Partial Identification and Confidence Sets for Functionals of the Joint Distribution of Potential Outcomes*

Yanqin Fan
Department of Economics
Vanderbilt University
VU Station B #351819
2301 Vanderbilt Place
Nashville, TN 37235-1819

Dongming Zhu
HSBC School of Business
Peking University Campus
University Town, Nanshan District
Shenzhen, 518055
China

First version: May 2009
This version: December 2009

Abstract

In this paper, we study partial identification and inference for a general class of functionals of the joint distribution of potential outcomes of a binary treatment under the strong ignorability assumption or the selection on observables assumption commonly used in evaluating average treatment effects. Members of this class of functionals include the correlation coefficient between the potential outcomes and many inequality measures of the distribution of treatment effects. We establish sharp bounds on functionals in this class and characterize conditions under which our lower and upper bounds coincide and thus point identify the true functionals. We propose nonparametric estimators of the sharp bounds, establish their asymptotic distributions, and construct asymptotically valid confidence sets for the true functionals. Another interesting finding of this paper is that under the selection on observables assumption, although the average treatment effect can be point identified from the observable covariates or from the propensity score, in general, the sharp bounds on functionals of the joint distribution of potential outcomes based on the observable covariates are tighter than the corresponding sharp bounds based on the propensity score.

Keywords: Partial identification; Inequality measure; Treatment effect; Selection on observables; Propensity score; Correlation coefficient

*We thank Stephane Bonhomme, Yingyao Hu, Simon Lee, Konrad Menzel, Stephen Shore, Richard Spady, Joerg Stoye, Tiemen Woutersen, participants of Bates White Sixth Annual Antitrust Conference, and seminar participants at City University of Hong Kong, Johns Hopkins University, and New York University for helpful comments and discussions.

1 Introduction

The fundamental problem in the evaluation of a treatment/program/policy is that we never observe the potential outcomes in all possible states. Because of this missing data problem, evaluating joint distributions of potential outcomes or distributions of treatment effects poses more challenges than evaluating average treatment effects, the latter being the focus of most work in the treatment effect literature, see Lee (2005), Abbring and Heckman (2007), Heckman and Vytlacil (2007a, b) for discussions and references. Heckman, Smith, and Clements (1997) and Abbring and Heckman (2007), among others, provide many examples demonstrating the need for evaluating joint distributions of potential outcomes, distributions of treatment effects, or other features of the distributions of treatment effects than various average treatment effects.

Recently, some progress has been made on econometric methods for evaluating joint distributions of potential outcomes or distributions of treatment effects. Depending on the assumptions made, the relevant distributions may or may not be point identified. Examples on the former typically impose restrictions on the dependence structure between the potential outcomes such that the joint distributions of potential outcomes and thus distributions of treatment effects are point identified, see, e.g., Heckman, Smith, and Clements (1997), Biddle, Boden, and Reville (2003), Carneiro, Hansen, and Heckman (2003), Aakvik, Heckman, and Vytlacil (2005), and Abbring and Heckman (2007), among others. In this case, functionals of the joint distributions of potential outcomes are point identified as well. Without imposing strong enough restrictions on the dependence structure between the potential outcomes, their joint distributions and the distributions of treatment effects are in general partially identified, see Manski (1997), Fan and Park (2007, 2008a, b), Fan and Wu (2007), Firpo and Ridder (2008), and Fan (2008). Assuming monotone treatment response, Manski (1997) developed sharp bounds on the distributions of treatment effects, while assuming the availability of ideal randomized data, Fan and Park (2007, 2008a) studied partial identification and inference tools for the distributions of treatment effects and their sharp bounds and Fan and Park (2008b) investigated partial identification and inference for the quantile of treatment effects. In the context of switching regime models, Fan and Wu (2007) studied partial identification and inference for conditional distributions of treatment effects given observable covariates. Fan (2008) extended Fan and Park (2007, 2008a) by taking into account individuals' observable characteristics. In particular, under the selection-on-observables assumption commonly used in evaluating average treatment effects, see, e.g., Rosenbaum and Rubin (1983a, b), Hahn (1998), Heckman, Ichimura, Smith, and Todd (1998a, b), Dehejia and Wahba (1999), among others, Fan (2008) provided nonparametric estimators of sharp bounds and developed inference tools for both the conditional distribution of treatment effects given the observable covariates and the unconditional distribution of treatment effects taking into account the observable covariates.

Often parameters defined as functionals of the joint distribution of potential outcomes may be of direct interest. One class of such parameters is that of inequality measures of the distribution of treatment effects. Under the selection on observables assumption, Fan and Park (2007, 2008), Fan (2008), and Firpo and Ridder (2008) established sharp bounds on the distribution of treatment effects. It is known that sharp bounds on the distribution of treatment effects in general do not imply sharp bounds on functionals of the distribution of treatment effects, as these distribution bounds are not uniformly sharp, see Firpo and Ridder (2008) and Stoye (2008). In this paper, we contribute to the recent literature on the evaluation of distributional treatment effects by studying partial identification and inference for a general class of functionals of the joint distribution of potential outcomes of a binary treatment under the selection on observables assumption. This class of functionals include the correlation coefficient between the potential outcomes and many inequality measures of the distribution of treatment effects. We establish sharp bounds on functionals in this class and characterize conditions under which our lower and upper bounds coincide and thus point identify the true functionals. We propose nonparametric estimators of the sharp bounds, establish their asymptotic distributions, and construct asymptotically valid confidence sets (CSs) for the true functionals. Another interesting finding of this paper is the role of the propensity score. It is well known that under the selection on observables assumption, the average treatment effect can be point identified from the observable covariates or from the propensity score. In contrast, we show in this paper that in general, the sharp bounds on functionals of the joint distribution of potential outcomes using the observable covariates are tighter than the corresponding sharp bounds using the propensity score.

Our partial identification results or sharp bounds on functionals of the joint distribution of potential outcomes build on the work by Cambanis, Simons, and Stout (1976), see also Tchen (1980) and Rachev and Ruschendorf (1998). These authors establish sharp bounds on the same class of functionals of a bivariate distribution as in this paper assuming fixed marginal distributions. Under the selection on observables assumption, it is well known that both the marginal distributions of the potential outcomes conditional on the observable covariates and the distribution of the observable covariates are point identified. Thus, the marginal distributions of the potential outcomes are identified as well, implying that the bounds in Cambanis, Simons, and Stout (1976) are still valid. However, we show that under the selection on observables assumption, the bounds in Cambanis, Simons, and Stout (1976) are no longer sharp and we establish sharp bounds taking into account the identified conditional marginal distributions of the potential outcomes and the distribution of the observable covariates. By comparing the bounds in Cambanis, Simons, and Stout (1976) and our sharp bounds on the correlation coefficient between the potential outcomes, we see clearly the advantage of our sharp bounds over the bounds in Cambanis, Simons, and Stout (1976). In general, the bounds in Cambanis, Simons, and Stout (1976) can not identify the sign of the correlation coefficient, but our sharp bounds can and may even point identify the correlation coefficient. For ideal randomized experiments, Heckman, Smith, and Clements (1997) concluded that the bounds on the correlation coefficient between the potential outcomes implied by the results in Cambanis, Simons, and Stout (1976) are often too wide to be informative. Our results show that by exploiting the information in the observable covariates, these bounds can be narrowed greatly provided that the selection on observables assumption holds.

This paper complements Firpo (2007) in which he defined an inequality treatment effect as the difference between the corresponding inequality measures of the marginal distributions of the potential outcomes and developed inference tools under the selection on observables assumption. Also related to this paper is Stoye (2008) in which he derived sharp bounds on many spread parameters (inequality measures) of distributions that are known to be mixtures of a known distribution and a distribution whose distribution function can be bounded. The paper that is most closely related to this paper is Firpo and Ridder (2008) in which they considered bounding a general functional of the distribution of treatment effects. Note that the bounds on a functional of the distribution of treatment effects obtained from the bounds on the distribution of treatment effects in Fan (2008), Fan and Park (2007, 2008), and Firpo and Ridder (2008) are in general not sharp, as the bounds on the distribution of treatment effects are pointwise sharp, but not uniformly sharp. Firpo and Ridder (2008) presented a general approach to establishing bounds on functionals of the distribution of treatment effects that are tighter than bounds obtained directly from bounds on the distribution of treatment effects. However, the bounds in Firpo and Ridder (2008) are not sharp. In addition, Firpo and Ridder (2008) focused exclusively on partial identification.

The rest of this paper is organized as follows. In Section 2, we study partial identification of a general class of functionals of the joint distribution of potential outcomes of a binary treatment under the selection on observables assumption. In Section 3, we propose nonparametric estimators of sharp bounds on functionals of the conditional joint distribution of potential outcomes given the covariates and construct asymptotically valid confidence sets for these functionals. In Section 4, we propose nonparametric estimators of sharp bounds on functionals of the unconditional joint distribution of potential outcomes and construct asymptotically valid confidence sets for these functionals. Section 5 concludes and presents some extensions. Technical proofs are collected in the Appendix.

Throughout the paper, we use $\Longrightarrow$ to denote weak convergence. All the limits are taken as the sample size goes to ∞.

2 Sharp Bounds on $E\left[\mu\left(Y_1, Y_0\right)\right]$

We consider a binary treatment and use $Y_1 \in \mathcal{Y}_1$ to denote the potential outcome from receiving the treatment and $Y_0 \in \mathcal{Y}_0$ the potential outcome without receiving the treatment. Let $F(y_1, y_0)$

denote the joint distribution of Y_1, Y_0 with marginals $F_1(\cdot)$ and $F_0(\cdot)$ respectively.

Let $\theta = E[\mu(Y_1, Y_0)]$, where $\mu(\cdot,\cdot)$ is a quasi-monotone (quasi-antitone) and right continuous function[1]. A function $\mu(\cdot,\cdot)$ is called quasi-monotone if for all $y_1 \leq y_1'$ and $y_0 \leq y_0'$,

$$\mu(y_1, y_0) + \mu(y_1', y_0') - \mu(y_1, y_0') - \mu(y_1', y_0) \geq 0,$$

and quasi-antitone if $-\mu(\cdot,\cdot)$ is quasi-monotone. If $\mu(\cdot,\cdot)$ is absolutely continuous, then it is quasi-monotone if and only if $\frac{\partial^2 \mu(y_1, y_0)}{\partial y_1 \partial y_0} \geq 0$ a.e. Cambanis, Simons, and Stout (1976) provide many examples of quasi-monotone or quasi-antitone functions, see also Tchen (1980). Many parameters of the joint distribution of potential outcomes of interest, including the correlation coefficient between the potential outcomes and many inequality measures of the distribution of treatment effects, can be written as functions of θ and parameters that depend on the marginal distributions of potential outcomes only.

Example 2.1. Let $\mu(Y_1, Y_0) = Y_1 Y_0$. Then μ is continuous and quasi-monotone. Let $\sigma_j^2 = Var(Y_j)$ for $j = 1, 0$. Then the correlation coefficient between Y_1 and Y_0 is given by

$$\rho_{10} = \frac{E[\mu(Y_1, Y_0)] - E(Y_1)E(Y_0)}{\sigma_1 \sigma_0}.$$

Example 2.2. Let $\Delta = Y_1 - Y_0$ denote the individual treatment effect and $\mu(Y_1, Y_0) = \nu(\Delta)$ for some function ν. Then $\mu(\cdot,\cdot)$ is continuous, quasi-monotone if $\nu(\cdot)$ is a continuous, concave function and $\mu(\cdot,\cdot)$ is continuous, quasi-antitone if $\nu(\cdot)$ is a continuous, convex function. Many inequality measures of the distribution of treatment effect Δ can be expressed as $g(E[\nu(\Delta)], \mu_\Delta)$, where $\mu_\Delta = E(\Delta)$, $g(\cdot,\cdot)$ is increasing in its first argument, and $\nu(\cdot)$ is continuous and convex, see Stoye (2008) and references therein. For instance, the coefficient of variation defined as

$$\theta_{CV} = \frac{\sqrt{Var(\Delta)}}{\mu_\Delta} = \frac{\sqrt{E(\Delta^2) - \mu_\Delta^2}}{\mu_\Delta}$$

can be written as $g(E[\nu(\Delta)], \mu_\Delta)$, where $\nu(\Delta) = \Delta^2$ is continuous and convex and $g(z, \mu_\Delta) = \frac{\sqrt{z - \mu_\Delta^2}}{\mu_\Delta}$ is increasing in z. A general class of inequality measures of the distribution of Δ is that of generalized entropy measures. For a given real-valued parameter γ, let $\nu_\gamma(\Delta) = \Delta^\gamma$ and

$$g_\gamma(z, \mu_\Delta) = \frac{1}{\gamma^2 - \gamma}\left[\frac{z}{\mu_\Delta^\gamma} - 1\right].$$

Then $\{g_\gamma(E[\nu_\gamma(\Delta)], \mu_\Delta) : \gamma \in R\}$ is the class of generalized entropy measures of the distribution of Δ. Theorem 5 in Cowell (1995) shows that any inequality measure that simultaneously satisfies the properties of the weak principle of transfers, decomposability, scale independence and the population principle must be expressible in the form $g_\gamma(E[\nu_\gamma(\Delta)], \mu_\Delta)$ or some ordinally equivalent

[1] Quasi-monotone functions are also referred to as super-modular functions or super-additive functions. Quasi-monotone and right continuous functions satisfy the “Monge” condition, see Rachev and Ruschendorf (1998).

transformation of $g_\gamma(E[\nu_\gamma(\Delta)], \mu_\Delta)$ for some $\gamma \in R$. If γ is even, then $\nu(\cdot)$ is continuous and convex.

In this section, we establish sharp bounds on $\theta = E[\mu(Y_1, Y_0)]$, where $\mu(\cdot,\cdot)$ is a quasi-monotone (quasi-antitone) and right continuous function, under the selection on observables assumption. Applying the sharp bounds on θ for the corresponding functions $\mu(\cdot,\cdot)$ in Examples 2.1 and 2.2 lead to sharp bounds on the correlation coefficient between the potential outcomes and on inequality measures of the distribution of Δ including the coefficient of variation and generalized entropy measures with even values of γ. We show via the sharp bounds on the correlation coefficient that our sharp bounds are in general much tighter than existing bounds using the marginals only.

2.1 Sharp Bounds with Fixed Marginals

Suppose the marginals F_1, F_0 are given and there is no other information available about the potential outcomes Y_1, Y_0. For example, access to data from an ideal randomized experiment would allow us to identify the marginals F_1, F_0. In this case, sharp bounds on θ for a quasi-monotone (antitone) and right continuous function $\mu(\cdot,\cdot)$ are available, see Cambanis, Simons, and Stout (1976), Tchen (1980), and Rachev and Ruschendorf (1998). Below we restate the result in Rachev and Ruschendorf (1998).

Lemma 2.1 *Let (Y_1', Y_0') be independent with the same marginals as (Y_1, Y_0). Suppose that $E[\mu(Y_1, Y_0)]$ and $E[\mu(Y_1', Y_0')]$ exist and are finite.*

(i) If $\mu(y_1, y_0)$ is quasi-monotone and right continuous, then

$$\int_0^1 \mu\left(F_1^{-1}(u), F_0^{-1}(1-u)\right) du \le \theta \le \int_0^1 \mu\left(F_1^{-1}(u), F_0^{-1}(u)\right) du,$$

where $F_j^{-1}(u) = \inf\{y : F_j(y) \ge u\}$ is the quantile function of Y_j, $j = 1, 0$.

(ii) If $\mu(y_1, y_0)$ is quasi-antitone and right continuous, then

$$\int_0^1 \mu\left(F_1^{-1}(u), F_0^{-1}(u)\right) du \le \theta \le \int_0^1 \mu\left(F_1^{-1}(u), F_0^{-1}(1-u)\right) du.$$

For $(u, v) \in [0,1]^2$, let $C^L(u,v) = \max(u + v - 1, 0)$ and $C^U(u,v) = \min(u,v)$ denote the Fréchet-Hoeffding lower and upper bounds for a copula. Suppose $\mu(y_1, y_0)$ is quasi-monotone and right continuous. Then the lower and upper bounds on θ in Lemma 2.1 (i) are achieved when $F(y_1, y_0) = C^L(F_1(y_1), F_0(y_0))$ and $F(y_1, y_0) = C^U(F_1(y_1), F_0(y_0))$ respectively. Since the latter functions are proper distribution functions with marginals F_1, F_0, the bounds on θ in Lemma 2.1 are sharp when the only information available on (Y_1, Y_0) are the marginals F_1, F_0.

Remark 2.1. Let $\left(Y_1^P, Y_0^P\right) \sim C^U(F_1(\cdot), F_0(\cdot))$. Then Y_1^P, Y_0^P have the same marginal distributions as (Y_1, Y_0) and are perfectly positively dependent. Similarly, let $\left(Y_1^N, Y_0^N\right) \sim C^L(F_1(\cdot), F_0(\cdot))$.

Then Y_1^N, Y_0^N have the same marginal distributions as (Y_1, Y_0) and are perfectly negatively dependent. The sharp bounds in Lemma 2.1 are respectively $E\left[\mu\left(Y_1^P, Y_0^P\right)\right]$ and $E\left[\mu\left(Y_1^N, Y_0^N\right)\right]$.

Let

$$\theta^L = \int_0^1 \mu\left(F_1^{-1}(u), F_0^{-1}(1-u)\right) du \text{ and } \theta^U = \int_0^1 \mu\left(F_1^{-1}(u), F_0^{-1}(u)\right) du.$$

They are respectively the lower and upper bounds on θ when $\mu(\cdot,\cdot)$ is quasi-monotone and right continuous. In the rest of this paper, we'll focus exclusively on quasi-monotone functions, as with quasi-antitone functions, we only need to switch the lower and upper bounds. If $\mu(\cdot,\cdot)$ is additively separable in its arguments, then $\theta^L = \theta^U$ in which case θ is point identified for all the marginal distribution functions F_1, F_0 satisfying the conditions of Lemma 2.1. If $\mu(\cdot,\cdot)$ is not additively separable in its arguments, then in general $\theta^L \neq \theta^U$ and θ is only partially identified. Below we show that for a general class of functions $\mu(\cdot,\cdot)$, θ is point identified only in trivial cases, i.e., when at least one of the marginal distributions F_1, F_0 is degenerate at a finite value.

Definition 2.1 *A function $\mu(\cdot,\cdot)$ is called "strict quasi-monotone" if it is quasi-monotone and for all $y_1 < y_1'$ and $y_0 < y_0'$,*

$$\mu(y_1, y_0) + \mu\left(y_1', y_0'\right) - \mu\left(y_1, y_0'\right) - \mu\left(y_1', y_0\right) > 0,$$

and is strict quasi-antitone if $-\mu(\cdot,\cdot)$ is strict quasi-monotone.

A quasi-monotone or quasi-antitone function can be additively separable in its arguments, but a strict quasi-monotone or a strict quasi-antitone function can not be additively separable in its arguments. For example, $\mu(y_1, y_0) = y_1 - y_0$ is quasi-monotone but not strict quasi-monotone, while $\mu(y_1, y_0) = y_1 y_0$ is strict quasi-monotone.

Proposition 2.2 *Suppose the conditions of Lemma 2.1 hold. (i) If $\mu(\cdot,\cdot)$ is additively separable, then $\theta^L = \theta^U$. (ii) If $\mu(\cdot,\cdot)$ is strict quasi-monotone or strict quasi-antitone, then $\theta^L = \theta^U$ if and only if at least one of the marginal distributions F_1, F_0 is degenerate at a finite value.*

Consider the treatment effect $\Delta = Y_1 - Y_0$. Proposition 2.2 (i) implies that the sharp bounds on $E(\Delta)$ in Lemma 2.1 coincide and point identify $E(\Delta)$. In contrast, Proposition 2.2 (ii) implies that the sharp bounds on $Var(\Delta)$ coincide only in the case where one of the marginal distributions F_1, F_0 is degenerate at a finite value. As a result, inferences for $E(\Delta)$ and $Var(\Delta)$ are fundamentally different.

When the marginals are given, functionals that depend on the marginals only are identified. Using Lemma 2.1, we can establish sharp bounds on the correlation coefficient between the potential outcomes and inequality measures of the distribution of Δ introduced in Example 2.2.

Example 2.1 (Contd). Applying Lemma 2.1 (i) to the correlation coefficient between the potential outcomes, we obtain $\rho^L \leq \rho_{10} \leq \rho^U$, where

$$\rho^L = \frac{\int_0^1 F_1^{-1}(u) F_0^{-1}(1-u)\,du - E(Y_1)E(Y_0)}{\sigma_1\sigma_0},$$
$$\rho^U = \frac{\int_0^1 F_1^{-1}(u) F_0^{-1}(u)\,du - E(Y_1)E(Y_0)}{\sigma_1\sigma_0}.$$

For ideal randomized experiments, Heckman, Smith, and Clements (1997) used these bounds to bound the variance of treatment effects and the correlation coefficient between the two potential outcomes. They found that the bounds are typically too wide to be informative. For example, since $\rho^L \leq 0$ and $\rho^U \geq 0$, we can not identify the sign of ρ_{10} using ρ^L, ρ^U. In fact, we can show that if F_1 and F_0 are in the same location-scale family, then $\rho_L = -1$ and $\rho_U = 1$.

Example 2.2 (Contd). Suppose γ is even. Since $g_\gamma(\cdot,\cdot)$ is increasing in its first argument, we obtain:

$$\begin{aligned} & g_\gamma\left(\int_0^1 \nu_\gamma\left(F_1^{-1}(u) - F_0^{-1}(1-u)\right) du, \mu_\Delta\right) \\ \leq\ & g_\gamma\left(E\left[\nu_\gamma(\Delta)\right], \mu_\Delta\right) \\ \leq\ & g_\gamma\left(\int_0^1 \nu_\gamma\left(F_1^{-1}(u) - F_0^{-1}(u)\right) du, \mu_\Delta\right). \end{aligned}$$

Noting that $\mu_\Delta = E(Y_1) - E(Y_0)$, we conclude that the bounds on $g_\gamma(E[\nu_\gamma(\Delta)], \mu_\Delta)$ are point identified and $g_\gamma(E[\nu_\gamma(\Delta)], \mu_\Delta)$ is partially identified.

2.2 Sharp Bounds With Covariates

In this subsection, we show that the bounds in Lemma 2.1 can be tightened when covariates are available. Let $X \in \mathcal{X} \subset \mathcal{R}^d$ denote the observable vector of covariates and D the binary variable indicating participation; $D = 1$ if the individual belongs to the treatment group and $D = 0$ if the individual belongs to the control group. Further let $Y = Y_1 D + Y_0(1-D)$ denote the observable outcome for the individual. In the literature on program evaluation, the following two assumptions are often used to evaluate average treatment effects, see e.g., Rosenbaum and Rubin (1983a,b), Hahn (1998), Heckman, Ichimura, Smith, and Todd (1998a, b), Dehejia and Wahba (1999), and Hirano, Imbens, and Ridder (2003), to name only a few.

(C1) Let (Y_1, Y_0, D, X) have a joint distribution. For all $x \in \mathcal{X}$, (Y_1, Y_0) is jointly independent of D conditional on $X = x$.

(C2) For all $x \in \mathcal{X}$, $0 < p(x) < 1$, where $p(x) = P(D = 1|x)$.

(C1) is a conditional independence assumption and (C2) is a support assumption. Together, they are referred to as the strong ignorability assumption or the selection on observables assumption.

Under (C1) and (C2), the conditional distributions $F_1(y|x)$ and $F_0(y|x)$ are point identified:

$$\begin{aligned} F_1(y|x) &= P(Y_1 \leq y|X=x) = P(Y_1 \leq y|X=x, D=1) \\ &= P(Y \leq y|X=x, D=1), \end{aligned} \tag{1}$$

and

$$F_0(y|x) = P(Y \leq y|X=x, D=0). \tag{2}$$

Moreover, the marginal distributions $F_1(y)$, $F_0(y)$ are also point identified[2]:

$$F_1(y) = E\left[\frac{D}{p(X)} I\{Y \leq y\}\right] \text{ and } F_0(y) = E\left[\frac{1-D}{1-p(X)} I\{Y \leq y\}\right].$$

Since the marginal distributions F_1, F_0 are point identified under the selection on observables assumption, the bounds on θ in Lemma 2.1 are still valid. However, we show in the following proposition that in general they are not sharp when covariates are available such that (C1) and (C2) hold. Intuitively, this is because under (C1) and (C2), we have more information on (Y_1, Y_0) than just the marginals F_1, F_0 as characterized by the conditional marginals $F_1(\cdot|x), F_0(\cdot|x)$, and the distribution of X.

Proposition 2.3 *Suppose the conditions of Lemma 2.1 and (C1)-(C2) hold. Let $\mu(y_1, y_0)$ be a quasi-monotone and right continuous function. Then (i)*

$$E\left[\int_0^1 \mu\left(F_1^{-1}(u|X), F_0^{-1}(1-u|X)\right) du\right] \leq \theta \leq E\left[\int_0^1 \mu\left(F_1^{-1}(u|X), F_0^{-1}(u|X)\right) du\right],$$

where $F_j^{-1}(u|x) = \inf\{y : F_j(y|x) \geq u\}$ is the quantile function of Y_j conditional on $X = x$, $j = 1, 0$. Moreover the bounds are sharp.

(ii) Let

$$\begin{aligned} \theta_L &= E\left[\int_0^1 \mu\left(F_1^{-1}(u|X), F_0^{-1}(1-u|X)\right) du\right], \\ \theta_U &= E\left[\int_0^1 \mu\left(F_1^{-1}(u|X), F_0^{-1}(u|X)\right) du\right]. \end{aligned}$$

Suppose $\mu(\cdot,\cdot)$ is strict quasi-monotone. Then $\theta_L = \theta_U$ if and only if at least one of the conditional marginal distributions $F_1(\cdot|x), F_0(\cdot|x)$ is degenerate at a finite value for every $x \in \mathcal{X}$.

Proposition 2.3 establishes sharp bounds on θ under the selection on observables assumption. As the bounds on θ in Lemma 2.1 are still valid, the bounds in Proposition 2.3 in general improve on the corresponding bounds in Lemma 2.1. As we show in the next subsection, using the sharp bounds on the correlation coefficient between the potential outcomes obtained from Proposition 2.3, it is possible to identify the sign of the correlation coefficient between Y_1 and Y_0.

[2] We refer the reader to Firpo (2007) for a detailed discussion on this.

Remark 2.2. Often we may be interested in $\theta(x_1) = E\left[\mu(Y_1, Y_0)|X^1 = x_1\right]$, where $X = \left(X^1, X^{-1}\right)$. When (C1) and (C2) hold, the sharp bounds on $\theta(x_1)$ are given by

$$E\left[\int_0^1 \mu\left(F_1^{-1}\left(u|x_1, X^{-1}\right), F_0^{-1}\left(1-u|x_1, X^{-1}\right)\right) du\right],$$
$$E\left[\int_0^1 \mu\left(F_1^{-1}\left(u|x_1, X^{-1}\right), F_0^{-1}\left(u|x_1, X^{-1}\right)\right) du\right].$$

Rosenbaum and Rubin (1983a,b) show that under (C1) and (C2), (Y_1, Y_0) is also jointly independent of D conditional on the propensity score $p(X)$. Thus, if $\mu(y_1, y_0)$ is quasi-monotone and right continuous, then $\theta_{LP} \leq \theta \leq \theta_{UP}$, where

$$\begin{aligned}
\theta_{LP} &= E\left[\int_0^1 \mu\left(F_1^{-1}(u|p(X)), F_0^{-1}(1-u|p(X))\right) du\right], \\
\theta_{UP} &= E\left[\int_0^1 \mu\left(F_1^{-1}(u|p(X)), F_0^{-1}(u|p(X))\right) du\right].
\end{aligned}$$

If the conditional distribution functions of Y_1, Y_0 given X are the same as the conditional distribution functions of Y_1, Y_0 given $p(X)$, then the bounds on θ based on the propensity score are identical to the sharp bounds based on X; otherwise the bounds on θ based on the propensity score are wider than the sharp bounds based on X as shown in the following proposition.

Proposition 2.4 *Suppose the conditions of Lemma 2.1 and (C1)-(C2) hold. Let $\mu(y_1, y_0)$ be a quasi-monotone and right continuous function. Then $\theta_{LP} \leq \theta_L \leq \theta_U \leq \theta_{UP}$.*

The result in Proposition 2.4 is an interesting observation; it is in sharp contrast to the identification of the average treatment effect, as under (C1) and (C2), μ_Δ can be identified either by conditioning on X or conditioning on $p(X)$:

$$\begin{aligned}
\mu_\Delta &= E\left(E[Y_1|X, D=1] - E[Y_0|X, D=0]\right) \\
&= E\left(E[Y_1|p(X), D=1] - E[Y_0|p(X), D=0]\right).
\end{aligned}$$

Proposition 2.4 shows that for parameter θ that is partially identified, the use of the full vector of covariates X provides tighter bounds than the propensity score $p(X)$.

Remark 2.3. A generalization of Proposition 2.4 implies that the sharp bounds on θ using the maximal relevant information set such that (C1) and (C2) hold are the tightest, see Heckman and Navarro-Lozano (2004) for discussions on the maximal relevant information set.

2.3 Sharp Bounds on the Correlation Coefficient

In this subsection, we present two examples to show that with covariate X, the sharp bounds in Proposition 2.3 are more informative than the bounds in Lemma 2.1. Suppose X is univariate. Assume (X, W_j) follows a bivariate normal distribution:

$$\begin{pmatrix} W_j \\ X \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_j^2 & \sigma_j\rho_{jX} \\ \sigma_j\rho_{jX} & 1 \end{pmatrix}\right], \; j = 1, 0.$$

Then $W_j|X = x \sim N\left(\sigma_j \rho_{jX} x, \sigma_j^2 \left(1-\rho_{jX}^2\right)\right)$.

Example 2.1.1. Let $Y_j = W_j$, $j = 1, 0$. Then $Y_j \sim N\left(0, \sigma_j^2\right)$. Suppose $\sigma_j^2 > 0$, $j = 1, 0$. Using Lemma 2.1, we get $\rho_{10} \in [-1, 1]$, so ρ_{10} is not identified. Now, we know $Y_j|X = x \sim N\left(\sigma_j \rho_{jX} x, \sigma_j^2 \left(1-\rho_{jX}^2\right)\right)$ and $X \sim N(0,1)$. Proposition 2.3 yields: $\rho_L \leq \rho_{10} \leq \rho_U$, where

$$\begin{aligned}
\rho_L &= \rho_{0X}\rho_{1X} - \sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)}, \\
\rho_U &= \rho_{0X}\rho_{1X} + \sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)}.
\end{aligned}$$

Obviously, when $\rho_{0X}\rho_{1X} > 0$ and $\rho_{0X}^2 + \rho_{1X}^2 > 1$, we have $0 < \rho_L \leq \rho_U$, so ρ_{10} is positive and when $\rho_{0X}\rho_{1X} < 0$ and $\rho_{0X}^2 + \rho_{1X}^2 > 1$, we have $\rho_L \leq \rho_U < 0$, so ρ_{10} is negative. In addition, if one of ρ_{jX}^2 is one, then ρ_{10} is point identified at either $\rho_{0X} sign(\rho_{1X})$ or $\rho_{1X} sign(\rho_{0X})$.

Example 2.1.2. Let $F_1(\cdot|x)$ and $F_0(\cdot|x)$ denote univariate log-normal distribution functions with parameters $\left(\sigma_1\rho_{1X}x, \sigma_1^2\left(1-\rho_{1X}^2\right)\right)$ and $\left(\sigma_0\rho_{0X}x, \sigma_0^2\left(1-\rho_{0X}^2\right)\right)$ respectively. Let $\rho_{10}(x)$ denote the conditional correlation coefficient between Y_1 and Y_0 given $X = x$. Using Lemma 2.1 (i), we obtain $\rho_L(x) \leq \rho_{10}(x) \leq \rho_U(x)$, where

$$\rho_L(x) = \frac{\exp(-\sigma_1\sigma_0\sqrt{\left(1-\rho_{1X}^2\right)\left(1-\rho_{0X}^2\right)}) - 1}{\sqrt{(\exp(\sigma_1^2\left(1-\rho_{1X}^2\right)) - 1)(\exp(\sigma_0^2\left(1-\rho_{0X}^2\right)) - 1)}} \leq 0, \tag{3}$$

$$\rho_U(x) = \frac{\exp(\sigma_1\sigma_0\sqrt{\left(1-\rho_{1X}^2\right)\left(1-\rho_{0X}^2\right)}) - 1}{\sqrt{(\exp(\sigma_1^2\left(1-\rho_{1X}^2\right)) - 1)(\exp(\sigma_0^2\left(1-\rho_{0X}^2\right)) - 1)}} \geq 0. \tag{4}$$

Now consider the unconditional correlation coefficient between Y_1 and Y_0 denoted by ρ_{10}. It is easy to show that

$$\rho_{10} = \frac{E\left[\rho_{10}(X)\sqrt{Var(Y_1|X)\,Var(Y_0|X)}\right] + E\left[E(Y_1|X)\,E(Y_0|X)\right] - E(Y_1)\,E(Y_0)}{\sqrt{Var(Y_1)\,Var(Y_0)}}.$$

Thus, it satisfies: $\rho_L \leq \rho_{10} \leq \rho_U$, where

$$\rho_L = \frac{E\left[\rho_L(X)\sqrt{Var(Y_1|X)\,Var(Y_0|X)}\right] + E\left[E(Y_1|X)\,E(Y_0|X)\right] - E(Y_1)\,E(Y_0)}{\sqrt{Var(Y_1)\,Var(Y_0)}}, \tag{5}$$

$$\rho_U = \frac{E\left[\rho_U(X)\sqrt{Var(Y_1|X)\,Var(Y_0|X)}\right] + E\left[E(Y_1|X)\,E(Y_0|X)\right] - E(Y_1)\,E(Y_0)}{\sqrt{Var(Y_1)\,Var(Y_0)}}. \tag{6}$$

Now we evaluate the individual components in the expressions for ρ_L and ρ_U. Noting that

$$E(Y_j|X) = \exp\left[\sigma_j\rho_{jX}X + \frac{1}{2}\sigma_j^2\left(1-\rho_{jX}^2\right)\right], \tag{7}$$

$$E\left(Y_j^2|X\right) = \exp\left[2\sigma_j\rho_{jX}X + 2\sigma_j^2\left(1-\rho_{jX}^2\right)\right], \tag{8}$$

we obtain:

$$Var(Y_j|X) = \left(\exp\left[\sigma_j^2\left(1-\rho_{jX}^2\right)\right] - 1\right)\exp\left[2\sigma_j\rho_{jX}X + \sigma_j^2\left(1-\rho_{jX}^2\right)\right]. \tag{9}$$

Then

$$\begin{aligned}
E\left(Y_j\right) &= E\left[E\left(Y_j|X\right)\right] = \exp\left[\frac{1}{2}\sigma_j^2\left(1-\rho_{jX}^2\right)\right]E\left(\exp\left[\sigma_j\rho_{jX}X\right]\right) \\
&= \exp\left[\frac{1}{2}\sigma_j^2\left(1-\rho_{jX}^2\right)\right]\exp\left[\frac{1}{2}\sigma_j^2\rho_{jX}^2\right] = \exp\left[\frac{1}{2}\sigma_j^2\right], \\
E\left(Y_j^2\right) &= E\left[E\left(Y_j^2|X\right)\right] = \exp\left[2\sigma_j^2\left(1-\rho_{jX}^2\right)\right]E\left(\exp\left[2\sigma_j\rho_{jX}X\right]\right) \\
&= \exp\left[2\sigma_j^2\left(1-\rho_{jX}^2\right)\right]\exp\left[\frac{1}{2}4\sigma_j^2\rho_{jX}^2\right] = \exp\left[2\sigma_j^2\right].
\end{aligned}$$

Using the facts in (7)-(9), we obtain:

$$\sqrt{Var\left(Y_1\right)Var\left(Y_0\right)} = \sqrt{\left(\exp\left[\sigma_0^2\right]-1\right)\left(\exp\left[\sigma_1^2\right]-1\right)}\exp\left[\frac{1}{2}\left(\sigma_0^2+\sigma_1^2\right)\right], \tag{10}$$

$$\begin{aligned}
& E\left[E\left(Y_1|X\right)E\left(Y_0|X\right)\right] - E\left(Y_1\right)E\left(Y_0\right) \\
=\ & \exp\left(\frac{1}{2}\left[\sigma_1^2\left(1-\rho_{1X}^2\right)+\sigma_0^2\left(1-\rho_{0X}^2\right)\right]\right)E\exp\left[\left(\sigma_1\rho_{1X}+\sigma_0\rho_{0X}\right)X\right] - \exp\left[\frac{1}{2}\left(\sigma_0^2+\sigma_1^2\right)\right] \\
=\ & \exp\left(\frac{1}{2}\left[\sigma_1^2\left(1-\rho_{1X}^2\right)+\sigma_0^2\left(1-\rho_{0X}^2\right)\right]\right)\exp\left(\frac{1}{2}\left[\sigma_1\rho_{1X}+\sigma_0\rho_{0X}\right]^2\right) - \exp\left[\frac{1}{2}\left(\sigma_0^2+\sigma_1^2\right)\right] \\
=\ & \exp\left(\frac{1}{2}\left[\sigma_1^2+\sigma_0^2\right]\right)\left[\exp\left(\sigma_1\sigma_0\rho_{1X}\rho_{0X}\right)-1\right], \qquad (11)
\end{aligned}$$

and

$$\begin{aligned}
& E\left[\rho_L\left(X\right)\sqrt{Var\left(Y_1|X\right)Var\left(Y_0|X\right)}\right] \\
=\ & \left(\exp\left[-\sigma_0\sigma_1\sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)}\right]-1\right)\exp\left[\frac{\sigma_0^2\left(1-\rho_{0X}^2\right)+\sigma_1^2\left(1-\rho_{1X}^2\right)}{2}\right] \\
& \cdot E\left\{\exp\left[\left(\sigma_0\rho_{0X}+\sigma_1\rho_{1X}\right)X\right]\right\} \\
=\ & \left(\exp\left[-\sigma_0\sigma_1\sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)}\right]-1\right)\exp\left[\frac{\sigma_0^2+\sigma_1^2+2\sigma_0\sigma_1\rho_{0X}\rho_{1X}}{2}\right]. \qquad (12)
\end{aligned}$$

Similarly,

$$\begin{aligned}
& E\left[\rho_U\left(X\right)\sqrt{Var\left(Y_1|X\right)Var\left(Y_0|X\right)}\right] \\
=\ & \left(\exp\left[\sigma_0\sigma_1\sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)}\right]-1\right)\exp\left[\frac{\sigma_0^2+\sigma_1^2+2\sigma_0\sigma_1\rho_{0X}\rho_{1X}}{2}\right]. \qquad (13)
\end{aligned}$$

Substituting (10), (11), (12) and (13) into (5) and (6) yields

$$\begin{aligned}
\rho_L &= \frac{\exp\left[\sigma_0\sigma_1\left(\rho_{0X}\rho_{1X}-\sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)}\right)\right]-1}{\sqrt{\left(\exp\left[\sigma_0^2\right]-1\right)\left(\exp\left[\sigma_1^2\right]-1\right)}}, \\
\rho_U &= \frac{\exp\left[\sigma_0\sigma_1\left(\rho_{0X}\rho_{1X}+\sqrt{\left(1-\rho_{0X}^2\right)\left(1-\rho_{1X}^2\right)}\right)\right]-1}{\sqrt{\left(\exp\left[\sigma_0^2\right]-1\right)\left(\exp\left[\sigma_1^2\right]-1\right)}}.
\end{aligned}$$

Again, when $\rho_{0X}\rho_{1X} > 0$ and $\rho_{0X}^2 + \rho_{1X}^2 > 1$, we have $0 < \rho_L \leq \rho_U$; when $\rho_{0X}\rho_{1X} < 0$ and $\rho_{0X}^2 + \rho_{1X}^2 > 1$, we have $\rho_L \leq \rho_U < 0$. This implies that we can identify the sign of ρ_{10} when $\rho_{0X}^2 + \rho_{1X}^2 > 1$.

Note that by (9), $Var(Y_j|X = x) = 0$ holds for every x if and only if $\rho_{jX} = \pm 1$, $j = 1, 0$. Suppose $\rho_{1X} = 1$, then $F_1(\cdot|x)$ is degenerate at $\exp[\sigma_1 x]$ for all x and ρ_{10} is point identified at

$$\rho_{10} = \rho_L = \rho_U = \frac{\exp[\sigma_0\sigma_1\rho_{0X}] - 1}{\sqrt{\left(\exp\left[\sigma_0^2\right] - 1\right)\left(\exp\left[\sigma_1^2\right] - 1\right)}}.$$

Further assuming $\sigma_0 = \sigma_1 = \sigma$, we have

$$\rho_{10} = \rho_L = \rho_U = \frac{\exp\left[\sigma^2\rho_{0X}\right] - 1}{\exp\left[\sigma^2\right] - 1} \in \left[-\exp\left(-\sigma^2\right), 1\right],$$

implying that if σ is small enough, a large range of ρ_{10} can be point identified.

3 Estimation and Confidence Sets for $\theta(x)$

Suppose $\mu(\cdot,\cdot)$ is strict quasi-monotone and right continuous. Let

$$\theta_L(x) = \int_0^1 \mu\left(F_1^{-1}(u|x), F_0^{-1}(1-u|x)\right) du, \; \theta_U(x) = \int_0^1 \mu\left(F_1^{-1}(u|x), F_0^{-1}(u|x)\right) du.$$

Lemma 2.1 and Proposition 2.2 imply that $\theta(x)$ is partially identified: $\theta_L(x) \leq \theta(x) \leq \theta_U(x)$ and $\theta_L(x) = \theta_U(x)$ if and only if one of the conditional marginal distributions $F_1(\cdot|x), F_0(\cdot|x)$ is degenerate.

Suppose a random sample $\{Y_i, X_i, D_i\}_{i=1}^n$ on $\{Y, X, D\}$ is available. For $j = 1, 0$, let

$$\begin{aligned}
\widehat{F}_j(y|x) &= \frac{\sum_{i=1}^n I\{Y_i \leq y\} K\left(\frac{X_i - x}{a}\right) I\{D_i = j\}}{\sum_{i=1}^n K\left(\frac{X_i - x}{a}\right) I\{D_i = j\}} \\
&= \frac{\sum_{i=1}^n I\{Y_i \leq y\} K_{aj}(X_i - x)}{\sum_{i=1}^n K_{aj}(X_i - x)},
\end{aligned}$$

where

$$K_{aj}(X_i - x) = K\left(\frac{X_i - x}{a}\right) I\{D_i = j\},$$

in which $K(\cdot)$ is a kernel function and $a = a_n \to 0$ is a bandwidth. The lower and upper bounds on $\theta(x)$ can be estimated by:

$$\widehat{\theta}_L(x) = \int_0^1 \mu\left(\widehat{F}_1^{-1}(u|x), \widehat{F}_0^{-1}(1-u|x)\right) du, \; \widehat{\theta}_U(x) = \int_0^1 \mu\left(\widehat{F}_1^{-1}(u|x), \widehat{F}_0^{-1}(u|x)\right) du.$$

Remark 3.1. For a given x, if $F_j(\cdot|x)$ is degenerate at $E(Y_j|X = x)$, then with probability one, we get: $Y_{ji} = E(Y_j|X = x)$ for $i = 1, ..., n$, which implies: with probability one,

$$\widehat{F}_j(y|x) = \begin{cases} 1 & \text{if } y \geq E(Y_j|X = x) \\ 0 & \text{if } y < E(Y_j|X = x) \end{cases}$$

and hence $\widehat{\theta}_L(x) = \widehat{\theta}_U(x)$ with probability one. In addition, if both $F_1(\cdot|x)$ and $F_0(\cdot|x)$ are degenerate, then $\widehat{\theta}_L(x) = \widehat{\theta}_U(x) = \theta_L(x) = \theta_U(x)$ with probability one.

Throughout the rest of this section, we assume that x is a fixed point in the interior of $\mathcal{X}$. To establish the asymptotic distribution of $\left(\widehat{\theta}_L(x), \widehat{\theta}_U(x)\right)'$, we introduce the following assumptions. Let $f_j(y \mid x) \equiv \partial F_j(y \mid x)/\partial y$.

(A1P) (i) Suppose X is continuous with density function $f_X(\cdot)$ and $f_X(x) > 0$. Further, assume both p and f_X are bounded with bounded second order partial derivatives. (ii) The kernel function $K(\cdot)$ is a bounded and symmetric density function satisfying: $\lim_{u\to\infty} \|u\|^d K(u) = 0$ and $\left\|\int vv^T K(v)\,dv\right\| < \infty$. (iii) The bandwidth sequence a_n satisfies $a_n \to 0$, $na_n^d/\log n \to \infty$ and $na_n^{4+d} \to 0$.

(A2P) Assume (i) for each $y \in \mathcal{Y}_j$, $F_j(y|\cdot)$ is bounded with bounded second order partial derivatives such that $\sup_{y\in\mathcal{Y}_j} \left|\left(\partial^2/\partial x_i \partial x_l\right) F_j(y|x)\right| < \infty$ for all $1 \le i, l \le d$, (ii) $f_j(\cdot|x)$ exists, is bounded with bounded second order partial derivatives, and (iii) $\mathcal{Y}_j$ is compact and $f_j(\cdot|x)$ satisfies $\inf_{y\in\mathcal{Y}_j} f_j(y \mid x) > 0$.

(A3P) $\mu(y_1, y_0)$ is differentiable on $\mathcal{Y}_1 \times \mathcal{Y}_0$ with uniformly continuous and bounded partial derivatives.

Let $p_1(x) = p(x)$, $p_0(x) = 1 - p(x)$, and $f_{jX}(x) = f_X(x)\,p_j(x)$, $j = 1, 0$. Further for $j = 1, 0$, let $\mu_j'(y_1, y_0) \equiv \partial\mu(y_1, y_0)/\partial y_j$,

$$G_{1L}(x,y) \equiv \mu_1'\left(y, F_0^{-1}\left(1 - F_1(y|x)\,|x\right)\right), \quad G_{0L}(x,y) \equiv \mu_0'\left(F_1^{-1}\left(1 - F_0(y|x)\,|x\right), y\right),$$

$$G_{1U}(x,y) \equiv \mu_1'\left(y, F_0^{-1}\left(F_1(y|x)\,|x\right)\right), \text{ and } G_{0U}(x,y) \equiv \mu_0'\left(F_1^{-1}\left(F_0(y|x)\,|x\right), y\right).$$

We are now ready to state the main theorem of this section.

THEOREM 3.1 *Suppose (C1)-(C2) and (A1P)-(A3P) hold. Then*

$$\sqrt{na_n^d}\begin{pmatrix} \widehat{\theta}_L(x) - \theta_L(x) \\ \widehat{\theta}_U(x) - \theta_U(x) \end{pmatrix} \Longrightarrow N\left[0, \begin{pmatrix} \sigma_L^2(x) & \sigma_{LU}(x) \\ \sigma_{LU}(x) & \sigma_U^2(x) \end{pmatrix}\right],$$

where for $j = 1, 0$ and $k = L, U$, $\sigma_k^2(x) = \sigma_{0k}^2(x) + \sigma_{1k}^2(x)$, $\sigma_{LU}(x) = \sigma_{0LU}(x) + \sigma_{1LU}(x)$ in which

$$\sigma_{jk}^2(x) = \frac{\int K^2(t)\,dt}{f_{jX}(x)} \int\int G_{jk}(x,u) G_{jk}(x,v) \left[F_j(\min(u,v)\,|x) - F_j(u|x)\,F_j(v|x)\right] dudv, \tag{14}$$

$$\sigma_{jLU}(x) = \frac{\int K^2(t)\,dt}{f_{jX}(x)} \int\int G_{jL}(x,u) G_{jU}(x,v) \left[F_j(\min(u,v)\,|x) - F_j(u|x)\,F_j(v|x)\right] dudv. \tag{15}$$

Below we provide a discussion of the role of each of the Assumptions (A1P)-(A3P) in the proof of Theorem 3.1 which is accomplished in the following steps.

Step 1. We show that the random function

$$(y_1, y_0) \mapsto \sqrt{na_n^d} \begin{pmatrix} \widehat{F}_1(y_1|x) - F_1(y_1|x) \\ \widehat{F}_0(y_0|x) - F_0(y_0|x) \end{pmatrix}$$

converges weakly to a bivariate Gaussian process $\mathcal{G} = (\mathcal{G}_1, \mathcal{G}_0)'$ with mean zero and covariance function

$$E\left(\mathcal{G}(y_1, y_0)\mathcal{G}^T\left(y_1', y_0'\right)\right) = E\begin{pmatrix} \mathcal{G}_1(y_1)\mathcal{G}_1^T(y_1') & \mathcal{G}_0(y_0)\mathcal{G}_1^T(y_1') \\ \mathcal{G}_1(y_1)\mathcal{G}_0^T(y_0') & \mathcal{G}_0(y_0)\mathcal{G}_0^T(y_0') \end{pmatrix},$$

where $E\left(\mathcal{G}_0(y_0)\mathcal{G}_1^T(y_1')\right) = E\left(\mathcal{G}_1(y_1)\mathcal{G}_0^T(y_0')\right) = 0$ and for $j = 1, 0$,

$$E\left(\mathcal{G}_j(y_j)\mathcal{G}_j^T\left(y_j'\right)\right) = \frac{1}{f_{jX}(x)}\left[\int K^2(t)\,dt\right]\left\{F_j\left(\min\left\{y_j, y_j'\right\}|x\right) - F_j(y_j|x)\,F_j\left(y_j'|x\right)\right\},$$

and the convergence is in $D(\mathcal{Y}_1) \times D(\mathcal{Y}_0)$. (A1P) and (A2P) (i)(ii) are used in this step. They are the same as the corresponding assumptions used in Polonik and Yao (2002).

Step 2. We show that the map $\phi_{F_1,F_0} : D(\mathcal{Y}_1) \times D(\mathcal{Y}_0) \to \mathcal{R}^2$ defined as

$$\phi_{F_1,F_0} = \begin{pmatrix} \int_0^1 \mu\left(F_1^{-1}(u|x), F_0^{-1}(1-u|x)\right) du \\ \int_0^1 \mu\left(F_1^{-1}(u|x), F_0^{-1}(u|x)\right) du \end{pmatrix}$$

is Hadamard differentiable at (F_1, F_0) tangentially to $C(\mathcal{Y}_1) \times C(\mathcal{Y}_0)$ with the derivative

$$\phi_{F_1,F_0}'(h_1, h_0) \mapsto -\begin{pmatrix} \int_0^1 G_{1L}(x,y)\,h_1(y)\,dy + \int_0^1 G_{0L}(x,y)\,h_0(y)\,dy \\ \int_0^1 G_{1U}(x,y)\,h_1(y)\,dy + \int_0^1 G_{0U}(x,y)\,h_0(y)\,dy \end{pmatrix}.$$

(A2P)(iii) and (A3P) are used in this step. (A2P)(iii) ensures that the quantile functions $F_1^{-1}(\cdot|x), F_0^{-1}(\cdot|x)$ are Hadamard differentiable at (F_1, F_0) tangentially to $C(\mathcal{Y}_1) \times C(\mathcal{Y}_0)$, see van der Vaart and Wellner (1996). (A3P) ensures that the map $\psi_{Q_1,Q_0} : C((0,1)) \times C((0,1)) \to \mathcal{R}^2$ defined as

$$\psi_{Q_1,Q_0} = \begin{pmatrix} \int_0^1 \mu\left(Q_1(u|x), Q_0(1-u|x)\right) du \\ \int_0^1 \mu\left(Q_1(u|x), Q_0(u|x)\right) du \end{pmatrix}$$

is Hadamard differentiable at (Q_1, Q_0).

Step 3. By the Delta method, we obtain

$$\sqrt{na_n^d}\begin{pmatrix} \widehat{\theta}_L(x) - \theta_L(x) \\ \widehat{\theta}_U(x) - \theta_U(x) \end{pmatrix} \Longrightarrow \phi_{F_1,F_0}'(\mathcal{G}_1, \mathcal{G}_0) \sim N\left[0, \begin{pmatrix} \sigma_L^2(x) & \sigma_{LU}(x) \\ \sigma_{LU}(x) & \sigma_U^2(x) \end{pmatrix}\right].$$

Remark 3.2. We note that Assumption (A2P) excludes the case that either $F_1(\cdot|x)$ or $F_0(\cdot|x)$ is degenerate. (A2P) (iii) may be relaxed at the expense of a more tedious proof analogous to the proof of Claim 1 in Bhattacharya (2007), see also Goldie (1977).

Theorem 3.1 can be used to make inference on $\theta(x)$. Let $\nabla(x) = \theta_U(x) - \theta_L(x)$. It follows from Proposition 2.2 that $\nabla(x) = 0$ if and only if at least one of $F_1(\cdot|x), F_0(\cdot|x)$ is degenerate. As

(A2P) excludes the case that at least one of $F_1(\cdot|x), F_0(\cdot|x)$ is degenerate, we only need to consider the case $\nabla(x) > 0$. Following Horowitz and Manski (2000), Imbens and Manski (2004), we define

$$CS_n(x) = \left[\widehat{\theta}_L(x) - \frac{\widehat{\sigma}_L^2(x)}{\sqrt{na_n^d}} z_{1-\alpha}, \widehat{\theta}_U(x) + \frac{\widehat{\sigma}_U^2(x)}{\sqrt{na_n^d}} z_{1-\alpha}\right],$$

where $z_{1-\alpha}$ is the $(1-\alpha)$ quantile of the standard normal distribution and for $k = L, U$, $\widehat{\sigma}_k^2(x) = \widehat{\sigma}_{0k}^2(x) + \widehat{\sigma}_{1k}^2(x)$,

$$\widehat{\sigma}_{jk}^2(x) = \frac{\int K^2(t)\,dt}{\widehat{f}_{jX}(x)} \int\int \widehat{G}_{jk}(x,u)\widehat{G}_{jk}(x,v)\left[\widehat{F}_j(\min(u,v)|x) - \widehat{F}_j(u|x)\widehat{F}_j(v|x)\right] dudv,$$

in which

$$\widehat{G}_{1L}(x,y) = \mu_1'\left(y, \widehat{F}_0^{-1}\left(1 - \widehat{F}_1(y|x)|x\right)\right), \quad \widehat{G}_{0L}(x,y) = \mu_0'\left(\widehat{F}_1^{-1}\left(1 - \widehat{F}_0(y|x)|x\right), y\right),$$

$$\widehat{G}_{1U}(x,y) = \mu_1'\left(y, \widehat{F}_0^{-1}\left(\widehat{F}_1(y|x)|x\right)\right), \quad \widehat{G}_{0U}(x,y) = \mu_0'\left(\widehat{F}_1^{-1}\left(\widehat{F}_0(y|x)|x\right), y\right).$$

THEOREM 3.2 *Suppose the conditions of Theorem 3.1 hold and $0 < \alpha < 1$. Then the nominal level $1-\alpha$ CS defined as $CS_n(x)$ satisfies $\lim_{n\to\infty} \inf_{\theta_L(x)\le\theta_0\le\theta_U(x)} P(\theta_0 \in CS_n(x)) = 1-\alpha$.*

Remark 3.3. Theorem 3.2 is a pointwise result in the true probability measure characterizing the population. It would be interesting to extend the CS and Theorem 3.2 to allow for the possibility that at least one of the $F_j(\cdot|x)$ is degenerate at a finite value. We'll leave this to future work.

4 Estimation and Confidence Sets for θ

Suppose $\mu(\cdot,\cdot)$ is strict quasi-monotone and right continuous. Proposition 2.3 implies: $\theta_L \le \theta \le \theta_U$, where

$$\theta_L = E\left[\int_0^1 \mu\left(F_1^{-1}(u|X), F_0^{-1}(1-u|X)\right) du\right], \ \theta_U = E\left[\int_0^1 \mu\left(F_1^{-1}(u|X), F_0^{-1}(u|X)\right) du\right].$$

Throughout this section, we assume $\mathcal{X} = \mathcal{R}^d$.

To estimate θ_L, θ_U, we use the following leave-one-out kernel estimators of $F_j(y|x)$ for $j = 1, 0$:

$$\widehat{F}_{j,-l}(y|x) = \frac{\sum_{i=1,i\neq l}^n I\{Y_i \le y\} K_{aj}(X_i - x)}{\sum_{i=1,i\neq l}^n K_{aj}(X_i - x)},$$

$$\widehat{f}_{jX,-l}(x) = \frac{1}{(n-1)a_n^d} \sum_{i=1,i\neq l}^n K_{aj}(X_i - x).$$

Let $I_{bi} = I\{\|X_i\| \le b\}$ for a positive sequence $b = b_n \to \infty$, where $\|x\| = \max\{|x_1|, ..., |x_d|\}$ for $x = (x_1, ..., x_d)'$. The lower and upper bounds on θ can be estimated by the plug-in estimators below:

$$\widehat{\theta}_L = \frac{1}{n}\sum_{i=1}^{n} I_{bi}\left[\int_0^1 \mu\left(\widehat{F}_{1,-i}^{-1}(u|X_i), \widehat{F}_{0,-i}^{-1}(1-u|X_i)\right)du\right] = \frac{1}{n}\sum_{i=1}^{n} I_{bi}\widehat{\theta}_L(X_i),$$

$$\widehat{\theta}_U = \frac{1}{n}\sum_{i=1}^{n} I_{bi}\left[\int_0^1 \mu\left(\widehat{F}_{1,-i}^{-1}(u|X_i), \widehat{F}_{0,-i}^{-1}(u|X_i)\right)du\right] = \frac{1}{n}\sum_{i=1}^{n} I_{bi}\widehat{\theta}_U(X_i).$$

To establish the asymptotic properties of $\widehat{\theta}_L$, $\widehat{\theta}_U$, we make the following assumptions.

(A1U) (i) Suppose that X is continuous with density function $f_X(\cdot)$ satisfying $\sup_x f_X(x) < \infty$. (ii) The functions $f_X(x)$ and $p(x)$ are m times differentiable and their m-th derivatives are uniformly continuous and bounded. (iii) The kernel function $K(x) : \mathcal{R}^d \to \mathcal{R}$ is an m-th order kernel for which $\left|K^{(m)}(u)\right|$ is bounded and integrable. (iv) The bandwidth sequence a_n satisfies $a_n \to 0$ and $\sqrt{n}a_n^m \to 0$.

(A2U) Assume (i) for each $y \in \mathcal{Y}_j$, $F_j(y|x)$ is m times differentiable with respect to x and the derivatives of $f_{jX}(x)F_j(y|x)$ are uniformly continuous and bounded. (ii) $f_j(\cdot|x)$ exists, is bounded with bounded second order partial derivatives, and (iii) $\mathcal{Y}_j$ is compact and $f_j(\cdot|\cdot)$ satisfies $\inf_{x,y\in\mathcal{Y}_j} f_j(y \mid x) > 0$.

Proposition 4.1 *Assume (A1U), (A2U) and (A3P) hold. For $q > 0$, let $b_n = O\left((\ln n)^{1/d} n^{1/2q}\right)$, $\delta_n = \min\{\delta_{n1}, \delta_{n0}\}$, and $c_n = \left(\frac{\log n}{na_n^d}\right)^{1/2} + a_n^m$, where $\delta_{nj} = \inf_{\|x\|\le b} f_{jX}(x)$ for $j = 1, 0$. Then*

$$\begin{aligned}
&\left[\widehat{\theta}_L(x) - \theta_L(x)\right] I\{\|x\| \le b\} = \\
&-\frac{1}{f_{1X}(x)(na_n^d)} I\{\|x\| \le b\}\sum_{i=1}^{n} K_{a1}(X_i - x)\int G_{1L}(x,y)\left[I(Y_i \le y) - F_1(y|x)\right]dy \\
&-\frac{1}{f_{0X}(x)(na_n^d)} I\{\|x\| \le b\}\sum_{i=1}^{n} K_{a0}(X_i - x)\int G_{0L}(x,y)\left[I(Y_i \le y) - F_0(y|x)\right]dy \\
&+R_n^L(x),
\end{aligned} \tag{16}$$

and

$$\begin{aligned}
&\left[\widehat{\theta}_U(x) - \theta_U(x)\right] I\{\|x\| \le b\} = \\
&-\frac{1}{f_{1X}(x)(na_n^d)} I\{\|x\| \le b\}\sum_{i=1}^{n} K_{a1}(X_i - x)\int G_{1U}(x,y)\left[I(Y_i \le y) - F_1(y|x)\right]dy \\
&-\frac{1}{f_{0X}(x)(na_n^d)} I\{\|x\| \le b\}\sum_{i=1}^{n} K_{a0}(X_i - x)\int G_{0U}(x,y)\left[I(Y_i \le y) - F_0(y|x)\right]dy \\
&+R_n^U(x),
\end{aligned} \tag{17}$$

where $\sup_{x\in\mathcal{X}}\left|R_n^L(x)\right| = O_p\left(\delta_n^{-2}c_n^2\right)$ *and* $\sup_{x\in\mathcal{X}}\left|R_n^U(x)\right| = O_p\left(\delta_n^{-2}c_n^2\right)$.

The additional conditions in Proposition 4.1 are similar to those in Theorems 6 and 8 in Hansen (2008). Together with (A1U) and (A2U), they ensure

$$\sup_{\|x\|\leq b}\left|\widehat{f}_{jX}(x)-f_{jX}(x)\right|=O_p(c_n), \quad \sup_{\|x\|\leq b, y\in\mathcal{Y}_j}\left|\widehat{F}_j(y|x)-F_j(y|x)\right|=O_p\left(\delta_{nj}^{-1}c_n\right),$$

where $c_n=\left(\frac{\log n}{na_n^d}\right)^{1/2}+a_n^m$. In the Appendix we use Proposition 4.1 to establish the joint asymptotic normality of $\widehat{\theta}_L$ and $\widehat{\theta}_U$. For notational convenience, let $W_i=(X_i,Y_i,D_i)$ and $r_L(W_i)=r_{1L}(W_i)+r_{0L}(W_i)$, where

$$\begin{aligned}
r_{1L}(W_i) &= \frac{I\{D_i=1\}}{p(X_i)}\int G_{1L}(X_i,y)\left[I(Y_i\leq y)-F_1(y|X_i)\right]dy,\\
r_{0L}(W_i) &= \frac{I\{D_i=0\}}{1-p(X_i)}\int G_{0L}(X_i,y)\left[I(Y_i\leq y)-F_0(y|X_i)\right]dy.
\end{aligned}$$

Similarly, we let $r_U(W_i)=r_{1U}(W_i)+r_{0U}(W_i)$, where

$$\begin{aligned}
r_{1U}(W_i) &= \frac{I\{D_i=1\}}{p(X_i)}\int G_{1U}(X_i,y)\left[I(Y_i\leq y)-F_1(y|X_i)\right]dy,\\
r_{0U}(W_i) &= \frac{I\{D_i=0\}}{1-p(X_i)}\int G_{0U}(X_i,y)\left[I(Y_i\leq y)-F_0(y|X_i)\right]dy.
\end{aligned}$$

In the Appendix, we show:

$$\sqrt{n}\left(\widehat{\theta}_L-\theta_L\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[\theta_L(X_i)-r_L(W_i)-E\left[\theta_L(X_i)-r_L(W_i)\right]\right]+o_p(1),$$

$$\sqrt{n}\left(\widehat{\theta}_U-\theta_U\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[\theta_U(X_i)-r_U(W_i)-E\left[\theta_U(X_i)-r_U(W_i)\right]\right]+o_p(1).$$

THEOREM 4.2 *Suppose the assumptions of Proposition 4.1 hold. In addition, if* $\sqrt{n}\delta_n^{-2}c_n^2\to 0$ *and* $\sqrt{n}E(\theta_k(X_i)I_{bi}-\theta_k)\to 0$ *for* $k=L,U$, *then we obtain:*

$$\sqrt{n}\begin{pmatrix}\widehat{\theta}_L-\theta_L\\ \widehat{\theta}_U-\theta_U\end{pmatrix}\Longrightarrow N\left(0,Var\left(\begin{pmatrix}\theta_L(X_i)-r_L(W_i)\\ \theta_U(X_i)-r_U(W_i)\end{pmatrix}\right)\right).$$

The condition: $\sqrt{n}E(\theta_k(X_i)I_{bi}-\theta_k)=o(1)$ for $k=L,U$, ensures that the bias introduced by trimming I_{bi} in the estimator $\widehat{\theta}_k$ for $k=L,U$ approaches zero at a faster rate than $n^{-1/2}$. Similar conditions are used in Khan and Tamer (2009), see also Andrews and Schafgans (1998) and Schafgans and Zinde-Walsh (2002).

The asymptotic variance of $\widehat{\theta}_L$ depends on three terms: $\theta_L(X_i)$, $r_{1L}(W_i)$, and $r_{0L}(W_i)$. The latter two terms are due to the estimation of the conditional marginal distributions $F_1(\cdot|\cdot), F_0(\cdot|\cdot)$ respectively. To make inference on θ, we need to estimate the asymptotic variances of $\widehat{\theta}_L$ and $\widehat{\theta}_U$.

For $k = L, U$, let $\widehat{\sigma}_k^2$ denote the sample variance of $\left\{ \left[\widehat{\theta}_k (X_i) - \widehat{r}_k (W_i) \right] I_{bi} \right\}_{i=1}^n$, where $\widehat{r}_k (W_i) = \widehat{r}_{1k} (W_i) + \widehat{r}_{0k} (W_i)$ in which

$$\begin{aligned} \widehat{r}_{1k} (W_i) &= \frac{I\{D_i = 1\}}{\widehat{p}_{-i} (X_i)} \int \widehat{G}_{1k,-i} (X_i, y) \left[I (Y_i \leq y) - \widehat{F}_{1,-i} (y|X_i) \right] dy, \\ \widehat{r}_{0k} (W_i) &= \frac{I\{D_i = 0\}}{1 - \widehat{p}_{-i} (X_i)} \int \widehat{G}_{0k,-i} (X_i, y) \left[I (Y_i \leq y) - \widehat{F}_{0,-i} (y|X_i) \right] dy, \end{aligned}$$

and

$$\widehat{p}_{-l} (x) = \frac{\sum_{i=1, i \neq l}^n K_{aj} (X_i - x)}{\sum_{i=1, i \neq l}^n K \left(\frac{X_i - x}{a} \right)},$$

$$\widehat{G}_{1L,-i}(x, y) = \mu_1' \left(y, \widehat{F}_{0,-i}^{-1} \left(1 - \widehat{F}_{1,-i} (y|x) |x \right) \right), \quad \widehat{G}_{0L,-i}(x, y) = \mu_0' \left(\widehat{F}_{1,-i}^{-1} \left(1 - \widehat{F}_{0,-i} (y|x) |x \right), y \right),$$

$$\widehat{G}_{1U,-i}(x, y) = \mu_1' \left(y, \widehat{F}_{0,-i}^{-1} \left(\widehat{F}_{1,-i} (y|x) |x \right) \right), \quad \widehat{G}_{0U,-i}(x, y) = \mu_0' \left(\widehat{F}_{1,-i}^{-1} \left(\widehat{F}_{0,-i} (y|x) |x \right), y \right).$$

Let

$$CS_n = \left[\widehat{\theta}_L - \frac{\widehat{\sigma}_L}{\sqrt{n}} z_{1-\alpha}, \widehat{\theta}_U + \frac{\widehat{\sigma}_U}{\sqrt{n}} z_{1-\alpha} \right].$$

THEOREM 4.3 *Suppose the conditions of Theorem 4.2 hold and* $0 < \alpha < 1$*. Then the nominal level* $1 - \alpha$ *CS defined as* CS_n *satisfies* $\lim_{n \to \infty} \inf_{\theta_L \leq \theta_0 \leq \theta_U} P(\theta_0 \in CS_n) = 1 - \alpha$.

5 Conclusion and Extensions

In this paper, we have presented a systematic study of partial identification and inference for a general class of functionals of the joint distribution of potential outcomes of a binary treatment under the selection on observables assumption. Specifically we have established sharp bounds on functionals in this class and identified conditions under which the true functionals are point identified. For both functionals of the conditional joint distribution function and the unconditional joint distribution function of potential outcomes in this class, we have proposed nonparametric estimators of their sharp bounds and established asymptotic properties of these estimators. Based on nonparametric estimators of the sharp bounds, we have constructed asymptotically valid CSs for functionals of both the conditional and unconditional joint distributions of potential outcomes.

Several extensions of the results we have provided in this paper are possible. First, as mentioned in Remark 2.2, we might be interested in the parameter $\theta(x_1) = E\left[\mu(Y_1, Y_0) | X^1 = x_1\right]$, where $X = \left(X^1, X^{-1}\right)$ and $\mu(\cdot, \cdot)$ is quasi-monotone and right continuous. When (C1) and (C2) hold, the sharp bounds on $\theta(x_1)$ are given by $\theta_L(x_1)$ and $\theta_U(x_1)$, where

$$\begin{aligned} \theta_L (x_1) &= E\left[\int_0^1 \mu\left(F_1^{-1}\left(u|x_1, X^{-1}\right), F_0^{-1}\left(1 - u|x_1, X^{-1}\right)\right) du \right], \\ \theta_U (x_1) &= E\left[\int_0^1 \mu\left(F_1^{-1}\left(u|x_1, X^{-1}\right), F_0^{-1}\left(u|x_1, X^{-1}\right)\right) du \right]. \end{aligned}$$

Then $\theta_L(x_1)$ and $\theta_U(x_1)$ can be estimated by

$$\widehat{\theta}_L(x_1) = \frac{1}{n}\sum_{i=1}^{n} I\left\{\left\|X_i^{-1}\right\| \leq b\right\} \int_0^1 \mu\left(\widehat{F}_{1,-i}^{-1}\left(u|x_1, X_i^{-1}\right), \widehat{F}_{0,-i}^{-1}\left(1-u|x_1, X_i^{-1}\right)\right) du,$$

$$\widehat{\theta}_U(x_1) = \frac{1}{n}\sum_{i=1}^{n} I\left\{\left\|X_i^{-1}\right\| \leq b\right\} \int_0^1 \mu\left(\widehat{F}_{1,-i}^{-1}\left(u|x_1, X_i^{-1}\right), \widehat{F}_{0,-i}^{-1}\left(u|x_1, X_i^{-1}\right)\right) du.$$

The asymptotic properties of $\widehat{\theta}_L(x_1), \widehat{\theta}_U(x_1)$ can be established similarly to those of $\widehat{\theta}_L, \widehat{\theta}_U$. Instead of converging at $n^{-1/2}$ rate, they converge at $(na^r)^{-1/2}$ rate, where r is the dimension of X^1. This is similar to the partial means studied in Newey (1994). CSs for $\theta(x_1)$ can be constructed similarly to CSs for θ.

Second, as discussed in Example 2.2, many inequality measures of the distribution of treatment effects Δ can be expressed as $g_\gamma(E[\nu_\gamma(\Delta)], \mu_\Delta)$, where

$$g_\gamma(z, \mu_\Delta) = \frac{1}{\gamma^2-\gamma}\left[\frac{z}{\mu_\Delta^\gamma} - 1\right],$$

and $\nu_\gamma(\Delta) = \Delta^\gamma$ for a given real-valued parameter γ. Suppose γ is even. Then $\nu(\cdot)$ is continuous and convex. In addition, $g(\cdot,\cdot)$ is increasing in its first argument. Let $\mu_\Delta(x) = E(\Delta|X=x)$. Given that $\mu_\Delta(x)$ and μ_Δ are point identified, we have:

$$g_\gamma(\theta_{L\gamma}(x), \mu_\Delta(x)) \leq g_\gamma(E[\nu_\gamma(\Delta)|x], \mu_\Delta(x)) \leq g_\gamma(\theta_{U\gamma}(x), \mu_\Delta(x)),$$

and

$$g_\gamma(\theta_{L\gamma}, \mu_\Delta) \leq g_\gamma(E[\nu_\gamma(\Delta)], \mu_\Delta) \leq g_\gamma(\theta_{U\gamma}, \mu_\Delta),$$

where

$$\begin{aligned}
\theta_{L\gamma}(x) &= \int_0^1 \nu_\gamma\left(F_1^{-1}(u|x) - F_0^{-1}(u|x)\right) du, \\
\theta_{U\gamma}(x) &= \int_0^1 \nu_\gamma\left(F_1^{-1}(u|x) - F_0^{-1}(1-u|x)\right) du, \\
\theta_{L\gamma} &= E\left[\int_0^1 \nu_\gamma\left(F_1^{-1}(u|X) - F_0^{-1}(u|X)\right) du\right], \\
\theta_{U\gamma} &= E\left[\int_0^1 \nu_\gamma\left(F_1^{-1}(u|X) - F_0^{-1}(1-u|X)\right) du\right].
\end{aligned}$$

Let $\widehat{\mu}_\Delta(x)$ and $\widehat{\mu}_\Delta$ denote consistent estimators of $\mu_\Delta(x)$ and μ_Δ respectively and

$$\begin{aligned}
\widehat{\theta}_{L\gamma}(x) &= \int_0^1 \nu_\gamma\left(\widehat{F}_1^{-1}(u|x) - \widehat{F}_0^{-1}(u|x)\right) du, \\
\widehat{\theta}_{U\gamma}(x) &= \int_0^1 \nu_\gamma\left(\widehat{F}_1^{-1}(u|x) - \widehat{F}_0^{-1}(1-u|x)\right) du, \\
\widehat{\theta}_{L\gamma} &= \frac{1}{n}\sum_{i=1}^{n} I_{bi} \int_0^1 \nu_\gamma\left(\widehat{F}_{1,-i}^{-1}(u|X_i) - \widehat{F}_{0,-i}^{-1}(u|X_i)\right) du, \\
\widehat{\theta}_{U\gamma} &= \frac{1}{n}\sum_{i=1}^{n} I_{bi} \int_0^1 \nu_\gamma\left(\widehat{F}_{1,-i}^{-1}(u|X_i) - \widehat{F}_{0,-i}^{-1}(1-u|X_i)\right) du.
\end{aligned}$$

Then consistent estimators of sharp bounds on $g_\gamma\left(E\left[\nu_\gamma\left(\Delta\right)|x\right], \mu_\Delta\left(x\right)\right)$ and $g_\gamma\left(E\left[\nu_\gamma\left(\Delta\right)\right], \mu_\Delta\right)$ can be constructed by the plug-in approach. To make inferences on inequality measures $g_\gamma\left(E\left[\nu_\gamma\left(\Delta\right)|x\right], \mu_\Delta\left(x\right)\right)$ and $g_\gamma\left(E\left[\nu_\gamma\left(\Delta\right)\right], \mu_\Delta\right)$, we need to establish the joint asymptotic distributions of $\widehat{\theta}_{L\gamma}\left(x\right), \widehat{\theta}_{U\gamma}\left(x\right)$, $\widehat{\mu}_\Delta\left(x\right)$ and of $\widehat{\theta}_{L\gamma}, \widehat{\theta}_{U\gamma}$, $\widehat{\mu}_\Delta$. The results for $\widehat{\theta}_L\left(x\right)$, $\widehat{\theta}_U\left(x\right)$, $\widehat{\theta}_L$, $\widehat{\theta}_U$ established in Sections 3 and 4 and the following assumption on $\widehat{\mu}_\Delta\left(x\right)$ and $\widehat{\mu}_\Delta$ suffice.

Assumption (ATE). (i) There exists $0 < \delta \leq 1/2$ and $\psi_x\left(W_i\right)$ with $E\left[\ \psi_x\left(W_i\right)\right] = 0$ such that

$$\widehat{\mu}_\Delta\left(x\right) - \mu_\Delta\left(x\right) = \frac{1}{n^\delta}\sum_{i=1}^n \psi_x\left(W_i\right) + o_p\left(n^{-\delta/2}\right),$$

(ii) There exists $\psi\left(W_i\right)$ with $E\left[\psi\left(W_i\right)\right] = 0$ such that

$$\widehat{\mu}_\Delta - \mu_\Delta = \frac{1}{n}\sum_{i=1}^n \psi\left(W_i\right) + o_p\left(n^{-1/2}\right).$$

Assumption (ATE) is known to hold for most existing estimators of average treatment effects, see e.g., Hahn (1997), Heckman, Ichimura, Smith, and Todd (1998a, b), and Hirano, Imbens, and Ridder (2003). The asymptotic distributions of the plug-in estimators of sharp bounds on $g_\gamma\left(E\left[\nu_\gamma\left(\Delta\right)|x\right], \mu_\Delta\left(x\right)\right)$ and $g_\gamma\left(E\left[\nu_\gamma\left(\Delta\right)\right], \mu_\Delta\right)$ can then be established by using the Delta method and CSs for $g_\gamma\left(E\left[\nu_\gamma\left(\Delta\right)|x\right], \mu_\Delta\left(x\right)\right)$ and $g_\gamma\left(E\left[\nu_\gamma\left(\Delta\right)\right], \mu_\Delta\right)$ can be constructed similarly to CSs for $\theta\left(x\right)$ and θ.

Finally, we point out that the sharp bounds on θ in Proposition 2.3 employ the covariate information only. If in addition to the covariate information, other information such as dependence information on the potential outcomes is available, it may be used to further tighten the bounds in Proposition 2.3.

Appendix: Technical Proofs

Proof of Proposition 2.2: (i) The result follows directly from additive separability of $\mu(\cdot,\cdot)$ and a change of variables. **(ii)** We prove this for strict quasi-monotone functions only. Noting that

$$\begin{aligned}\theta^U &= \int_0^{1/2} \mu\left(F_1^{-1}(u), F_0^{-1}(u)\right) du + \int_{1/2}^{1} \mu\left(F_1^{-1}(u), F_0^{-1}(u)\right) du \\ &= \int_0^{1/2} \mu\left(F_1^{-1}(u), F_0^{-1}(u)\right) du + \int_0^{1/2} \mu\left(F_1^{-1}(1-u), F_0^{-1}(1-u)\right) du\end{aligned}$$

and

$$\theta^L = \int_0^{1/2} \mu\left(F_1^{-1}(u), F_0^{-1}(1-u)\right) du + \int_0^{1/2} \mu\left(F_1^{-1}(1-u), F_0^{-1}(u)\right) du,$$

we obtain:

$$\theta^U - \theta^L = \int_0^{1/2} \left[\begin{array}{c} \mu\left(F_1^{-1}(u), F_0^{-1}(u)\right) + \mu\left(F_1^{-1}(1-u), F_0^{-1}(1-u)\right) \\ -\mu\left(F_1^{-1}(u), F_0^{-1}(1-u)\right) - \mu\left(F_1^{-1}(1-u), F_0^{-1}(u)\right) \end{array}\right] du. \tag{A.1}$$

Since $\mu(\cdot,\cdot)$ is a quasi-monotone function, we get:

$$\begin{aligned}&\mu\left(F_1^{-1}(u), F_0^{-1}(u)\right) + \mu\left(F_1^{-1}(1-u), F_0^{-1}(1-u)\right) \\ &-\mu\left(F_1^{-1}(u), F_0^{-1}(1-u)\right) - \mu\left(F_1^{-1}(1-u), F_0^{-1}(u)\right) \geq 0, \ \forall u \in [0,1]\,.\end{aligned} \tag{A.2}$$

It then follows from (A.1) and (A.2) that $\theta^L = \theta^U$ if and only if for almost all $u \in [0, 1/2]$,

$$\begin{aligned}&\mu\left(F_1^{-1}(u), F_0^{-1}(u)\right) + \mu\left(F_1^{-1}(1-u), F_0^{-1}(1-u)\right) \\ &-\mu\left(F_1^{-1}(u), F_0^{-1}(1-u)\right) - \mu\left(F_1^{-1}(1-u), F_0^{-1}(u)\right) = 0\,.\end{aligned} \tag{A.3}$$

It is obvious that (A.3) holds when one of $F_1(\cdot)$ and $F_0(\cdot)$ is degenerate.

Now we show that if $\theta^U = \theta^L$, then at least one of $F_1(\cdot)$ and $F_0(\cdot)$ is degenerate. We will prove this by contradiction. Suppose that both $F_1(\cdot)$ and $F_0(\cdot)$ are non-degenerate. Then there are y_j, y_j' $(j = 1, 0)$ satisfying: $y_j < y_j'$ and $0 < F_j(y_j) \leq F_j\left(y_j'\right) < 1$. Define

$$u_* \equiv \min\left\{F_1(y_1), 1 - F_1\left(y_1'\right), F_0(y_0), 1 - F_0\left(y_0'\right)\right\}.$$

Then, $0 < u_* \leq 1/2$ and for all $u \in [0, u_*)$, we have

$$F_j^{-1}(u) \leq y_j < y_j' < F_j^{-1}(1-u)\,.$$

It follows from the "strict quasi-monotone" assumption that

$$\begin{aligned}&\mu\left(F_1^{-1}(u), F_0^{-1}(u)\right) + \mu\left(F_1^{-1}(1-u), F_0^{-1}(1-u)\right) \\ &-\mu\left(F_1^{-1}(u), F_0^{-1}(1-u)\right) - \mu\left(F_1^{-1}(1-u), F_0^{-1}(u)\right) > 0, \ \forall u \in [0, u_*)\,.\end{aligned}$$

This contradicts with (A.3), a sufficient and necessary condition for $\theta^U - \theta^L = 0$ to hold. ■

Proof of Proposition 2.3: (i) For $x \in \mathcal{X}$, let $\theta(x) = E[\mu(Y_1, Y_0)|X = x]$,

$$\theta_L(x) = \int_0^1 \mu\left(F_1^{-1}(u|x), F_0^{-1}(1-u|x)\right) du,$$

and

$$\theta_U(x) = \int_0^1 \mu\left(F_1^{-1}(u|x), F_0^{-1}(u|x)\right) du.$$

Then Lemma 2.1 (i) implies: $\theta_L(x) \le \theta(x) \le \theta_U(x)$ for all $x \in \mathcal{X}$. Taking expectations with respect to X leads to the bounds in (i). These bounds are sharp, as they are achieved at $C^L(F_1(\cdot|x), F_0(\cdot|x)), C^U(F_1(\cdot|x), F_0(\cdot|x))$ respectively.

(ii) Define

$$\begin{aligned}\Delta(u|x) \equiv\ & \mu\left(F_1^{-1}(u|x), F_0^{-1}(u|x)\right) + \mu\left(F_1^{-1}(1-u|x), F_0^{-1}(1-u|x)\right) \\ & -\mu\left(F_1^{-1}(u|x), F_0^{-1}(1-u|x)\right) - \mu\left(F_1^{-1}(1-u|x), F_0^{-1}(u|x)\right).\end{aligned}$$

Similar to (A.2) and (A.1), we have $\Delta(u|x) \ge 0$ for all u and x, and

$$\theta_U - \theta_L = E\left[\int_0^{1/2} \Delta(u|X)\, du\right] = \int \left(\int_0^{1/2} \Delta(u|x)\, du\right) dF_X(x) \ge 0,$$

where $F_X(\cdot)$ is the distribution function of X. Obviously, $\theta_U = \theta_L$ *if and only if* $\Delta(u|X) = 0$ with probability one for almost all $u \in [0, 1/2]$. When one of $F_1(\cdot|x)$ and $F_0(\cdot|x)$ is degenerate for every $x \in \mathcal{X}$, we have $\Delta(u|x) = 0$ for almost all $u \in [0, 1/2]$, implying $\theta_U = \theta_L$. Now we show that under the "strict quasi-monotone" assumption, if $\theta_U = \theta_L$, then one of $F_1(\cdot|x)$ and $F_0(\cdot|x)$ is degenerate for almost all $x \in \mathcal{X}$. By contradiction, assuming that there is a set $A \subset \mathcal{X}$ such that $\Pr(A) > 0$ and for every $x \in A$ both $F_1(\cdot|x)$ and $F_0(\cdot|x)$ are non-degenerate, then by a similar proof to that of Proposition 2.2 (ii), we have $\int_0^{1/2} \Delta(u|x)\, du > 0$ for every $x \in A$, implying $\theta_U > \theta_L$, which is a contradiction. ∎

Proof of Proposition 2.4: Let $U \sim U[0, 1]$, independent of X. Then

$$\begin{aligned}\theta_L &= E\left[\mu\left(F_1^{-1}(U|X), F_0^{-1}(1-U|X)\right)\right], \\ \theta_{LP} &= E\left[\mu\left(F_1^{-1}(U|p(X)), F_0^{-1}(1-U|p(X))\right)\right].\end{aligned}$$

Define the following random variables:

$$\begin{aligned}V_1 &= F_1^{-1}(U|X), \quad V_0 = F_0^{-1}(1-U|X), \\ V_{1P} &= F_1^{-1}(U|p(X)), \quad V_{0P} = F_0^{-1}(1-U|p(X)).\end{aligned}$$

It is easy to show that

$$V_1 \sim F_1, \quad V_0 \sim F_0, \quad V_{1P} \sim F_1, \quad V_{0P} \sim F_0.$$

It follows from Theorem 3.1.2 (a) in Rachev and Ruschendorf (1998) that:

$$\theta_L - \theta_{LP} = \int \int \left[G\left(y_1, y_0\right) - G_P\left(y_1, y_0\right) \right] \mu_c\left(dy_1, dy_0\right),$$

where G, G_P are the joint distribution functions of (V_1, V_0) and (V_{1P}, V_{0P}) respectively, and μ_c is a nonnegative measure generated by μ. Noting that

$$\begin{aligned} G\left(y_1, y_0\right) &= E\left[\Pr\left(V_1 \leq y_1, V_0 \leq y_0 | X\right)\right] = E\left[C^L\left(F_1\left(y_1|X\right), F_0\left(y_0|X\right)\right)\right], \\ G_P\left(y_1, y_0\right) &= E\left[\Pr\left(V_{1P} \leq y_1, V_{0P} \leq y_0 | p\left(X\right)\right)\right] = E\left[C^L\left(F_1\left(y_1|p\left(X\right)\right), F_0\left(y_0|p\left(X\right)\right)\right)\right], \end{aligned}$$

we get

$$\begin{aligned} G_P\left(y_1, y_0\right) &= E\left[\max\left\{F_1\left(y_1|p\left(X\right)\right) + F_0\left(y_0|p(X)\right) - 1, 0\right\}\right] \\ &= E\left[\max\left\{E\left[F_1\left(y_1|X\right) + F_0\left(y_0|X\right) - 1|p\left(X\right)\right], 0\right\}\right] \\ &\leq E\left[E\left[\max\left\{F_1\left(y_1|X\right) + F_0\left(y_0|X\right) - 1, 0\right\}|p\left(X\right)\right]\right] \\ &= E\left[\max\left\{F_1\left(y_1|X\right) + F_0\left(y_0|X\right) - 1, 0\right\}\right] \\ &= G\left(y_1, y_0\right), \end{aligned}$$

where the inequality is obtained by Jensen's inequality, as $\max\{x, 0\}$ is a convex function of x. Thus, $\theta_L - \theta_{LP} \geq 0$. Similarly, by using the fact that $\min\{x, 0\}$ is a concave function of x, we obtain:

$$E\left[C^U\left(F_1\left(y_1|p\left(X\right)\right), F_0\left(y_0|p(X)\right)\right)\right] \geq E\left[C^U\left(F_1\left(y_1|X\right), F_0\left(y_0|X\right)\right)\right].$$

■

Proof of Theorem 3.1. We will complete the proof in three steps. **Step 1.** We show that the random function

$$(y_1, y_0) \mapsto \sqrt{na_n^d} \begin{pmatrix} \widehat{F}_1(y_1|x) - F_1(y_1|x) \\ \widehat{F}_0(y_0|x) - F_0(y_0|x) \end{pmatrix}$$

converges weakly to a bivariate Gaussian process $\mathcal{G} = (\mathcal{G}_1, \mathcal{G}_0)'$ with mean zero and covariance function

$$E\left(\mathcal{G}\left(y_1, y_0\right) \mathcal{G}^T\left(y_1', y_0'\right)\right) = E\begin{pmatrix} \mathcal{G}_1\left(y_1\right) \mathcal{G}_1^T\left(y_1'\right) & \mathcal{G}_0\left(y_0\right) \mathcal{G}_1^T\left(y_1'\right) \\ \mathcal{G}_1\left(y_1\right) \mathcal{G}_0^T\left(y_0'\right) & \mathcal{G}_0\left(y_0\right) \mathcal{G}_0^T\left(y_0'\right) \end{pmatrix},$$

where $E\left(\mathcal{G}_0\left(y_0\right) \mathcal{G}_1^T\left(y_1'\right)\right) = E\left(\mathcal{G}_1\left(y_1\right) \mathcal{G}_0^T\left(y_0'\right)\right) = 0$ and for $j = 1, 0$,

$$E\left(\mathcal{G}_j\left(y_j\right) \mathcal{G}_j^T\left(y_j'\right)\right) = \frac{1}{f_{jX}\left(x\right)}\left[\int K^2\left(t\right) dt\right]\left\{F_j\left(\min\left\{y_j, y_j'\right\}|x\right) - F_j\left(y_j|x\right) F_j\left(y_j'|x\right)\right\},$$

and the convergence is in $D\left(\mathcal{Y}_1\right) \times D\left(\mathcal{Y}_0\right)$.

We accomplish this by using similar arguments to those used in the proofs of Theorems 2.2 and 2.3 in Polonik and Yao (2002). First, note that under the conditions of Theorem 3.1, we have:

$$\frac{1}{na^d}\sum_{i=1}^{n} I\left\{D_i = j\right\} K\left(\frac{X_i - x}{a}\right) - p_j\left(x\right) f_X\left(x\right) = O_p\left(\left(na^d\right)^{1/2}\right).$$

Thus, we get

$$\begin{aligned}
&\widehat{F}_j(y_j|x) - F_j(y_j|x)\\
=\ & \frac{\left(na^d\right)^{-1}\sum_{i=1}^{n} I\left\{D_i = j\right\} K\left(\frac{X_i - x}{a}\right)\left[I\left\{Y_i \le y_j\right\} - F_j(y_j|x)\right]}{\left(na^d\right)^{-1}\sum_{i=1}^{n} I\left\{D_i = j\right\} K\left(\frac{X_i - x}{a}\right)}\\
=\ & \frac{\left(na^d\right)^{-1}\sum_{i=1}^{n} I\left\{D_i = j\right\} K\left(\frac{X_i - x}{a}\right)\left[I\left\{Y_i \le y_j\right\} - F_j(y_j|x)\right]}{p_j\left(x\right) f_X\left(x\right)} + o_p\left(\left(na^d\right)^{1/2}\right). \qquad \text{(A.4)}
\end{aligned}$$

The asymptotic normality of the finite dimensional distribution of the process:

$$(y_1, y_0) \mapsto \sqrt{na_n^d}\begin{pmatrix} \widehat{F}_1(y_1|x) - F_1(y_1|x) \\ \widehat{F}_0(y_0|x) - F_0(y_0|x) \end{pmatrix}$$

follows from (A.4) and tightness of the above process follows from the same argument as in the proof of Theorem 2.3 in Polonik and Yao (2002). It remains to evaluate the mean and variance of the Gaussian process. For $j = 1, 0$, let

$$\varepsilon_{j,i}\left(y\right) = \frac{1}{f_{jX}\left(x\right)\sqrt{na_n^d}} K_{aj}\left(X_i - x\right)\left[I\left(Y_i \le y\right) - F_j\left(y|x\right)\right]. \qquad \text{(A.5)}$$

Then

$$\sqrt{na_n^d}\begin{pmatrix} \widehat{F}_1(y_1|x) - F_1(y_1|x) \\ \widehat{F}_0(y_0|x) - F_0(y_0|x) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \varepsilon_{1,i}\left(y_1\right) \\ \sum_{i=1}^{n} \varepsilon_{0,i}\left(y_0\right) \end{pmatrix}.$$

First we show that $E\left(\sum_{i=1}^{n} \varepsilon_{j,i}\left(y\right)\right) = o\left(1\right)$ for $j = 1, 0$. In fact,

$$\begin{aligned}
&\sqrt{na_n^d}E(\varepsilon_{j,i}\left(y\right))\\
=\ & \frac{1}{f_{jX}\left(x\right)} E\left[K_{aj}\left(X_i - x\right)\left[I\left(Y_i \le y\right) - F_j\left(y|x\right)\right]\right]\\
=\ & \frac{1}{f_{jX}\left(x\right)} E\left[E\left\{K_{aj}\left(X_i - x\right)\left[I\left(Y_i \le y\right) - F_j\left(y|x\right)\right]|X_i, D_i\right\}\right]\\
=\ & \frac{1}{f_{jX}\left(x\right)} E\left[K_{aj}\left(X_i - x\right)\left[E\left\{I\left(Y_i \le y\right) \mid X_i, D_i\right\} - F_j\left(y|x\right)\right]\right]\\
=\ & \frac{1}{f_{jX}\left(x\right)} E\left[K_{aj}\left(X_i - x\right)\left[F_{D_i}\left(y|X_i\right) - F_j\left(y|x\right)\right]\right]\\
=\ & \frac{1}{f_{jX}\left(x\right)} \int K\left(\frac{z - x}{a_n}\right)\left[F_j\left(y|z\right) - F_j\left(y|x\right)\right] f_{jX}\left(z\right) dz.
\end{aligned}$$

Since x is an interior point of $\mathcal{X}$, by a change of variables and Taylor series expansions, we get

$$\begin{aligned} F_j(y|x+a_n u) - F_j(y|x) &= a_n \nabla_x F_j(y|x)' u + \frac{1}{2} a_n^2 u' \nabla_{xx}^2 F_j(y|\xi_{xu}) u, && \text{(A.6)} \\ f_{jX}(x+a_n u) &= f_{jX}(x) + a_n \nabla_x f_{jX}(\eta_{xu})' u, && \text{(A.7)} \end{aligned}$$

where ξ_{xu} and η_{xu} are two points on the line segment from x to $x+a_n u$. Therefore, we have

$$\begin{aligned} &\sqrt{n a_n^d} E(\varepsilon_{j,i}) \\ = &\frac{a_n^d}{f_{jX}(x)} \int K(u) \left[F_j(y|x+a_n u) - F_j(y|x)\right] f_{jX}(x+a_n u)\, du \\ = &\frac{a_n^{d+1}}{f_{jX}(x)} \int K(u) \left(\begin{array}{c} \left[\nabla_x F_j(y|x)' u + \frac{1}{2} a_n u' \nabla_{xx}^2 F_j(y|\xi_{xu}) u\right] \\ \cdot \left[f_{jX}(x) + a_n \nabla_x f_{jX}(\eta_{xu})' u\right] \end{array} \right) du \\ = &\frac{a_n^{d+2}}{f_{jX}(x)} \nabla_x F_j(y|x)' \left(\int K(u) u \nabla_x f_{jX}(\eta_{xu})' u du \right) \\ &+ \frac{a_n^{d+2}}{2} \int K(u) u' \nabla_{xx}^2 F_j(y|\xi_{xu}) u du + O\left(a_n^{d+3}\right), \end{aligned}$$

implying that $nE(\varepsilon_{j,i}) = O\left(\left(na^{d+4}\right)\right) = o(1)$.

Next, we compute

$$\begin{aligned} &Cov\left(\left(\begin{array}{c} \sum_{i=1}^n \varepsilon_{1,i}(y_1) \\ \sum_{i=1}^n \varepsilon_{0,i}(y_0) \end{array} \right), \left(\begin{array}{c} \sum_{i=1}^n \varepsilon_{1,i}(y_1') \\ \sum_{i=1}^n \varepsilon_{0,i}(y_0') \end{array} \right) \right) \\ = &\left(\begin{array}{ll} nE\left(\varepsilon_{1,i}(y_1) \varepsilon_{1,i}(y_1')\right) & 0 \\ 0 & nE\left(\varepsilon_{0,i}(y_0) \varepsilon_{0,i}(y_0')\right) \end{array} \right), \end{aligned}$$

where we have used the fact that $E(\varepsilon_{0,i}(y_0)\varepsilon_{1,i}(y_1)) = 0$. By a straightforward calculation, we have

$$\begin{aligned} &n a_n^d f_{jX}^2(x) E\left[\varepsilon_{j,i}(y) \varepsilon_{j,i}(y')\right] \\ = &E\left[K_{aj}^2(X_i - x)\left[I(Y_i \le y) - F_j(y|x)\right]\left[I(Y_i \le y') - F_j(y'|x)\right]\right] \\ = &E\left[K_{aj}^2(X_i - x) \left\{ \begin{array}{c} I(Y_i \le \min\{y, y'\}) - I(Y_i \le y) F_j(y'|x) \\ -I(Y_i \le y') F_j(y|x) + F_j(y|x) F_j(y'|x) \end{array} \right\}\right] \\ = &E\left[K_{aj}^2(X_i - x) E\left\{ \left\{ \begin{array}{c} I(Y_i \le \min\{y, y'\}) - I(Y_i \le y) F_j(y'|x) \\ -I(Y_i \le y') F_j(y|x) + F_j(y|x) F_j(y'|x) \end{array} \right\} \mid X_i, D_i \right\}\right] \\ = &E\left[K_{aj}^2(X_i - x) \left\{ \begin{array}{c} F_{D_i}(\min\{y, y'\} | X_i) - F_{D_i}(y|X_i) F_j(y'|x) \\ -F_{D_i}(y'|X_i) F_j(y|x) + F_j(y|x) F_j(y'|x) \end{array} \right\}\right] \\ = &\int K^2\left(\frac{z-x}{a_n}\right) \left\{ \begin{array}{c} F_j(\min\{y, y'\} | z) - F_j(y|z) F_j(y'|x) \\ -F_j(y'|z) F_j(y|x) + F_j(y|x) F_j(y'|x) \end{array} \right\} f_{jX}(z)\, dz. \end{aligned}$$

Similarly, by a change of variables and Taylor series expansions in (A.6) and (A.7), we get

$$\begin{aligned} &n a_n^d f_{jX}^2(x) E\left[\varepsilon_{j,i}(y) \varepsilon_{j,i}(y')\right] \\ = &a_n^d \int K^2(t) \left\{ \begin{array}{c} F_j(\min\{y, y'\} | x + a_n t) - F_j(y|x + a_n t) F_j(y'|x) \\ -F_j(y'|x + a_n t) F_j(y|x) + F_j(y|x) F_j(y'|x) \end{array} \right\} f_{jX}(x + a_n t)\, dt \\ = &a_n^d f_{jX}(x) \left[\int K^2(t)\, dt\right] \left\{F_j\left(\min\{y, y'\} | x\right) - F_j(y|x) F_j\left(y'|x\right)\right\} + O\left(a_n^{d+1}\right), \end{aligned}$$

implying that

$$nE\left[\varepsilon_{j,i}\left(y\right)\varepsilon_{j,i}\left(y'\right)\right]=\frac{1}{f_{jX}\left(x\right)}\left[\int K^{2}\left(t\right)dt\right]\left\{F_{j}\left(\min\left\{y,y'\right\}|x\right)-F_{j}\left(y|x\right)F_{j}\left(y'|x\right)\right\}+O\left(a_{n}\right).$$

Step 2. We show that the map $\phi_{F_1,F_0}:D\left(\mathcal{Y}_1\right)\times D\left(\mathcal{Y}_0\right)\to\mathcal{R}^2$ defined as

$$\phi_{F_1,F_0}=\begin{pmatrix}\int_0^1\mu\left(F_1^{-1}\left(u|x\right),F_0^{-1}\left(1-u|x\right)\right)du\\ \int_0^1\mu\left(F_1^{-1}\left(u|x\right),F_0^{-1}\left(u|x\right)\right)du\end{pmatrix}$$

is Hadamard differentiable at (F_1,F_0) tangentially to $C\left(\mathcal{Y}_1\right)\times C\left(\mathcal{Y}_0\right)$ with the derivative

$$\begin{aligned}
&\phi'_{F_1,F_0}\left(h_1,h_0\right)\\
\mapsto\ &-\begin{pmatrix}\int_0^1\begin{bmatrix}\mu'_1\left(F_1^{-1}\left(u|x\right),F_0^{-1}\left(1-u|x\right)\right)\frac{h_1\left(F_1^{-1}(u|x)\right)}{f_1\left(F_1^{-1}(u|x)|x\right)}\\+\mu'_0\left(F_1^{-1}\left(u|x\right),F_0^{-1}\left(1-u|x\right)\right)\frac{h_0\left(F_0^{-1}(1-u|x)\right)}{f_0\left(F_0^{-1}(1-u|x)|x\right)}\end{bmatrix}du\\
\int_0^1\begin{bmatrix}\mu'_1\left(F_1^{-1}\left(u|x\right),F_0^{-1}\left(u|x\right)\right)\frac{h_1\left(F_1^{-1}(u|x)\right)}{f_1\left(F_1^{-1}(u|x)|x\right)}\\+\mu'_0\left(F_1^{-1}\left(u|x\right),F_0^{-1}\left(u|x\right)\right)\frac{h_0\left(F_0^{-1}(u|x)\right)}{f_0\left(F_0^{-1}(u|x)|x\right)}\end{bmatrix}du\end{pmatrix}\\
=\ &-\begin{pmatrix}\int\mu'_1\left(y,F_0^{-1}\left(1-F_1\left(y|x\right)|x\right)\right)h_1\left(y\right)dy+\int\mu'_0\left(F_1^{-1}\left(1-F_0\left(y|x\right)|x\right),y\right)h_0\left(y\right)dy\\ \int\mu'_1\left(y,F_0^{-1}\left(F_1\left(y|x\right)|x\right)\right)h_1\left(y\right)dy+\int\mu'_0\left(F_1^{-1}\left(F_0\left(y|x\right)|x\right),y\right)h_0\left(y\right)dy\end{pmatrix}\\
=\ &-\begin{pmatrix}\int G_{1L}\left(x,y\right)h_1\left(y\right)dy+\int G_{0L}\left(x,y\right)h_0\left(y\right)dy\\ \int G_{1U}\left(x,y\right)h_1\left(y\right)dy+\int G_{0U}\left(x,y\right)h_0\left(y\right)dy\end{pmatrix}.
\end{aligned}$$

This follows from Chain rule applied to

$$\phi_{F_1,F_0}=\begin{pmatrix}\varphi^L_{Q_1,Q_0}\circ\left(Q_1,Q_0\right)\\ \varphi^U_{Q_1,Q_0}\circ\left(Q_1,Q_0\right)\end{pmatrix},$$

where

$$\begin{aligned}
\varphi^L_{Q_1,Q_0}&=\int_0^1\mu\left(Q_1\left(u|x\right),Q_0\left(1-u|x\right)\right)du,\\
\varphi^U_{Q_1,Q_0}&=\int_0^1\mu\left(Q_1\left(u|x\right),Q_0\left(u|x\right)\right)du,
\end{aligned}$$

and

$$\left(Q_1\left(u_1\right),Q_0\left(u_0\right)\right)=\left(F_1^{-1}\left(u_1|x\right),F_0^{-1}\left(u_0|x\right)\right).$$

(A2P)(iii) ensures that the map: $(Q_1,Q_0):D\left(\mathcal{Y}_1\right)\times D\left(\mathcal{Y}_0\right)\to l^\infty\left((0,1)\right)$ is Hadamard differentiable at (F_1,F_0) tangentially to $C\left(\mathcal{Y}_1\right)\times C\left(\mathcal{Y}_0\right)$. (A3P) ensures that the map $\psi_{Q_1,Q_0}:C\left((0,1)\right)\times C\left((0,1)\right)\to\mathcal{R}^2$ defined as

$$\psi_{Q_1,Q_0}=\begin{pmatrix}\int_0^1\mu\left(Q_1\left(u|x\right),Q_0\left(1-u|x\right)\right)du\\ \int_0^1\mu\left(Q_1\left(u|x\right),Q_0\left(u|x\right)\right)du\end{pmatrix}$$

is Hadamard differentiable at (Q_1,Q_0). By Chain rule, ϕ_{F_1,F_0} is Hadamard differentiable at (F_1,F_0) tangentially to $C\left(\mathcal{Y}_1\right)\times C\left(\mathcal{Y}_0\right)$.

Step 3. By the Delta method, we obtain

$$\sqrt{na_n^d}\begin{pmatrix} \widehat{\theta}_L(x)-\theta_L(x) \\ \widehat{\theta}_U(x)-\theta_U(x) \end{pmatrix} \Longrightarrow \phi'_{F_1,F_0}(\mathcal{G}_1,\mathcal{G}_0).$$

We now show that $\phi'_{F_1,F_0}(\mathcal{G}_1,\mathcal{G}_0)$ has the distribution stated in Theorem 3.1. First, $E\left[\phi'_{F_1,F_0}(\mathcal{G}_1,\mathcal{G}_0)\right]=0$. Now,

$$Var\left[\phi'_{F_1,F_0}(\mathcal{G}_1,\mathcal{G}_0)\right]=Var\left[\begin{pmatrix} \int G_{1L}(x,y)\mathcal{G}_1(y)\,dy+\int G_{0L}(x,y)\mathcal{G}_0(y)\,dy \\ \int G_{1U}(x,y)\mathcal{G}_1(y)\,dy+\int G_{0U}(x,y)\mathcal{G}_0(y)\,dy \end{pmatrix}\right],$$

where for $k=L,U$, it is easy to see that

$$\begin{aligned} & E\left(\left\{\int G_{1k}(x,y)\mathcal{G}_1(y)\,dy+\int G_{0k}(x,y)\mathcal{G}_0(y)\,dy\right\}^2\right) \\ = & \int\int G_{1k}(x,y)G_{1k}(x,y')E\left[\mathcal{G}_1(y)\mathcal{G}_1(y')\right]dydy' \\ & +\int\int G_{0k}(x,y)G_{0k}(x,y')E\left[\mathcal{G}_0(y)\mathcal{G}_0(y')\right]dydy' \\ = & \sigma_{1k}^2(x)+\sigma_{0k}^2(x), \end{aligned}$$

and

$$\begin{aligned} & E\left(\left\{\int\left\{G_{1L}(x,y)\mathcal{G}_1(y)+G_{0L}(x,y)\mathcal{G}_0(y)\right\}dy\right\}\left\{\int\left\{G_{1U}(x,y)\mathcal{G}_1(y)+G_{0U}(x,y)\mathcal{G}_0(y)\right\}dy\right\}\right) \\ = & \int\int G_{1L}(x,y)G_{1U}(x,y)E\left[\mathcal{G}_1(y)\mathcal{G}_1(y')\right]dydy' \\ & +\int\int G_{0L}(x,y)G_{0U}(x,y)E\left[\mathcal{G}_0(y)\mathcal{G}_0(y')\right]dydy' \\ = & \sigma_{1LU}(x)+\sigma_{0LU}(x). \end{aligned}$$

■

Proof of Theorem 3.2. Under (A2P), Proposition 2.2 (ii) implies that $\theta_L(x)<\theta_U(x)$. It follows from Imbens and Manski (2004) and Stoye (2009) that an asymptotically valid CS for $\theta(x)$ can be constructed from the corresponding one-sided CSs for $\theta_L(x),\theta_U(x)$ respectively. In view of Theorem 3.1, it suffices to prove consistency of $\widehat{\sigma}_L^2(x)$ and $\widehat{\sigma}_U^2(x)$. The latter follows from uniform consistency of $\widehat{F}_j(\cdot|x)$, consistency of $\widehat{f}_{jX}(x)$, and the continuous mapping theorem. ■

Proof of Proposition 4.1. Under the conditions of the Proposition, the following results hold:

$$\sup_{\|x\|\leq b}\left|\widehat{f}_{jX}(x)-f_{jX}(x)\right|=O_p(c_n), \tag{A.8}$$

$$\sup_{\|x\|\leq b,y\in\mathcal{Y}_j}\left|\widehat{F}_j(y|x)-F_j(y|x)\right|=O_p\left(\delta_{nj}^{-1}c_n\right), \tag{A.9}$$

where $c_n=\left(\frac{\log n}{na_n^d}\right)^{1/2}+a_n^m$. The claim in (A.8) follows from Theorem 2 in Hansen (2008) and the fact that $E\left[\widehat{f}_{jX}(x)\right]-f_{jX}(x)=O(a_n^m)$. The claim in (A.9) can be proved by adapting the proofs of Theorem 2 and Theorem 8 in Hansen (2008). The latter is an extension of similar results in Hardle, Janssen, and Serfling (1988) or Akritas and van Keilegom (2001).

We prove (16) only; the proof of (17) is similar. By a second-order Taylor series expansion, we obtain:

$$
\begin{aligned}
& \left[\widehat{\theta}_L(x) - \theta_L(x)\right] I\{\|x\| \le b\} \\
= & \int_0^1 \left[\mu\left(\widehat{F}_1^{-1}(u|x), \widehat{F}_0^{-1}(1-u|x)\right) - \mu\left(F_1^{-1}(u|x), F_0^{-1}(1-u|x)\right)\right] I\{\|x\| \le b\}\, du \\
= & \int_0^1 J_1(u \mid x)\left[\widehat{F}_1^{-1}(u|x) - F_1^{-1}(u|x)\right] I\{\|x\| \le b\}\, du \\
& + \int_0^1 J_0(u \mid x)\left[\widehat{F}_0^{-1}(1-u|x) - F_0^{-1}(1-u|x)\right] I\{\|x\| \le b\}\, du \\
& + \frac{1}{2}\int_0^1 \mu_{11}''(\xi_{ux})\left[\widehat{F}_1^{-1}(u|x) - F_1^{-1}(u|x)\right]^2 I\{\|x\| \le b\}\, du \\
& + \frac{1}{2}\int_0^1 \mu_{00}''(\xi_{ux})\left[\widehat{F}_0^{-1}(1-u|x) - F_0^{-1}(1-u|x)\right]^2 I\{\|x\| \le b\}\, du \\
& + \int_0^1 \mu_{10}''(\xi_{ux})\left[\widehat{F}_1^{-1}(u|x) - F_1^{-1}(u|x)\right]\left[\widehat{F}_0^{-1}(1-u|x) - F_0^{-1}(1-u|x)\right] I\{\|x\| \le b\}\, du,
\end{aligned} \tag{A.10}
$$

where $J_j(u \mid x) \equiv \mu_j'\left(F_1^{-1}(u|x), F_0^{-1}(1-u|x)\right)$ $(j = 0, 1)$, ξ_{ux} is a point on the line segment from $\left(F_1^{-1}(u|x), F_0^{-1}(1-u|x)\right)$ to $(\widehat{F}_1^{-1}(u|x), \widehat{F}_0^{-1}(1-u|x))$.

We now show that the last three terms in (A.10) are $O_p(\delta_n^{-2}c_n^2)$ uniformly in $x \in \mathcal{X}$. Let $e(u|x) = \widehat{F}_1^{-1}(u|x) - F_1^{-1}(u|x)$ and $y = F_1^{-1}(u|x)$. Then there is $Y_k^{(n)}$ (the kth order statistic) such that $y + e(u|x) = \widehat{F}_1^{-1}(u|x) = Y_k^{(n)}$ and $\widehat{F}_j(Y_k^{(n)}|x) \ge u > \widehat{F}_j(Y_{k-1}^{(n)}|x)$; and by (A2U)(ii) we have

$$
\widehat{F}_1(y + e(u|x)|x) \ge u = F_1(y|x) = F_1(y + e(u|x)|x) - f_1(\zeta_{xy}|x)\, e(u|x) > \widehat{F}_j(Y_{k-1}^{(n)}|x),
$$

where ζ_{xy} is between y and $y + e(u|x)$, implying for any u that

$$
\begin{aligned}
& \left|\widehat{F}_1^{-1}(u|x) - F_1^{-1}(u|x)\right| I\{\|x\| \le b\} \\
\le & \frac{\left|\widehat{F}_1(y + e(u|x)|x) - F_1(y + e(u|x)|x)\right| I\{\|x\| \le b\} + \left|\widehat{F}_1(Y_k^{(n)}|x) - \widehat{F}_1(Y_{k-1}^{(n)}|x)\right| I\{\|x\| \le b\}}{f_1(\zeta_{xy}|x)} \\
\le & \frac{\sup_{y \in \mathcal{Y}_1}\left|\widehat{F}_1(y|x) - F_1(y|x)\right| I\{\|x\| \le b\} + \max_k\left|\widehat{F}_1(Y_k^{(n)}|x) - \widehat{F}_1(Y_{k-1}^{(n)}|x)\right| I\{\|x\| \le b\}}{\inf_{y \in \mathcal{Y}_1} f_1(y|x)}
\end{aligned} \tag{A.11}
$$

It follows from (A2U)(iii) and (A.9) that $\sup_{x,u}\left|\widehat{F}_1^{-1}(u|x) - F_1^{-1}(u|x)\right| I\{\|x\| \le b\} = O_p\left(\delta_n^{-1}c_n\right)$. Similarly, we have $\sup_{x,u}\left|\widehat{F}_0^{-1}(1-u|x) - F_0^{-1}(1-u|x)\right| I\{\|x\| \le b\} = O_p\left(\delta_n^{-1}c_n\right)$. Further, by(A3U)(i), we get: uniformly in $x \in \mathcal{X}$,

$$\begin{aligned}
&\left[\widehat{\theta}_L(x) - \theta_L(x)\right] I\{\|x\| \le b\} \\
= &\int_0^1 J_1(u \mid x)\left[\widehat{F}_1^{-1}(u|x) - F_1^{-1}(u|x)\right] I\{\|x\| \le b\}\, du \\
&+ \int_0^1 J_0(u \mid x)\left[\widehat{F}_0^{-1}(1-u|x) - F_0^{-1}(1-u|x)\right] I\{\|x\| \le b\}\, du + O_p\left(\delta_n^{-2} c_n^2\right). \qquad \text{(A.12)}
\end{aligned}$$

Let $L_j(u \mid x) \equiv \int_0^u J_j(s \mid x)\, ds,\ j = 1, 0$. By using integration by parts and a second-order Taylor series expansion, we have

$$\begin{aligned}
&\int_0^1 J_1(u \mid x)\left[\widehat{F}_1^{-1}(u|x) - F_1^{-1}(u|x)\right] du I\{\|x\| \le b\} \\
= &\int \left[L_1(F_1(y|x) \mid x) - L_1\left(\widehat{F}_1(y|x) \mid x\right)\right] dy I\{\|x\| \le b\} \\
= &\int G_{1L}(x,y)\left[F_1(y|x) - \widehat{F}_1(y|x)\right] I\{\|x\| \le b\}\, dy \\
&- \frac{1}{2}\int J_1'(\zeta_{yx} \mid x)\left[F_1(y|x) - \widehat{F}_1(y|x)\right]^2 I\{\|x\| \le b\}\, dy,
\end{aligned}$$

where ζ_{yx} is between $F_1(y|x)$ and $\widehat{F}_1(y|x)$. The second term is $O_p(\delta_n^{-2} c_n^2)$ by (A.9) and the assumptions (A2U)(ii) and (A3P)(i). The first one equals

$$\begin{aligned}
&\int G_{1L}(x,y)\left[F_1(y|x) - \widehat{F}_1(y|x)\right] dy I\{\|x\| \le b\} \\
= &-\frac{\left(na_n^d\right)^{-1}}{\widehat{f}_{1X}(x)} I\{\|x\| \le b\} \sum_{i=1}^n K_{a1}(X_i - x) \int G_{1L}(x,y)\left[I(Y_i \le y) - F_1(y|x)\right] dy
\end{aligned}$$

where we used the expressions for $\widehat{F}_j(y|x)$ and $\widehat{f}_{jX}(x)$. Further, by using (A.8), (A.9), the assumptions (A1U)(i), (A2U)(ii) and (A3U)(ii), we have

$$\begin{aligned}
&\int G_{1L}(x,y)\left[F_1(y|x) - \widehat{F}_1(y|x)\right] dy I\{\|x\| \le b\} \\
= &-\frac{\left(na_n^d\right)^{-1}}{f_{1X}(x)} \sum_{i=1}^n K_{a1}(X_i - x) \int G_{1L}(x,y)\left[I(Y_i \le y) - F_1(y|x)\right] dy I\{\|x\| \le b\} \\
&+ \frac{\widehat{f}_{1X}(x) - f_{1X}(x)}{f_{1X}(x)} \int G_{1L}(x,y)\left[\widehat{F}_1(y|x) - F_1(y|x)\right] dy I\{\|x\| \le b\} \\
= &-\frac{\left(na_n^d\right)^{-1}}{f_{1X}(x)} I\{\|x\| \le b\} \sum_{i=1}^n K_{a1}(X_i - x) \int G_{1L}(x,y)\left[I(Y_i \le y) - F_1(y|x)\right] dy + O_p\left(\delta_n^{-2} c_n^2\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\int_0^1 J_1(u \mid x)\left[\widehat{F}_1^{-1}(u|x) - F_1^{-1}(u|x)\right] du I\{\|x\| \le b\} \\
= &-\frac{\left(na_n^d\right)^{-1}}{f_{1X}(x)} I\{\|x\| \le b\} \sum_{i=1}^n K_{a1}(X_i - x) \int G_{1L}(x,y)\left[I(Y_i \le y) - F_1(y|x)\right] dy \\
&+ O_p\left(\delta_n^{-2} c_n^2\right). \qquad \text{(A.13)}
\end{aligned}$$

Similarly, we have

$$
\begin{aligned}
& \int_0^1 J_0\left(u \mid x\right)\left[\widehat{F}_0^{-1}\left(1-u|x\right)-F_0^{-1}\left(1-u|x\right)\right] du I\left\{\|x\| \leq b\right\} \\
= & \int\left[L_0\left(1-\widehat{F}_0\left(y|x\right) \mid x\right)-L_0\left(1-F_0\left(y|x\right) \mid x\right)\right] dy I\left\{\|x\| \leq b\right\} \\
= & \int G_{0L}(x,y)\left[F_0\left(y|x\right)-\widehat{F}_0\left(y|x\right)\right] dy I\left\{\|x\| \leq b\right\}+O_p(\delta_n^{-2}c_n^2) \\
= & -\frac{\left(na_n^d\right)^{-1}}{f_{0X}\left(x\right)} I\left\{\|x\| \leq b\right\} \sum_{i=1}^n K_{a0}\left(X_i-x\right) \int G_{0L}(x,y)\left[I\left(Y_i \leq y\right)-F_0\left(y|x\right)\right] dy \\
& +O_p(\delta_n^{-2}c_n^2). \qquad \text{(A.14)}
\end{aligned}
$$

Combining (A.12), (A.13) and (A.14) yields (16). ■

Proof of Theorem 4.2. Note that

$$
\widehat{\theta}_L-\theta_L=\frac{1}{n}\sum_{l=1}^n\left[\widehat{\theta}_L\left(X_l\right)-\theta_L\left(X_l\right)\right] I_{bl}+\frac{1}{n}\sum_{l=1}^n\left[\theta_L\left(X_l\right) I_{bl}-\theta_L\right]. \qquad \text{(A.15)}
$$

For $j=1,0$, let $U_{ji}\left(y\right)=I\left(Y_i \leq y\right)-F_j\left(y|X_i\right)$. Then by (C1), we get: $E\left(U_{ji}\left(y\right)|X_i,D_i=j\right)=0$. It follows from Proposition 4.1 that

$$
\begin{aligned}
& \frac{1}{n}\sum_{l=1}^n\left[\widehat{\theta}_L\left(X_l\right)-\theta_L\left(X_l\right)\right] I_{bl}= \\
& -\frac{1}{n\left(n-1\right)a_n^d}\sum_{l=1}^n I_{bl}\sum_{i\neq l}^n\frac{1}{f_{1X}\left(X_l\right)}K_{a1}\left(X_i-X_l\right)\int G_{1L}\left(X_l,y\right)\left[I\left(Y_i \leq y\right)-F_1\left(y|X_l\right)\right] dy \\
& -\frac{1}{n\left(n-1\right)a_n^d}\sum_{l=1}^n I_{bl}\sum_{i\neq l}^n\frac{1}{f_{0X}\left(X_l\right)}K_{a0}\left(X_i-X_l\right)\int G_{0L}\left(X_l,y\right)\left[I\left(Y_i \leq y\right)-F_0\left(y|X_l\right)\right] dy \\
& +\frac{1}{n}\sum_{l=1}^n R_n^L\left(X_l\right) \\
= & -\frac{1}{n\left(n-1\right)a_n^d}\sum_{l=1}^n\sum_{i\neq l}^n\frac{1}{f_{1X}\left(X_l\right)}I_{bl}K_{a1}\left(X_i-X_l\right)\int G_{1L}\left(X_l,y\right)U_{1i}\left(y\right) dy \\
& -\frac{1}{n\left(n-1\right)a_n^d}\sum_{l=1}^n\sum_{i\neq l}^n\frac{1}{f_{0X}\left(X_l\right)}I_{bl}K_{a0}\left(X_i-X_l\right)\int G_{0L}\left(X_l,y\right)\left[U_{0i}\left(y\right)+F_0\left(y|X_i\right)-F_0\left(y|X_l\right)\right] dy \\
& -\frac{1}{n\left(n-1\right)a_n^d}\sum_{j=0,1}\sum_{l=1}^n\sum_{i\neq l}^n\frac{1}{f_{jX}\left(X_l\right)}I_{bl}K_{aj}\left(X_i-X_l\right)\int G_{jL}\left(X_l,y\right)\left[F_j\left(y|X_i\right)-F_j\left(y|X_l\right)\right] dy \\
& +o_p\left(n^{-1/2}\right),
\end{aligned}
$$

where we have used the assumption that $\sqrt{n}\delta_n^{-2}c_n^2 = o(1)$. Let

$$\begin{aligned} U_n &= \frac{2}{n(n-1)}\sum_{l=1}^{n}\sum_{i\geq l}^{n}\left[\begin{array}{c} \frac{1}{f_{1X}(X_l)a_n^d}K_{a1}(X_i-X_l)\,I_{bl}\int G_{1L}(X_l,y)\,U_{1i}(y)\,dy \\ +\frac{1}{f_{0X}(X_l)a_n^d}K_{a0}(X_i-X_l)\,I_{bl}\int G_{0L}(X_l,y)\,U_{0i}(y)\,dy \\ +\frac{1}{f_{1X}(X_i)a_n^d}K_{a1}(X_l-X_i)\,I_{bi}\int G_{1L}(X_i,y)\,U_{1l}(y)\,dy \\ +\frac{1}{f_{0X}(X_i)a_n^d}K_{a0}(X_l-X_i)\,I_{bi}\int G_{0L}(X_i,y)\,U_{0l}(y)\,dy \end{array}\right] \\ &= \frac{2}{n(n-1)}\sum_{l=1}^{n}\sum_{i\geq l}^{n}P_n(W_i,W_l) \end{aligned}$$

and

$$B_n = \frac{1}{n(n-1)a_n^d}\sum_{j=0,1}\sum_{l=1}^{n}\sum_{i\neq l}^{n}\frac{1}{f_{jX}(X_l)}I_{bl}K_{aj}(X_i-X_l)\int G_{jL}(X_l,y)\left[F_j(y|X_i)-F_j(y|X_l)\right]dy$$

where $W_i = (X_i, Y_i, D_i)$. Then

$$\frac{1}{n}\sum_{l=1}^{n}\left[\widehat{\theta}_L(X_l)-\theta_L(X_l)\right]I_{bl} = -U_n - B_n + o_p\left(n^{-1/2}\right).$$

Let $r_n(W_i) = E\left[P_n(W_i,W_l)\,|W_i\right]$ and $\widehat{U}_n = 2n^{-1}\sum_{i=1}^{n} r_n(W_i)$. It is easy to show that $E\left[P_n^2(W_i,W_l)\right] = O\left(a^{-d}\right)$, so $n^{-1}E\left[P_n^2(W_i,W_l)\right] = O\left(\left(na^d\right)^{-1}\right) = o(1)$. By using Lemma 3.1 in Powell, Stock, and Stoker (1989), we have:$\sqrt{n}U_n = \sqrt{n}\widehat{U}_n + o_p(1)$. Now we show:

$$r_n(W_i) = r_L(W_i) + t_n(W_i),$$

where

$$\begin{aligned} r_L(W_i) &= \frac{I\{D_i=1\}}{p(X_i)}I_{bi}\int G_{1L}(X_i,y)\,U_{1i}(y)\,dy + \frac{I\{D_i=0\}}{1-p(X_i)}I_{bi}\int G_{0L}(X_i,y)\,U_{0i}(y)\,dy, \\ t_n(W_i) &= t_{n1}(W_i) + t_{n2}(W_i), \end{aligned}$$

and for $j = 1, 0$,

$$t_{nj}(W_i) = I\{D_i=j\}\int\left[\int\frac{I\{\|x\|\leq b\}}{p_j(x)\,a_n^d}K\left(\frac{X_i-x}{a}\right)G_{jL}(x,y)\,dx - \frac{1}{p_j(X_i)}I_{bi}G_{jL}(X_i,y)\right]U_{ji}(y)\,dy.$$

To see this, we note that

$$\begin{aligned}
r_n(W_i) &= E\left[P_n(W_i,W_l)|W_i\right] \\
&= E\left[\frac{1}{f_{1X}(X_l)a_n^d}K_{a1}(X_i-X_l)I_{bl}\int G_{1L}(X_l,y)U_{1i}(y)dy|W_i\right] \\
&\quad +E\left[\frac{1}{f_{0X}(X_l)a_n^d}K_{a0}(X_i-X_l)I_{bl}\int G_{0L}(X_l,y)U_{0i}(y)dy|W_i\right] \\
&\quad +E\left[\frac{1}{f_{1X}(X_i)a_n^d}K_{a1}(X_l-X_i)I_{bi}\int G_{1L}(X_i,y)U_{1l}(y)dy|W_i\right] \\
&\quad +E\left[\frac{1}{f_{0X}(X_i)a_n^d}K_{a0}(X_l-X_i)I_{bi}\int G_{0L}(X_i,y)U_{0l}(y)dy|W_i\right] \\
&= E\left[\frac{1}{f_{1X}(X_l)a_n^d}K_{a1}(X_i-X_l)I_{bl}\int G_{1L}(X_l,y)U_{1i}(y)dy|W_i\right] \\
&\quad +E\left[\frac{1}{f_{0X}(X_l)a_n^d}K_{a0}(X_i-X_l)I_{bl}\int G_{0L}(X_l,y)U_{0i}(y)dy|W_i\right] \\
&\equiv r_{n1}(W_i)+r_{n0}(W_i),
\end{aligned}$$

where we have used the fact that $E(U_{jl}(y)|X_l,D_l=j)=0$. For $j=1,0$, we have:

$$\begin{aligned}
&r_{nj}(W_i) \\
&= E\left[\frac{1}{f_{jX}(X_l)a_n^d}K_{a1}(X_i-X_l)I_{bl}\int G_{jL}(X_l,y)U_{ji}(y)dy|W_i\right] \\
&= I\{D_i=j\}\int\left[\int\frac{I\{\|x\|\le b\}}{p_j(x)a_n^d}K\left(\frac{X_i-x}{a}\right)G_{jL}(x,y)dx\right]U_{ji}(y)dy \\
&= \frac{I\{D_i=j\}}{p_j(X_i)}I_{bi}\int G_{jL}(X_i,y)U_{ji}(y)dy+t_{nj}(W_i).
\end{aligned}$$

It remains to show: $n^{-1}\sum_{i=1}^n t_n(W_i)=o_p\left(n^{-1/2}\right)$. Since $E[t_n(W_i)]=0$, we only need to show: $Var[t_n(W_i)]=o(1)$. This follows from Lemma 6.4 in Gine, Mason, and Zaitsev (2003). Thus, we have

$$\begin{aligned}
U_n &= 2n^{-1}\sum_{i=1}^n r_n(W_i)+o_p\left(n^{-1/2}\right) \\
&= 2n^{-1}\sum_{i=1}^n \left[r_L(W_i)+t_n(W_i)\right]+o_p\left(n^{-1/2}\right) \\
&= 2n^{-1}\sum_{i=1}^n \left[r_L(W_i)-E(r_L(W_i))\right]+o_p\left(n^{-1/2}\right).
\end{aligned}$$

Now we consider the bias term B_n. For $i \neq l$, we have:

$$\begin{aligned}
E(B_n) &= \sum_{j=0,1} E\left\{\frac{1}{f_{jX}(X_l)\, a_n^d} I_{bl} K_{aj}(X_i - X_l) \int G_{jL}(X_l, y)\left[F_j(y|X_i) - F_j(y|X_l)\right] dy\right\} \\
&= \sum_{j=0,1} E\left\{\frac{1}{f_X(X_l)\, a_n^d} I_{bl} K\left(\frac{X_i - X_l}{a}\right) \int G_{jL}(X_l, y)\left[F_j(y|X_i) - F_j(y|X_l)\right] dy\right\} \\
&= \sum_{j=0,1} E\left\{\frac{1}{f_X(X_l)\, a_n^d} I_{bl} \int G_{jL}(X_l, y)\, E\left[K\left(\frac{X_i - X_l}{a}\right)\left[F_j(y|X_i) - F_j(y|X_l)\right] | X_l\right] dy\right\} \\
&= O(a^m),
\end{aligned}$$

where we have used Lemma B.1 in Fan and Li (1996). As a result, we obtain:

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\theta}_L - \theta_L\right) \\
=\ & -\sqrt{n}\frac{U_n}{2} - \sqrt{n}B_n + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\theta_L(X_l)\, I_{bl} - \theta_L\right] + o_p(1) \\
=\ & \frac{1}{\sqrt{n}}\sum_{l=1}^{n}\left[\theta_L(X_l)\, I_{bl} - \theta_L - r_L(W_i) - \left\{E(-r_L(W_i)) + E(\theta_L(X_l)\, I_{bl} - \theta_L)\right\}\right] \\
& + \sqrt{n}E\left(\theta_L(X_l)\, I_{bl} - \theta_L\right) + o_p(1).
\end{aligned}$$

Similarly, we can show that

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\theta}_U - \theta_U\right) \\
=\ & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\theta_U(X_i)\, I_{bi} - \theta_U - r_U(W_i) - \left\{E\left[\theta_U(X_i)\, I_{bi} - \theta_U - r_U(W_i)\right]\right\}\right] \\
& + \sqrt{n}E\left(\theta_U(X_i)\, I_{bi} - \theta_U\right) + o_p(1).
\end{aligned}$$

Theorem 4.2 follows. ■

Proof of Theorem 4.3. It is similar to that of Theorem 3.2 and thus omitted. ■

References

[1] Aakvik, A. , J. Heckman, and E. Vytlacil (2005), "Treatment Effects for Discrete Outcomes When Responses to Treatment Vary Among Observationally Identical Persons: An Application to Norwegian Vocational Rehabilitation Programs," *Journal of Econometrics* 125, 15-51.

[2] Abbring, J. H. and J. Heckman (2007), "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in *The Handbook of Econometrics* 6B, 5145-5301.

[3] Akritas, M.G. and I. van Keilegom (2001), "Nonparametric Estimation of the Residual Distribution," *Scandinavian Journal of Statistics* 28, 549-567.

[4] Andrews, D. W. K. and M. M. A. Schafgans (1998), "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies* 65, 497-517.

[5] Biddle, J., L. Boden and R. Reville (2003), "A Method for Estimating the Full Distribution of a Treatment Effect, With Application to the Impact of Workfare Injury on Subsequent Earnings," Mimeo, Michigan State University.

[6] Bhattacharya, D. (2007), "Inference on Inequality From Household Survey Data," *Journal of Econometrics* 137, 674-707.

[7] Cambanis, S., G. Simons and W. Stout (1976), "Inequalities for $\mathcal{E}k(X, Y)$ when the Marginals are Fixed," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 36, 285-294.

[8] Carneiro, P. , K. T. Hansen, and J. Heckman (2003), "Estimating Distributions of Treatment Effects With an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review* 44(2), 361-422.

[9] Cowell, F.A. (1995). *Measuring Inequality.* LSE Handbooks in Economics.

[10] Dehejia, R. and S. Wahba (1999), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94, 1053-1062.

[11] Fan, Y. (2008), "Confidence Sets for Distributions of Treatment Effects with Covariates," Manuscript, Vanderbilt University.

[12] Fan, Y. and Q. Li (1996), "Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms," *Econometrica* 64, 865-890.

[13] Fan, Y. and S. Park (2007), "Sharp Bounds on the Distribution of the Treatment Effect and Their Statistical Inference," Forthcoming in *Econometric Theory.*

[14] Fan, Y. and S. Park (2008a), "Partial Identification of the Distribution of Treatment Effects and its Confidence Sets," Forthcoming in *Advances in Econometrics: Nonparametric Econometric Methods* (vol. 24).

[15] Fan, Y. and S. Park (2008b), "Confidence Intervals for the Quantile of Treatment Effects in Randomized Experiments," Working Paper, Vanderbilt University.

[16] Fan, Y. and J. Wu (2007), "Sharp Bounds on the Distribution of the Treatment Effect in Switching Regimes Models," Forthcoming in the *Review of Economic Studies.*

[17] Firpo, S. (2007), "Inequality Treatment Effects," Manuscript.

[18] Firpo, S. and G. Ridder (2008), "Bounds on Functionals of the Distribution of Treatment Effects," Working Paper.

[19] Gine, E., D.M. Mason, and A. Yu. Zaitsev (2003), "The L_1-Norm Density Estimator Process," *The Annals of Probability* 31, 719-768.

[20] Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66, 315-331.

[21] Hardle, W., P. Janssen, and R. Serfling (1988), "Strong Uniform Consistency Rates for Estimators of Conditional Functionals," *The Annals of Statistics* 16, 1428-1449.

[22] Hansen, B.E. (2008), "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory* 24, 726-748.

[23] Hardle, W., P. Janssen, and R. Serfling (1988), "Strong Uniform Consistency Rates for Estimators of Conditional Functionals," *The Annals of Statistics* 16, 1428-1449.

[24] Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998a), "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66, 1017-1098.

[25] Heckman, J., H. Ichimura, and P. Todd (1998b), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261-294.

[26] Heckman, J. and S. Navarro-Lozano (2004), "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models," *The Review of Economics and Statistics* 86, 30-57.

[27] Heckman, J., J. Smith, and N. Clements (1997), "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *Review of Economic Studies* 64, 487-535.

[28] Heckman, J. and E. Vytlacil (2007a), "Econometric Evaluation of Social Programs. Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *The Handbook of Econometrics* 6B, 4779-4874.

[29] Heckman, J. and E. Vytlacil (2007b), "Econometric Evaluation of Social Programs. Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments," in *The Handbook of Econometrics* 6B, 4875-5143.

[30] Hirano, K., Imbens, G. W. and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71, 1161-1189.

[31] Horowitz, J. and C.F. Manski (2000), "Nonparametric Analysis of Randomized Experiments with Missing Covariates and Outcome Data," *Journal of the American Statistical Association* 95, 77-84.

[32] Imbens, G. W. and C. F. Manski (2004), "Confidence Intervals For Partially Identified Parameters," *Econometrica* 72, 1845–1857.

[33] Khan, S. and E. Tamer (2009), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," Working Paper, Northwestern University.

[34] Lee, M. J. (2005). *Micro-Econometrics for Policy, Program, and Treatment Effects*. Oxford University Press.

[35] Manski, C. F. (1997), "Monotone Treatment Effect," *Econometrica* 65, 1311-1334.

[36] Newey, W. (1994), "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory* 10, 1-21.

[37] Polonik, W. and Q. Yao (2002), "Set-Indexed Conditional Empirical and Quantile Processes Based on Dependent Data," *Journal of Multivariate Analysis* 80, 234-255.

[38] Powell, J.L., J.H. Stock, and T.M. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.

[39] Rachev, S.T. and L. Ruschendorf (1998). *Mass Transportation Problems, Vol. I: Theory.* Springer.

[40] Rosenbaum, P. R. and D. B. Rubin (1983a), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society, Series B* 45, 212-218.

[41] Rosenbaum, P. R. and D. B. Rubin (1983b), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70, 41-55.

[42] Schafgans, M. M. A. and V. Zinde-Walsh (2002), "On Intercept Estimation in the Sample Selection Model," *Econometric Theory* 18, 40-50.

[43] Stoye, J. (2008), "Partial Identification of Spread Parameters," Manuscript, NYU.

[44] Stoye, J. (2009), "More on Confidence Intervals for Partially Identified Parameters," Forthcoming in *Econometrica*.

[45] Tchen, A. H. (1980), "Inequalities for Distributions with Given Marginals," *Annals of Probability* 8, 814-827.

[46] van der Vaart , A. W. and Jon A. Wellner (1996). *Weak Convergence and Empirical Processes.* Springer.