proposition 2: Given the results in Lemmas 5 and 6, we only need to prove $\sup_{x\in\mathcal{X}}\left|\widehat{w}_d(x)-w_d^*(x)\right|=O_p\left(MJn^{-(1/4+s/2)}\right)$ and $\left\|\widehat{w}_d(x)-w_d^*(x)\right\|_{P,2}=O_p\left(\sqrt{M}Jn^{-(1/4+s/2)}\right)$ for $d=0$ and 1.

We first focus on $\sup_{x\in\mathcal{X}}\left|\widehat{w}_d(x)-w_d^*(x)\right|$. Notice that

$$\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|\widehat{w}_d(x)-w_d^*(x)\right|=\sup_{x\in\mathcal{X}}\left|1(D_i=d)\left[\rho'\left(\widetilde{\mathbf{U}}_d\left(x\right)^T\widehat{\boldsymbol{\lambda}}_d\right)-\rho'\left(\mathbf{U}_d\left(x\right)^T\boldsymbol{\lambda}_d^*\right)\right]\right|\\
\leq&\sup_{x\in\mathcal{X}}\left|\rho'\left(\widetilde{\mathbf{U}}_d\left(x\right)^T\widehat{\boldsymbol{\lambda}}_d\right)-\rho'\left(\mathbf{U}_d\left(x\right)^T\boldsymbol{\lambda}_d^*\right)\right|\lesssim\sup_{x\in\mathcal{X}}\left|\widetilde{\mathbf{U}}_d\left(x\right)^T\widehat{\boldsymbol{\lambda}}_d-\mathbf{U}_d\left(x\right)^T\boldsymbol{\lambda}_d^*\right|
\end{aligned}$$

and

$$
\begin{aligned}
&\sup_{x\in\mathcal{X}}\left|\widetilde{\mathbf{U}}_d(x)^T\widehat{\boldsymbol{\lambda}}_d-\mathbf{U}_d(x)^T\boldsymbol{\lambda}_d^*\right| \\
=&\sup_{x\in\mathcal{X}}\left|\left[\widetilde{\mathbf{U}}_d(x)-\mathbf{U}_d(x)+\mathbf{U}_d(x)\right]^T\left(\widehat{\boldsymbol{\lambda}}_d-\boldsymbol{\lambda}_d^*+\boldsymbol{\lambda}_d^*\right)-\mathbf{U}_d(x)^T\boldsymbol{\lambda}_d^*\right| \\
\leq&\sup_{x\in\mathcal{X}}\left|\left[\widetilde{\mathbf{U}}_d(x)-\mathbf{U}_d(x)\right]^T\left(\widehat{\boldsymbol{\lambda}}_d-\boldsymbol{\lambda}_d^*\right)\right|+\sup_{x\in\mathcal{X}}\left|\left[\widetilde{\mathbf{U}}_d(x)-\mathbf{U}_d(x)\right]^T\boldsymbol{\lambda}_d^*\right|+\sup_{x\in\mathcal{X}}\left|\mathbf{U}_d(x)^T\left(\widehat{\boldsymbol{\lambda}}_d-\boldsymbol{\lambda}_d^*\right)\right| \\
\leq&\sup_{x\in\mathcal{X}}\left\|\widetilde{\mathbf{F}}_1(\mathbf{q}|x)-\mathbf{F}_1(\mathbf{q}|x)\right\|_2\left\|\widehat{\boldsymbol{\lambda}}_d-\boldsymbol{\lambda}_d^*\right\|_2+\sup_{x\in\mathcal{X}}\left\|\widetilde{\mathbf{F}}_1(\mathbf{q}|x)-\mathbf{F}_1(\mathbf{q}|x)\right\|_2\left\|\boldsymbol{\lambda}_{d,J}^*\right\|_2 \\
&+\sup_{x\in\mathcal{X}}\left\|\mathbf{U}_d(x)\right\|_2\left\|\widehat{\boldsymbol{\lambda}}_d-\boldsymbol{\lambda}_d^*\right\|_2 \\
=&O_p(J^{1/2}n^{-(1/4+s/2)})\cdot O_p\left(\sqrt{M}Jn^{-(1/4+s/2)}\right)+O_p(Jn^{-(1/4+s/2)})+O_p\left(MJn^{-(1/4+s/2)}\right) \\
=&O_p\left(MJn^{-(1/4+s/2)}\right)
\end{aligned}
$$

according to the results in Proposition 1 and Lemma 4. Then, we have

$$\sup_{x\in\mathcal{X}}|\widehat{w}_d(x)-w_d^*(x)|=O_p\left(MJn^{-(1/4+s/2)}\right).$$

Similarly, for $\|\widehat{w}_d(x)-w_d^*(x)\|_{P,2}$, we have

$$
\begin{aligned}
&\|\widehat{w}_d(x)-w_d^*(x)\|_{P,2}\leq\left\|\rho'\left(\widetilde{\mathbf{U}}_d(x)^T\widehat{\boldsymbol{\lambda}}_d\right)-\rho'\left(\mathbf{U}_d(x)^T\boldsymbol{\lambda}_d^*\right)\right\|_{P,2} \\
\lesssim&\left\|\widetilde{\mathbf{U}}_d(x)^T\widehat{\boldsymbol{\lambda}}_d-\mathbf{U}_d(x)^T\boldsymbol{\lambda}_d^*\right\|_{P,2} \\
=&\left\|\left[\widetilde{\mathbf{U}}_d(x)-\mathbf{U}_d(x)+\mathbf{U}_d(x)\right]^T\left(\widehat{\boldsymbol{\lambda}}_d-\boldsymbol{\lambda}_d^*+\boldsymbol{\lambda}_d^*\right)-\mathbf{U}_d(x)^T\boldsymbol{\lambda}_d^*\right\|_{P,2} \\
\leq&\left\|\left[\widetilde{\mathbf{U}}_d(x)-\mathbf{U}_d(x)\right]^T\left(\widehat{\boldsymbol{\lambda}}_d-\boldsymbol{\lambda}_d^*\right)\right\|_{P,2}+\left\|\left[\widetilde{\mathbf{U}}_d(x)-\mathbf{U}_d(x)\right]^T\boldsymbol{\lambda}_d^*\right\|_{P,2}+\left\|\mathbf{U}_d(x)^T\left(\widehat{\boldsymbol{\lambda}}_d-\boldsymbol{\lambda}_d^*\right)\right\|_{P,2} \\
=&O_p(Jn_1^{-(1/4+s/2)})\cdot O_p\left(\sqrt{M}Jn_1^{-(1/4+s/2)}\right)+O_p(Jn^{-(1/4+s/2)})+O_p\left(\sqrt{M}Jn_1^{-(1/4+s/2)}\right) \\
=&O_p\left(\sqrt{M}Jn_1^{-(1/4+s/2)}\right).
\end{aligned}
$$

This concludes the proof of Proposition 2. □

B Proofs of Theorems 1 and 2

Before presenting proofs of Theorems 1 and 2, we first introduce some notations and inequalities. Let $\mathcal{G}=\{g:\mathcal{X}\mapsto\mathbb{R}\}$ be a class of measurable real-valued functions defined on $\mathcal{X}$. An envelop function of a class $\mathcal{G}$ is a function $G(x)$ such that $|g(x)|\leq G(x)$, for every

$x \in \mathcal{X}$ and $g \in \mathcal{G}$. The bracketing number $N_{[]}(\epsilon, \mathcal{G}, \|\cdot\|)$ is the minimum number of ϵ brackets needed to cover $\mathcal{G}$. The bracketing integral is defined as, for any $\delta > 0$,

$$J_{[]}\left(\delta, \mathcal{G}, L_2(P)\right) = \int_0^{\delta} \sqrt{\ln N_{[]}(\epsilon, \mathcal{G}, L_r(P))}\, d\epsilon < \infty. \tag{B.1}$$

Suppose $\{X_i\}_{i=1}^n$ are i.i.d following probability P and for any function $g \in \mathcal{G}$. Define the empirical process as

$$\mathbb{G}_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) - \mathbb{E}\left[g(X_i)\right].$$

Lemma 7. (Maximal Inequality.) *For any class $\mathcal{G}$ of measurable functions with an envelope function G, one has*

$$\mathbb{E}\left[\sup_{g \in \mathcal{G}} |\mathbb{G}_n(g)|\right] \lesssim J_{[]}(\|G\|_{P,2}, \mathcal{G}, L_2(P)).$$

where "$\lesssim$" denotes "less than or equal to" up to a universal constant.

Proof of Lemma 7: The result is given by Corollary 19.35 in van der Vaart (2000). □

Proof of Theorem 1: Without loss of generality, we focus on the case $d = 1$. Recall that $\mathbf{F}_1(\mathbf{q} \,|\, X_i) = (F_1(q_1 \,|\, X_i), \cdots, F_1(q_J \,|\, X_i))^T$ with $y_l = q_1 < q_2 < \cdots < q_J = y_u$ being equally-spaced grid points on $\mathcal{Y} = [y_l, y_u]$, $c_j(y) = \prod_{k=1, k\neq j}^{J} \frac{y - q_k}{q_j - q_k}$, and $\mathbf{c}(y) = (c_1(y), \cdots, c_J(y))^T$. Then, since $\frac{1}{n}\sum_i \left[D_i \widehat{w}_1(X_i) - 1\right] \mathbf{c}(y)^T \widetilde{\mathbf{F}}_1(\mathbf{q} \,|\, X_i) = 0$ as posited in the constraints of the optimization problem, $\sqrt{n}\left[\widehat{F}_1(y) - F_1(y)\right]$ can be written as

$$\begin{aligned}
&\sqrt{n}\left\{\widehat{F}_1(y) - F_1(y)\right\} = \sqrt{n}\left\{\frac{1}{n}\sum_i D_i \widehat{w}_1(X_i) 1\left(Y_i \leq y\right) - F_1(y)\right\} \\
=&\sqrt{n}\left\{\frac{1}{n}\sum_i \frac{D_i}{\pi(X_i)}\left[1\left(Y_i \leq y\right) - F_1\left(y \,|\, X_i\right)\right] + \frac{1}{n}\sum_i F_1\left(y \,|\, X_i\right) - F_1(y)\right. \\
&+ \frac{1}{n}\sum_i D_i \left[\widehat{w}_1(X_i) - \frac{1}{\pi(X_i)}\right]\left[1\left(Y_i \leq y\right) - F_1\left(y \,|\, X_i\right)\right] \\
&\left.+ \frac{1}{n}\sum_i \left[D_i \widehat{w}_1(X_i) - 1\right] \mathbf{c}(y)^T \left[\mathbf{F}_1(\mathbf{q} \,|\, X_i) - \widetilde{\mathbf{F}}_1(\mathbf{q} \,|\, X_i)\right]\right\} \\
&+ \frac{1}{n}\sum_i \left[D_i \widehat{w}_1(X_i) - 1\right]\left[F_1\left(y \,|\, X_i\right) - \mathbf{c}(y)^T \mathbf{F}_1(\mathbf{q} \,|\, X_i)\right] \\
:=&S_1 + R_1 + R_2 + R_3.
\end{aligned}$$

To prove the theorem, it suffices to show that S_1 asymptotically follows a normal distribution, and $R_1 + R_2 + R_3$ is $o_p(1)$.

First, we consider the term R_1, which is

$$\begin{aligned} R_1 =& \frac{1}{\sqrt{n}} \sum_i D_i \left[\widehat{w}_1(X_i) - \pi^{-1}(X_i)\right] \left[1\left(Y_i \leq y\right) - F_1\left(y \,|\, X_i\right)\right] \\ =& \frac{1}{\sqrt{n}} \sum_i D_i \left[\rho'\left(\widetilde{\mathbf{U}}_1\left(X_i\right)^T \widehat{\boldsymbol{\lambda}}_1\right) - \rho'\left(m_1^*(X_i)\right)\right] \left[1\left(Y_i \leq y\right) - F_1\left(y \,|\, X_i\right)\right]. \end{aligned}$$

Define

$$g_1(D, Y, X) := D\left[\rho'\left(m_1(X)\right) - \rho'\left(m_1^*(X)\right)\right]\left[1\left(Y \leq y\right) - F_1\left(y \,|\, X\right)\right],$$

where, $m_1 : \mathcal{X} \mapsto \mathbb{R}$ is a continuous bounded function. Since $\mathbb{E}\left[g_1(D, Y, X)\right] = 0$, the empirical process of g_1 is

$$\begin{aligned} \mathbb{G}_n(g_1) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{g_1(D_i, Y_i, X_i) - \mathbb{E}\left[g_1(D, Y, X)\right]\right\} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n D\left[\rho'\left(m_1(X)\right) - \rho'\left(m_1^*(X)\right)\right]\left[1\left(Y \leq y\right) - F_1\left(y \,|\, X\right)\right]. \end{aligned}$$

Consider a set of functions

$$\mathcal{G}_1 = \left\{ g_1 : \sup_{x \in \mathcal{X}} \left|m_1(x) - m_1^*(x)\right| \leq \eta_1 \right\}$$

with $\eta_1 = b_{11}\left(M J n_1^{-(1/4+s/2)} + M^{1/2 - r_\pi/2}\right)$ for some constant $b_{11} > 0$. Then, by the maximal inequality in Lemma 7, we have

$$R_1 \leq \sup_{g_1 \in \mathcal{G}_1} \mathbb{G}_n\left(g_1\right) \lesssim \mathbb{E} \sup_{g_1 \in \mathcal{G}_1} \mathbb{G}_n\left(g_1\right) \lesssim J_{[]}\left\{\left\|G_1\right\|_{P,2}, \mathcal{G}_1, L_2(P)\right\},$$

where $G_1 = b_{12} \cdot \eta_1 \geq \left|g_1(D_i, Y_i, X_i)\right|$ is an envelop function for some $b_{12} > 0$. Consequently, we have $\left\|G_1\right\|_{P,2} = \left[\mathbb{E}\left(G_1^2\right)\right]^{1/2} \lesssim \eta_1$. Then, we bound $J_{[]}\left\{\left\|G_1\right\|_{P,2}, \mathcal{G}_1, L_2(P)\right\}$ by $N_{[]}\left(\varepsilon, \mathcal{G}_1, L_2(P)\right)$.

To bound $\log N_{[]}\left(\varepsilon, \mathcal{G}_1, L_2(P)\right)$, define a new set of function $\mathcal{G}_1' = \{g_1 : \sup_{x \in \mathcal{X}} \left|m_1(x) - m_1^*(x)\right| \leq b_{13}\}$ for some constant $b_{13} > 0$. Then, it is easily seen that

$$\begin{aligned} \log N_{[]}\left(\varepsilon, \mathcal{G}_1, L_2(P)\right) \lesssim& \log N_{[]}\left(\varepsilon, \eta_1 \mathcal{G}_1', L_2(P)\right) \lesssim \log N_{[]}\left(\varepsilon/\eta_1, \mathcal{G}_1', L_2(P)\right) \\ \lesssim& \log N_{[]}\left(\varepsilon/\eta_1, \mathcal{M}_1, L_2(P)\right) \leq C_5 (\eta_1/\varepsilon)^{(1/\nu)}. \end{aligned}$$

where, $\eta_1\mathcal{G}_1' = \{\eta_1 g_1 : \sup_{x\in\mathcal{X}} |m_1(x) - m_1^*(x)| \leq b_{13}\}$, $\mathcal{M}_1$ is the function class to which $m_1^*(x)$ pertains, and the last inequality is due to Assumption 9 (iv). Thus, $J_{[]}\left\{\|G_1\|_{P,2}, \mathcal{G}_1, L_2(P)\right\}$ is bounded as follows,

$$J_{[]}\left\{\|G_1\|_{P,2}, \mathcal{G}_1, L_2(P)\right\} \lesssim \int_0^{\eta_1} \left[\log N_{[]}\left\{\varepsilon, \mathcal{G}_1, L_2(P)\right\}\right]^{1/2} \mathrm{d}\varepsilon \lesssim \int_0^{\eta_1} (\eta_1/\varepsilon)^{1/(2\nu)} \mathrm{d}\varepsilon \to 0,$$

where, the last term holds since η_1 goes to 0 and $2\nu > 1$. This complete the proof $R_1 = o_p(1)$.

Next, we consider the term R_2, which is

$$\begin{aligned} R_2 =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[D_1 \widehat{w}_1(X_i) - 1\right] \mathbf{c}(y)^T \left[\mathbf{F}_1(\mathbf{q} \,|\, X_i) - \widetilde{\mathbf{F}}_1(\mathbf{q} \,|\, X_i)\right] \\ =& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[D_1 \rho'\left(\widetilde{\mathbf{U}}_1 \left(X_i\right)^T \widehat{\boldsymbol{\lambda}}_1\right) - 1\right] \sum_{j=1}^{J} c_j(y) \left[F_1(q_j \,|\, X) - \widetilde{F}_1(q_j \,|\, X)\right]. \end{aligned}$$

Define

$$g_2(D, Y, X) = [D\rho'(m_1(X)) - 1] \sum_{j=1}^{J} c_j(y) \Delta_j(X),$$

where, $\Delta_j(X) = F_1(q_j \,|\, X) - \widetilde{F}_1(q_j \,|\, X)$. The empirical process of g_2 is

$$\mathbb{G}_n(g_2) = \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n} g_2(D_i, Y_i, X_i) - \mathbb{E}\left[g_2(D, Y, X)\right]\right\}.$$

Define a class of functions as follows

$$\mathcal{G}_2 = \left\{g_2 : \|m_1 - m_1^*\|_{P,2} \leq \eta_{21}, \|\Delta_j\|_\infty \leq \eta_{22} \text{ for all } j\right\},$$

where $\eta_{21} = b_{14}\left(\sqrt{M} J n_1^{-(1/4+s/2)} + M^{-r_\pi/2}\right)$ and $\eta_{22} = b_{15}\left(\left(\frac{\log n_1}{n_1 h_1^p}\right)^{1/2} + h_1^2\right)$ for some constants $b_{14} > 0$ and $b_{15} > 0$. Then, we have

$$R_2 \leq \sup_{g_2\in\mathcal{G}_2} \mathbb{G}_n\left(g_2\right) + n^{1/2} \sup_{g_2\in\mathcal{G}_2} \mathbb{E}(g_2). \tag{B.2}$$

For the first term $\sup_{g_2\in\mathcal{G}_2} \mathbb{G}_n\left(g_2\right)$ on the right side of (B.2), by the maximal inequality in Lemma 7, we have

$$\sup_{g_2\in\mathcal{G}_2} \mathbb{G}_n\left(g_2\right) \lesssim \mathbb{E} \sup_{g_2\in\mathcal{G}_2} \mathbb{G}_n\left(g_2\right) \lesssim J_{[]}\left\{\|G_2\|_{P,2}, \mathcal{G}_2, L_2(P)\right\},$$

where $G_2 = b_{16} \cdot J\eta_{22} \geq |g_2(D, Y, X)|$ is an envelop function for some constant $b_{16} > 0$. Con-

sequently, we have $\|G_2\|_{P,2} \lesssim J\eta_{22}$. Similar to characterizing R_1, $J_{[]}\left\{\|G_2\|_{P,2}, \mathcal{G}_2, L_2(P)\right\}$ is bounded by $N_{[]}(\varepsilon, \mathcal{G}_2, L_2(P))$.

To bound $N_{[]}(\varepsilon, \mathcal{G}_2, L_2(P))$, for some constant $b_{17} > 0$, we define three new classes: $\mathcal{G}_2' = \{g_2 : \|m_1 - m_1^*\|_{P,2} \leq b_{17}, \|\Delta_j\|_{P,2} \leq 1 \text{ for all } j\}$, $\mathcal{G}_{20}' = \{m_1 \in \mathcal{M}_1 + m_1^* : \|m_1\|_{P,2} \leq b_{17}\}$, and $\mathcal{G}_{21}' = \left\{\Delta_j \in \mathcal{F}_1 - \widetilde{F}_1(q_j \,|\, X) : \|\Delta_j\|_{P,2} \leq 1 \text{ for all } j\right\}$. Then, we have

$$\begin{aligned}
\log N_{[]}\{\varepsilon, \mathcal{G}_2, L_2(P)\} &\lesssim \log N_{[]}\{\varepsilon/(J\eta_{22}), \mathcal{G}_2', L_2(P)\} \\
&\lesssim \log N_{[]}\{\varepsilon/(J\eta_{22}), \mathcal{G}_{20}', L_2(P)\} + \log N_{[]}\{\varepsilon/(J\eta_{22}), \mathcal{G}_{21}', L_2(P)\} \\
&\lesssim \log N_{[]}\{\varepsilon/(J\eta_{22}), \mathcal{M}_1, L_2(P)\} + \log N_{[]}\{\varepsilon/(J\eta_{22}), \mathcal{F}_1, L_2(P)\} \\
&\leq C_5\left((J\eta_{22})/\varepsilon\right)^{1/\nu} + C_1\left((J\eta_{22})/\varepsilon\right)^{p/r}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
J_{[]}\left\{\|g_2\|_{P,2}, \mathcal{G}_2, L_2(P)\right\} &\lesssim \int_0^{J\eta_{22}} \left[\log N_{[]}\{\varepsilon, \mathcal{G}_2, L_2(P)\}\right]^{1/2} \mathrm{d}\varepsilon \\
&\lesssim \int_0^{J\eta_{22}} (J\eta_{22}/\varepsilon)^{1/(2\nu)}\,\mathrm{d}\varepsilon + \int_0^{J\eta_{22}} (J\eta_{22}/\varepsilon)^{p/(2r)}\,\mathrm{d}\varepsilon.
\end{aligned}$$

By $2\nu > 1$, $p/(2r) < 1$ and $J\eta_{22} \to 0$, we have $J_{[]}\left\{\|g_2\|_{P,2}, \mathcal{G}_2, L_2(P)\right\} = o(1)$, which implies

$$\sup_{g_2 \in \mathcal{G}_2} \mathbb{G}_n(g_2) = o_p(1). \tag{B.3}$$

For the second term $n^{1/2}\sup_{g_2\in\mathcal{G}_2}\mathbb{E}(g_2)$ on the right hand side of (B.2). Let $\mathcal{G}_{20} = \{m_1 : \|m_1 - m_1^*\|_{P,2} \leq \eta_{21}\}$ and $\mathcal{G}_{21} = \{\Delta : \|\Delta\|_\infty \leq \eta_{22}\}$. Thus, one has

$$\begin{aligned}
n^{1/2}\sup_{g_2\in\mathcal{G}_2}\mathbb{E}g_2 &= n^{1/2}\sup_{m_1\in\mathcal{G}_{20},\Delta_j\in\mathcal{G}_{21}} \mathbb{E}\left\{[\pi(X)\rho'\{m_1(X)\} - 1]\sum_{j=1}^J c_j(y)\Delta_j(X)\right\} \\
&= n^{1/2}\sup_{m_1\in\mathcal{G}_{20},\Delta\in\mathcal{G}_{21}} \mathbb{E}\left(\left[\rho'\{m_1(X)\} - \frac{1}{\pi(X)}\right]\pi(X)\sum_{j=1}^J c_j(y)\Delta_j(X)\right) \\
&\lesssim n^{1/2}J\sup_{m_1\in\mathcal{G}_{20}}\left\|\rho'\{m_1(X)\} - \frac{1}{\pi(X)}\right\|_{P,2}\sup_{\Delta_j\in\mathcal{G}_{21}}\|\Delta_j(X)\|_{P,2} \\
&\lesssim n^{1/2}J\eta_{21}\eta_{22} = o_p(1),
\end{aligned} \tag{B.4}$$

where the last equality is due to Assumptions 7 and 9.

Combining (B.2)-(B.4), we can conclude that $R_2 = o_p(1)$.

Now, for the term R_3, we have

$$\begin{aligned} R_3 =& \frac{1}{\sqrt{n}} \sum_i \left[D_i \widehat{w}_1(X_i) - 1\right] \left[F_1\left(y \,|\, X_i\right) - \mathbf{c}(y)^T \mathbf{F}_1(\mathbf{q} \,|\, X_i)\right] \\ \leq & \sqrt{n} \sup_i \left|D_i \widehat{w}_1(X_i) - 1\right| \sup_i \left|F_1\left(y \,|\, X_i\right) - \mathbf{c}(y)^T \mathbf{F}_1(\mathbf{q} \,|\, X_i)\right| \\ =& O_p\left(\sqrt{n}\, \frac{C^J}{J(J-1)^J}\right) = o(1), \end{aligned}$$

by assuming $J = O(\log(n))$ as in Assumption 9 (i).

Finally, for any $y \in \mathcal{Y}$,

$$\begin{aligned} & \sqrt{n}\left[\widehat{F}_d(y) - F_1(y)\right] \\ =& \frac{1}{\sqrt{n}} \sum_i \left\{ \frac{1(D_i = d)\left[1\left(Y_i \leq y\right) - F_1\left(y \,|\, X_i\right)\right]}{\pi(X_i)^d [1 - \pi(X_i)]^{(1-d)}} + F_1\left(y \,|\, X_i\right) - F_1(y) \right\} + o_p(1) \\ =& \frac{1}{\sqrt{n}} \sum_i \psi_d^F(y, \mathbf{Z}_i) + o_p(1), \end{aligned}$$

where $\mathbf{Z}_i = (Y_i, D_i, X_i)$. It can be shown that the function class $\mathcal{I} = \{y \in \mathcal{Y} : \psi_d^F(y, \mathbf{Z}_i)\}$ belongs to Donsker class by combining the results in Examples 19.6, 19.9 and 19.20 in van der Vaart (2000). Note that the Cartesian product of two Donsker classes is still a Donsker class (van der Vaart, 2000). Then, the result in Theorem 1 follows directly from the functional central limit theorem. □

Proof of Theorem 2: Denote $\widehat{\mathbb{F}}^b(\mathbf{y}) = \sqrt{n}\left[\widehat{\mathbf{F}}^{\mathbf{b}}(\mathbf{y}) - \widehat{\mathbf{F}}(\mathbf{y})\right]$. Since the function class $\mathcal{I} = \{y \in \mathcal{Y} : \psi_d^F(y, \mathbf{Z}_i)\}$ belongs to Donsker class, then, by Theorem 3.6.1 in van der Vaart and Wellner (1996), we have

$$\sup_{h \in \mathrm{BL}_1} \left| \mathbb{E}\left(h\left(\widehat{\mathbb{F}}^b\right) \mid \mathbf{Z}_i\right) - \mathbb{E}(h(\mathbb{F}))\right| \xrightarrow{p} 0, \tag{B.5}$$

where BL_1 is the set of bounded and Lipschitz functions from $\ell^\infty(\mathcal{I})$ to $\mathbb{R}$. It follows from the results in Section 1.12 in van der Vaart and Wellner (1996) that (B.5) implies weak convergence $\widehat{\mathbb{F}}^b(\cdot) \overset{p}{\Rightarrow} \mathbb{F}(\cdot)$ conditional on the data. Thus, Theorem 2 is proved. □

C Proofs of Propositions 3 - 5

Proof of Proposition 3: By the Bahadur representation, we have

$$\sqrt{n}\left[\widehat{Q}_d(\tau) - Q_d(\tau)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\psi_d^F(Q_d(\tau), \mathbf{Z}_i)}{f_d(Q_d(\tau))} + o_p(1),$$

and then, $\sqrt{n}\left[\widehat{\Delta}(\tau) - \Delta(\tau)\right]$ can be represented as:

$$\sqrt{n}\left[\widehat{\Delta}(\tau) - \Delta(\tau)\right] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\psi_1^F(Q_1(\tau), \mathbf{Z}_i)}{f_1(Q_1(\tau))} - \frac{\psi_0^F(Q_0(\tau), \mathbf{Z}_i)}{f_0(Q_0(\tau))} + o_p(1).$$

It can be shown easily that the function class $\left\{\tau \in [0,1] : \frac{\psi_1^F(Q_1(\tau),\mathbf{Z}_i)}{f_1(Q_1(\tau))} - \frac{\psi_0^F(Q_0(\tau),\mathbf{Z}_i)}{f_0(Q_0(\tau))}\right\}$ belongs to Donsker class by the results in Examples 19.6, 19.9 and 19.20 in van der Vaart (2000). Hence, Proposition 3 is proved by the functional central limit theorem. □

Proof of Proposition 4: Denote $\widehat{\mathbb{Q}}^b(\cdot) = \sqrt{n}\left[\widehat{\Delta}^b(\cdot) - \widehat{\Delta}(\cdot)\right]$. Since the function class $\mathcal{Q} = \left\{\tau \in [a,b] : \frac{\psi_1^F(Q_1(\tau),\mathbf{Z}_i)}{f_1(Q_1(\tau))} - \frac{\psi_0^F(Q_0(\tau),\mathbf{Z}_i)}{f_0(Q_0(\tau))}\right\}$ belongs to Donsker class, then, by Theorem 3.6.1 in van der Vaart and Wellner (1996), we have

$$\sup_{h\in \mathrm{BL}_1}\left|\mathbb{E}\left(h\left(\widehat{\mathbb{Q}}^b\right) \mid \mathbf{Z}_i\right) - \mathbb{E}(h(\mathbb{Q}))\right| \xrightarrow{p} 0, \tag{C.1}$$

where BL_1 is the set of bounded and Lipschitz functions from $\ell^\infty(\mathcal{Q})$ to $\mathbb{R}$. An application of the results in Section 1.12 in van der Vaart and Wellner (1996) gives that (C.1) implies weak convergence $\widehat{\mathbb{Q}}^b(\cdot) \overset{p}{\Rightarrow} \mathbb{Q}(\cdot)$ conditional on the data. Thus, Proposition 4 is proved. □

Proof of Proposition 5: From the proof of Theorem 1, we have

$$\sqrt{n}\left\{\left[\widehat{F}_1(y) - F_1(y)\right] - \left[\widehat{F}_0(y) - F_0(y)\right]\right\} = \frac{1}{\sqrt{n}}\sum_i\left[\psi_1^F(y, \mathbf{Z}_i) - \psi_0^F(y, \mathbf{Z}_i)\right] + o_p(1),$$

where the function class $\left\{y \in \mathcal{Y} : \psi_1^F(y, \mathbf{Z}_i) - \psi_0^F(y, \mathbf{Z}_i)\right\}$ belongs to Donsker class. Thus, by the functional central limit theorem, under Assumptions 1-10, uniformly for $y \in \mathcal{Y}$, we have

$$\sqrt{n}\left\{\left[\widehat{F}_1(y) - F_1(y)\right] - \left[\widehat{F}_0(y) - F_0(y)\right]\right\} \;\Rightarrow\; \mathbb{KS}(y), \tag{C.2}$$

where $\mathbb{KS}(y)$ is a Gaussian process with mean zero and covariance function $\Psi^{\mathrm{KS}}(y, y') =$

$\mathbb{E}\left[\psi^{\text{KS}}(y)\,\psi^{\text{KS}}(y')\right]$ with $\psi^{\text{KS}}(y) = \left[\psi_1^F(y, \mathbf{Z}_i) - \psi_0^F(y, \mathbf{Z}_i)\right]$ and ψ_d^F is given in (12). Define

$$\widehat{T} = \sqrt{n} \sup_y \left\{ \left[\widehat{F}_1(y) - F_1(y)\right] - \left[\widehat{F}_0(y) - F_0(y)\right] \right\}.$$

Then, it follows from the continuous mapping theory that,

$$\widehat{T} \xrightarrow{d} \sup_y \mathbb{KS}(y).$$

Next, let $\widehat{F}_1^b$ and $\widehat{F}_0^b$ be the Bootstrap estimates of the potential outcomes' distributions based on the Bootstrap scheme in Section 2.4. Denote

$$\widehat{\mathbb{KS}}^b(y) = \sqrt{n} \left\{ \left[\widehat{F}_1^b(y) - \widehat{F}_1(y)\right] - \left[\widehat{F}_0^b(y) - \widehat{F}_0(y)\right] \right\}.$$

Similar to the discussion in the proof of Theorem 2, we can obtain $\widehat{\mathbb{KS}}^b(\cdot) \Rightarrow \mathbb{KS}(\cdot)$ conditional on the sample, in probability. Then, the continuous mapping theory implies that,

$$\sup_y \widehat{\mathbb{KS}}^b(y) \Rightarrow \sup_y \mathbb{KS}(y), \tag{C.3}$$

conditional on the sample in probability.

Recall that

$$\widehat{\text{KS}} = \sqrt{n} \sup_{y \in \mathcal{Y}} \left[\widehat{F}_1(y) - \widehat{F}_0(y)\right].$$

Then, under $H_0: F_1(y) \leq F_0(y)$ for all $y \in \mathcal{Y}$, it is clear that $\widehat{\text{KS}} \leq \widehat{T}$. Using the same arguments as in the proof of Proposition 1 (A) in Barrett and Donald (2003), the distribution of $\sup_y \mathbb{KS}(y)$ is absolutely continuous on $(0, \infty)$, and $c(\alpha)$ defined by $\mathbb{P}\left(\sup_y \mathbb{KS}(y) > c(\alpha)\right) = \alpha$ is finite for $\alpha < 1/2$. Note that the event that $\widehat{p} < \alpha$ is equivalent to the event $\widehat{\text{KS}} > \widehat{c}(\alpha)$, where

$$\widehat{c}(\alpha) = \inf \left\{ t : \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{\text{KS}}^b > t \right) > \alpha \right\}.$$

The result in (C.3) implies $\widehat{c}(\alpha) \xrightarrow{p} c(\alpha)$. Then, under H_0,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\left(\widehat{p} < \alpha\right) = \lim_{n\to\infty} \mathbb{P}\left(\widehat{\text{KS}} > \widehat{c}(\alpha)\right) \\
&\leq \lim_{n\to\infty} \mathbb{P}\left(\widehat{T} > \widehat{c}(\alpha)\right) = \lim_{n\to\infty} \mathbb{P}\left(\widehat{T} > c(\alpha)\right) + \lim_{n\to\infty} \left[\mathbb{P}\left(\widehat{T} > \widehat{c}(\alpha)\right) - \mathbb{P}\left(\widehat{T} > c(\alpha)\right)\right] \\
&= \mathbb{P}\left(\sup_y \mathbb{KS}(y) > c(\alpha)\right) = \alpha,
\end{aligned}$$

where the second inequality holds since $\widehat{\mathrm{KS}} \leq \widehat{T}$ under H_0 as discussed above, and the last line follows from $\widehat{T} \xrightarrow{d} \sup_y \mathbb{KS}(y)$ and $\widehat{c}(\alpha) \xrightarrow{p} c(\alpha)$. This proves part(i) in Proposition 5.

For part(ii), we note that under H_1, there exist some $y' \in \mathcal{Y}$ such that $F_1(y') - F_0(y') = \zeta > 0$. The results in (C.2) imply that $\widehat{F}_1(y') - \widehat{F}_0(y') \xrightarrow{p} \zeta$. Then,

$$\widehat{\mathrm{KS}} \geq \sqrt{n}\left(\widehat{F}_1(y') - \widehat{F}_0(y')\right) \xrightarrow{p} \infty.$$

Since $\widehat{c}(\alpha)$ is bounded in probability, we have

$$\lim_{n\to\infty} \mathbb{P}(\widehat{p} > \alpha) = \mathbb{P}\left(\widehat{\mathrm{KS}} > \widehat{c}(\alpha)\right) = 1.$$

This completes the proof of part(ii). □