Double Machine Learning Density Estimation for Local Treatment Effects with Instruments

Yonghan Jung
Purdue University
jung222@purdue.edu

Jin Tian
Iowa State University
jtian@iastate.edu

Elias Bareinboim
Columbia University
eb@cs.columbia.edu

Abstract

It is common to quantify causal effects with mean values, which, however, may fail to capture significant distribution differences of the outcome under different treatments. We study the problem of estimating the density of the causal effect of a binary treatment on a continuous outcome given a binary instrumental variable in the presence of covariates. Specifically, we consider the local treatment effect, which measures the effect of treatment among those who comply with the assignment under the assumption of monotonicity (only the ones who were offered the treatment take it). We develop two families of methods for this task, *kernel-smoothing* and *model-based* approximations – the former smoothes the density by convoluting with a smooth kernel function; the latter projects the density onto a finite-dimensional density class. For both approaches, we derive double/debiased machine learning (DML) based estimators. We study the asymptotic convergence rates of the estimators and show that they are robust to the biases in nuisance function estimation. We illustrate the proposed methods on synthetic data and a real dataset called 401(k).

1 Introduction

Controlled experimentation is one powerful tool used throughout the biological, medical, and social sciences to infer the effect of a certain treatment on a given outcome. The idea is to randomize the treatment assignment so as to neutralize the effect of the unobserved confounders. In some practical settings, however, it may be challenging to ascertain that individuals who are selected for treatment will follow their recommendations. In fact, issues of non-compliance and unmeasured confounders are quite common and lead to the non-identification of treatment effects in such cases [29, 50, 32, 56].

An approach known as instrumental variables (IVs) has been proposed to try to circumvent this issue [68]. The idea is to find a variable (or set) that is not the target of the analysis by itself, but that will help to control for the unobserved confounding between the treatment and the outcome. In particular, IVs are special variables that (i) influence the treatment, (ii) do not directly influence the outcome, and (iii) are not affected by unmeasured confounders. For concreteness, consider a study of the effect of 401(k) participation (X) on the distribution of net financial assets (Y) [2]. This setting is represented in the causal graph in Fig. 1. Note that there exists a dashed-bidirected arrow between X and Y, which in graphical language represents unobserved confounding affecting both X and Y. The variable Z in this model represents the eligibility of 401(k). We note that Z qualifies as an instrument in this case – (i) it does affect the participation of 401(k) (X), (ii) has no direct influence on the net financial asset (Y), (iii) is not affected by unmeasured confounders between X and Y. The variable W represents observed covariates (e.g., income, gender, family size, etc.).

We are interested in the particular setting where only individuals who were offered the treatment may have access to it [31]. For instance, in the case of 401(k) participation ($X = 1$), only eligible individuals ($Z = 1$) would be allowed to join the program. This assumption is known as *monotonicity*,

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Figure 2: Densities of outcome Y among compliers under the treatment $X = 1$. All densities have a mean 0 and a variance 2.

which rules out the possibility that any units would respond contrary to the instrument. Under monotonicity, the causal effect in the subpopulation whose actual treatment X coincides with the assigned treatment Z (called compliers) is identifiable [31, 2]. The average treatment effect (ATE) for the compliers is called `Local ATE' (LATE) (or Complier average causal effects, CACE) [31].

The most common quantification of effects in IV settings found in practice is the average (e.g., LATE). The average is certainly an informative summary; however, it may fail to capture significant differences in the causal distributions of the outcome. For instance, consider Fig. 2 that shows the densities of outcomes Y under treatments $X = 1$ among compliers (generated from samples drawn from four synthetic data generating processes represented by the IV graph in Fig. 1, as discussed in Sec. 5). All of the distributions have the same mean 0 and variance 2. However, the difference in the distributions is self-evident.

Most of the prior work on quantifying treatment effects on outcome distributions focuses on estimating cumulative distribution functions (CDFs) or quantiles, and little attention has been given to estimating densities of treatment effects (refer to Sec. 1.1 for further comparison). As a complement to CDFs, densities have the benefits of providing more visually interpretable information of the distribution and enabling researchers/practitioners to generate counterfactual samples. One challenge with estimating densities is that while CDFs are pathwise-differentiable and enjoy $\sqrt{n}$-rate estimators (n is the size of data), densities are not (i.e., they are not regular), and therefore possess no influence functions nor $\sqrt{n}$-rate estimators without approximations [7, Ch. 3].

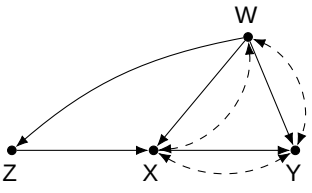

Figure 1: A causal graph for the IV setting. Bidirected arrows encode unmeasured confounders.

In this paper, our goal is to provide methods to estimate densities of local treatment effects in IV settings under the monotonicity assumption. We develop two families of methods for this task based on kernel-smoothing and model-based approximations. The former smooths the density by convolution with a smooth kernel function; the latter projects the density onto a finite-dimensional density class based on a distributional distance measure. For both approaches, we construct double/debiased machine learning (DML) style density estimators [43, 54, 52, 70, 13]. We analyze the asymptotic convergence properties of the estimators, showing that they can converge fast ($\sqrt{n}$-rate) even when nuisance estimates converge slowly (e.g., $n^{1/4}$ rate) (`debiasedness').[1] We illustrate the proposed methods on synthetic and real data.

1.1 Related work

Our work touches different areas, which we will discuss next.

DML-based causal effect estimators. The DML framework has been adapted for estimating the average causal effect under the setting where back-door criterion [50, Sec. 3.3.1] (also known as ignorability [57]) holds (e.g., [12, 19]). Recently, DML-based causal effect estimators have been developed for any identifiable causal functionals in a given causal graph [33, 34].

Local average & quantile treatment effect. The formal identification results for LATE under the monotonicity assumption in IV settings were developed by [3]. Building on these results, semiparametric estimation for LATE has received remarkable attention [60, 23, 62, 48], including robust LATE estimators that achieve debiasedness [43, 40, 38, 64]. As shown in Fig. 2, however,

[1] Also known as `nonparametric doubly robust' [37] or `rate doubly robust' [59].

the average is sometimes insufficient to capture the effects of the treatment on the distributions of outcomes. To address this issue, a problem of estimating quantiles or CDFs has taken attention.

A common approach to quantifying LTEs is to estimate quantiles or CDFs, which can be studied based on the LATE estimation [1, 2, 15, 24, 16, 30, 45, 18, 69]) using the fact that the expectation of $1_{Y \le y}(Y)$, an indicator that outcome Y falls short of threshold y, reduces to the quantiles (i.e., replacing Y in LATE with $1_{Y \le y}(Y)$).

Non-regular target estimand. Densities are an example of non-regular targets [7, Chap. 3]. One can approximate a non-regular target with smooth ones such that influence functions and the estimators can be derived. Two popular used approximation approaches are kernel-smoothing-based (e.g., [52, 6, 42, 19, 35]) and model-based (e.g., [46, 52, 21, 41, 40, 39]).

Causal density estimation. There is limited literature on estimating the density of treatment effects. Most of the results assume that the ignorability/backdoor admissibility holds [55, 49]. [22] used the kernel-smoothing technique to estimate the density of a treatment effect and [42] provided a kernel-smoothing-based density estimator that achieves doubly robustness and debiasedness building on top of the work in [53]. Recently, [39] investigated a model-based approach and developed estimators that achieve debiasedness properties. Under the IV setting, [10] provides a local polynomial regression-based density estimator for local treatment effects, and we are not aware of any work studying debiased density estimators. As mentioned, this paper investigates both kernel-smoothing and model-based approaches for estimating local treatment effects under IV settings and develops DML-style density estimators for both.

2 LTE Estimation – Problem setup

Each variable is represented with a capital letter (e.g., X) and its realized value with a small letter (x). For a discrete (e.g., binary) random variable X, we use $1_x(X)$ to represent the indicator function such that $1_x(X) = 1$ if $X = x$; $1_x(X) = 0$ otherwise. For a continuous variable X with a probability density $p(x)$ of a distribution P and a function $f(x)$, $E_P[f(X)] \coloneqq \int_X f(x)p(x)\,d[x]$ where X is the domain for X, and $\|f(X)\| \coloneqq \sqrt{E_P[(f(X))^2]}$. $\hat{f}_n$ is said to converge to f at rate r_n if $\|\hat{f}_n(x) - f(x)\| = O_P(1/r_n)$. For a dataset $\mathcal{D} = \{V_i\}_{i=1}^n$, we use $E_{\mathcal{D}}[f(V)] \coloneqq (1/n)\sum_{i=1}^n f(V_i)$ to denote the empirical mean of $f(V)$ with $\mathcal{D}$.

Structural Causal Models (SCMs). We use the language of SCMs as our basic semantic and inferential framework [50, 4]. An SCM M is a quadruple $M = \langle U, V, P(U), F \rangle$ where U is a set of exogenous (latent) variables following a joint distribution $P(u)$, and V is a set of endogenous (observable) variables whose values are determined by functions $F = \{f_{V_i}\}_{V_i \in V}$ such that $V_i \leftarrow f_{V_i}(pa_i, u_i)$ where $PA_i \subseteq V$ and $U_i \subseteq U$. Each SCM M induces a distribution $P(v)$ and a causal graph $G = G(M)$ over V in which there exists a directed edge from every variable in PA_i to V_i and dashed-bidirected arrows encode common latent variables (e.g., see Fig. 1). Within the structural semantics, performing an intervention and setting $X = x$ is represented through the do-operator, $do(X = x)$, which encodes the operation of replacing the original equation of X (i.e., $f_X(pa_x, u_x)$) by the constant x and induces a submodel M_x and an interventional distribution $P(v|do(x))$. For any variable $Y \in V$, the potential response $Y_x(u)$ is defined as the solution of Y in the submodel M_x given $U = u$, which induces a counterfactual variable Y_x.

Local Treatment Effect (LTE) with IV. We consider the IV setting represented by the causal graph G in Fig. 1[2], where Z is a binary instrument with domain $\{0, 1\}$, X is a binary treatment with domain $\{0, 1\}$, and Y is a (set of) continuous outcomes with bounded domain in $\mathbb{R}^d$, and W is a set of covariates. G satisfies the IV assumption that Z has no direct influence on outcome Y and is not affected by unmeasured confounders between X and Y.

The causal density $p(y|do(x))$ is not identifiable from the observed density $p(x, y, z, w)$ due to unobserved confounders between X and Y. However, the effect may be recovered for certain subpopulation under additional assumptions. Formally, a unit in the population is an always-taker if

[2] It is common in the literature to define IV assumptions in terms of conditional independences among counterfactuals [51, 9, 8, 2, 60, 47, 64], whose connection with the causal graph in Fig. 1 is discussed in Assumption A.1

$X_{Z=1} = X_{Z=0} = 1$, a never-taker if $X_{Z=1} = X_{Z=0} = 0$, a complier if $X_{Z=1} = 1; X_{Z=0} = 0$, and a defier if $X_{Z=1} = 0; X_{Z=0} = 1$ [3, 2]. We will make the following assumptions after the literature.

Assumption 1 (Monotonicity). There are no defiers: $X_{Z=1} \geq X_{Z=0}$.

Assumption 2 (Positivity). $P(x|z; w) > 0; P(z|w) > 0$ for any $x; z; w$.

Let C denote the event that a unit is a complier. For a given constant x and a variable X, let x^a denote the event $X = a$. The LTE $p(y_x|C)$ is identifiable under monotonicity and is given by [1, 2]:

$$p(y_x|C) = \frac{E_P\left[p(y|x, z^x; W)P(x|z^x; W) - p(y|x, z^{1-x}; W)P(x|z^{1-x}; W)\right]}{E_P\left[P(x^1|z^1; W) - P(x^1|z^0; W)\right]}; \quad (1)$$

where the expectation is over W. In this paper, we aim to estimate the LTE density $p(y_x|C)$ in Eq. (1). We will make the following mild assumption on the target, popularly employed in density estimation (e.g., [44, 25, 27, 61, 26, 42]).

Assumption 3. For any $x; z; w; y$, $p(y|w; z; x)$, $p(y|z; x)$ and $p(y_x|C)$ are bounded, and $p(y_x|C)$ is twice differentiable.

Double/Debiased Machine Learning (DML) method [13]. Let $\theta := \theta_{P^0}$ denote a functional of an arbitrary distribution P^0. We use P to denote the true distribution such that θ_P. Let $\theta_0 := \theta_P$ denote the true parameter to be estimated. To estimate θ_0, DML-based estimators use a Neyman Orthogonal score $\psi(V; \theta_0; \eta)$ (where $\eta := \eta_{P^0}$ is a set of nuisance parameters and $\eta_0 = \eta_P$ denotes the true nuisances), a function such that $E_P[\psi(V; \theta_0; \eta_0)] = 0$, $(\partial/\partial\eta)|_{\eta=\eta_0} E_P[\psi(V; \theta_0; \eta)] = 0$. Given ψ, an DML estimator is constructed using the cross-fitting technique as follows: For randomly split halves of D denoted $\{D_0; D_1\}$, let $\hat{\eta}_p$ for $p \in \{0; 1\}$ denote the estimates for η_0 from D_p. Let T_p denote a solution such that $E_{D_{1-p}}[\psi(V; T_p; \hat{\eta}_p)] = o_P(N^{-1/2})$. Then, $\bar{T} := (T_0 + T_1)/2$ is an DML estimator [13, Def. 3.1]. In addition to being consistent, the estimator $\bar{T}$ exhibits a robustness property called debiasedness: $\bar{T}$ converges to θ_0 in the root-N rate even when $\hat{\eta}$ converges to η_0 in slower $N^{-1/4}$ rate [13, Thm. 3.1]. A Neyman Orthogonal Score can be derived by adding to its influence function [14, Thm. 1]. An influence function of the functional θ_P is defined as a solution satisfying $E_P[\phi] = 0$, $E_P[\phi^2] < \infty$, and $(\partial/\partial t)|_{P_t|_{t=0}} = E_P[\phi(V; \theta; \eta)S_t(V; t = 0)]$ where $P_t := P(v)(1 + t g(v))$ for $t \in R$ and any bounded mean-zero function $g(v)$ over V, and $S_t(v; t = 0) := (\partial/\partial t) \log P_t(v)|_{t=0}$ [63, Chap. 25].

The proofs are provided in Appendix B in suppl. material.

3 Kernel-smoothing-based approach

In this section, we develop a kernel-smoothing-based approach for estimating the LTE density. The kernel-smoothing technique approximates a non-pathwise-differentiable target estimand with a differentiable estimand by convoluting the density with a kernel function $K(y)$. Properties of the kernel function includes symmetry about the origin (i.e., $\int_Y K(y)\, d[y] = 0$), non-negativity ($0 < K(y) < \infty; \forall y \in Y$), and integrates to 1 (i.e., $\int_Y K(y)\, d[y] = 1$) [66, Chap. 4.2].

We consider a product kernel $K_{h;y}(y^0) := h^{-d} \prod_{j=1}^{d} K((y_j - y_j^0)/h)$ with given bandwidth $h \in R$ and a fixed point $y = \{y_j\}_{j=1}^{d} \in R^d$. We assume that the kernel of interest has a bounded second moment and norm: i.e., $\sigma_2(K) := \int_Y y^2 K(y)\, d[y] < \infty$ and $\|K(y)\| < \infty$ following [27, 61]. Example of kernels include Gaussian kernel $K(u) = (1/\sqrt{2\pi}) \exp(-u^2/2)$ for $u \in R$, Epanechnikov kernel $K(u) = (3/4)(1 - u^2)1_{|u| \leq 1}(u)$, Quadratic kernel $K(u) = (15/16)(1 - u^2)^2 1_{|u| \leq 1}(u)$, Cosine kernel $k(u) = (\pi/4)\cos(\pi u/2)1_{|u| \leq 1}(u)$, etc.

For convenience, we denote the target estimand $p(y_x|C)$ by $\theta(y)$. In the kernel-smoothing-based approach, we will aim to estimate a kernel-smoothed approximation $\theta_h(y)$ defined as follows:

$$\theta_h(y) := \int_Y \theta(y^0) K_{h;y}(y^0)\, d[y^0] = [K_{h;y}(Y)]; \quad (2)$$

where $[f(Y)]$ is an expectation of a function $f(Y)$ w.r.t. $\theta(y)$, which is specified as

$$[f(Y)] := \frac{E_P\left[E_P[f(Y)1_x(X)|z^x; W]\right] - E_P\left[f(Y)1_x(X)|z^{1-x}; W\right]}{E_P\left[P(x^1|z^1; W) - P(x^1|z^0; W)\right]}. \quad (3)$$

The second equality in Eq. (2) is by Eq. (1). For a target estimand $\mathbb{E}[f(Y)]$, we will denote nuisances by $\pi_z(w) \equiv P(z|w)$, $\mu_x(z;w) \equiv P(x|z;w)$, and $\gamma(x;z;w)[f(Y)] \equiv \mathbb{E}_P[f(Y)1_x(X)|z;w]$, shortly $\gamma(\cdot;\cdot;\cdot)$.

We aim to construct a DML estimator for the estimand θ_h. Toward this goal, we will first derive a Neyman orthogonal score for θ_h. Since a Neyman orthogonal score can be constructed based on moment score functions (a function of parameters such that its expectation is 0 at the true parameters) [14, Thm. 1], we start by defining the moment score function. Let

$$\Delta_x \equiv \mathbb{E}_P[\mu_{x^1}(z^1;W) - \mu_{x^1}(z^0;W)]; \tag{4}$$

$$V_X(f,g) \equiv \frac{1_{z^1}(Z) - 1_{z^0}(Z)}{\pi_Z(W)}\{1_{x^1}(X) - \mu_{x^1}(Z;W)\} + \mu_{x^1}(z^1;W) - \mu_{x^1}(z^0;W). \tag{5}$$

Then, the following is a moment score function for θ_h:

$$m(\theta_0, \eta_h) \equiv \frac{1}{\Delta_X}(\theta_h - \theta_0)V_X; \tag{6}$$

where θ_h is given in Eq. (2) and θ_0 is an estimate of θ_h.

Next, we derive an influence function for the moment score function $m(\theta_0, \eta_h)$. We first define the following function: for a bounded function $f(Y) < \infty$, let

$$\gamma_{YX}[f(Y)] \equiv \mathbb{E}_P[\gamma(x;z^x;W)[f(Y)] - \gamma(x;z^{1-x};W)[f(Y)]]; \tag{7}$$

$$V_{YX}(f,g)[f(Y)] \equiv \frac{1_{z^x}(Z) - 1_{z^{1-x}}(Z)}{\pi_Z(W)}\{f(Y)1_x(X) - \gamma(x;Z;W)[f(Y)]\}$$
$$+ \gamma(x;z^x;W)[f(Y)] - \gamma(x;z^{1-x};W)[f(Y)]; \tag{8}$$

and

$$\phi(\cdot = f,g;\cdot;\cdot)[f(Y)] \equiv \frac{1}{\Delta_X}(V_{YX}(f,g)[f(Y)] - [f(Y)]V_X(f,g)); \tag{9}$$

where V_X is defined in Eq. (5). Then, the influence function for the expectation of the moment score function $m(\theta_0, \eta_h)$ in Eq. (6) is given as follows:

Lemma 1 (Influence function for $m(\theta_0, \eta_h)$). Let $m(\theta_0, \eta_h)$ be the score defined in Eq. (6). Then, the influence function for $\mathbb{E}_P[m(\theta_0, \eta_h)]$, denoted ψ_m, is given by

$$\psi_m(\cdot = f,g;\cdot;\cdot) \equiv \phi(\cdot;\cdot)[K_{h;y}(Y)] \tag{10}$$

where ϕ is in Eq. (9).

For any score function (e.g., m in Eq. (6)), its addition to the influence function of the expected score (e.g., ψ_m) is a Neyman orthogonal score[3] ([14, Thm.1], [13, Sec. 2.2.5]). Specifically,

Lemma 2 (Neyman orthogonal score for θ_h). Let $m(\theta_0, \eta_h)$ be the score function in Eq. (6), and $\psi_m(\cdot = f,g;\cdot;\eta_h)$ be the influence function for $\mathbb{E}_P[m(\theta_0, \eta_h)]$ given in Eq. (10). Then, a Neyman orthogonal score for θ_h is given as $\psi(\theta_0, \cdot = f,g) \equiv m(\theta_0, \eta_h) + \psi_m(\cdot;\cdot)$. Specifically,

$$\psi(\theta_0, \cdot = f,g) = \frac{1}{\Delta_X}(V_{YX}(f,g)[K_{h;y}(Y)] - \theta_0 V_X(f,g)). \tag{11}$$

Given the Neyman orthogonal score $\psi(\theta_0, \cdot)$, an estimate $\hat{\eta}_h$ satisfying $\mathbb{E}_D[\psi(\hat{\theta}_h; \hat{\eta} = \{\hat{\pi}, \hat{\mu}, \hat{\gamma}\})] = o_P(n^{-1/2})$ gives a DML estimator. Specifically, we propose the following kernel-smoothing based estimator for the LTE density, named `KLTE' (kernel-based estimator for LTE):

[3] A Neyman orthogonal score is a function ψ satisfying $\mathbb{E}_P[\psi(\cdot;\eta_0)] = 0$ and $\frac{\partial}{\partial\eta}\mathbb{E}_P[\psi(V;\cdot;\eta)]|_{\eta=\eta_0} = 0$, where η_0 denotes the true nuisances [13, Def.2.2]. In words, a score function that is not sensitive to local errors in nuisance models.

Definition 1 (KLTE estimator for θ_h). Let $\psi(\cdot;\theta,\eta) = \{\psi_\ell(\cdot;\theta,\eta) : \ell = 1,\dots,L\}$ be the Neyman orthogonal score for θ_h given in Eq. (11). Let $\{D, D^0\}$ denote the randomly split halves of the samples, where $|D| = |D^0| = n$. Let $\hat{\eta} = \{\hat{\gamma}, \hat{\alpha}, \hat{g}\}$ denote the estimates for the nuisances using D^0. Then, the KLTE estimator for $\theta_h(y)$ for all $y \in \mathcal{Y}$, denoted $\hat{\theta}_h(y)$, is given by

$$\hat{\theta}_h(y) \leftarrow \mathbb{E}_D\left[V_{YX}(\hat{\gamma},\hat{\alpha},\hat{g})[K_{h,y}(Y)]\right] = \mathbb{E}_D\left[V_X(\hat{\gamma},\hat{\alpha},\hat{g})\right], \tag{12}$$

where V_X and V_{YX} are given in Eqs. (5,8), respectively.

We will show that the KLTE is a DML estimator exhibiting debiasedness property. Detailed asymptotic properties are discussed next.

3.1 Asymptotic convergence

Now, we study the convergence rate of the estimator $\hat{\theta}_h(y)$. For any fixed $y \in \mathcal{Y}$, the error $\hat{\theta}_h(y) - \theta(y)$ will be analyzed in two folds: we will first analyze the error between the estimator in Eq. (12) and the smoothed estimand in Eq. (2) (i.e., $\hat{\theta}_h(y) - \theta_h(y)$), and then analyze the error between the smoothed estimand and the true estimand (i.e., $\theta_h(y) - \theta(y)$).

The following result gives the error analysis for $\hat{\theta}_h(y) - \theta_h(y)$:

Lemma 3 (Convergence rate of $\hat{\theta}_h$ to θ_h). For any fixed $y \in \mathcal{Y}$, suppose the estimators for nuisances are consistent; i.e., $\|\hat{\alpha} - \alpha\| = o_P(1)$ for $\alpha \in \{\gamma, \alpha, g\}$ for all (w,z,x). Suppose $h < 1$, and $nh^d \to \infty$ as $n \to \infty$. Then,

$$\hat{\theta}_h(y) - \theta_h(y) = O_P\left(1/\sqrt{nh^d} + R_2^k + 1/\sqrt{n}\right),$$

where

$$R_2^k \leftarrow \sum_z \left\{\|\hat{\gamma}_z - \gamma_z\|\,\|\hat{\alpha}_z - \alpha_z\| + \|\hat{\alpha}_z - \alpha_z\|\right\}; \tag{13}$$

where $\gamma_z \leftarrow \gamma_z(W)$, $\alpha_z \leftarrow \alpha_z(z;W)$ and $g_z \leftarrow g(x;z;W)[K_{h,y}(Y)]$.

The error analysis in Lemma. 3 implies the following:

Corollary 1 (Debiasedness property of $\hat{\theta}_h$ to θ_h). If all nuisances $\{\hat{\gamma},\hat{\alpha},\hat{g}\}$ for any given $(w;z;x;y)$ converge at rate $o\{nh^d\}^{-1/4}$, then the target estimator $\hat{\theta}_h(y)$ achieves $\sqrt{nh^d}$-rate convergence to θ_h.

We now analyze the gap between the smoothed estimand θ_h and the true estimand θ; i.e., $\theta_h - \theta$:

Lemma 4 ([66, Thm. 6.28]). The following holds:

$$\theta_h(y) - \theta(y) = B_y \leftarrow 0.5h^2\,\mu_2(K)\left(\partial^2/\partial y^2\right)\theta(y^0)|_{y^0=y} + O(h^2). \tag{14}$$

Combining the results of Lemma. (3,4), we have the following result:

Theorem 1 (Convergence rate of $\hat{\theta}_h$ to θ). For any fixed $y \in \mathcal{Y}$, suppose the estimators for nuisances are consistent; i.e., $\|\hat{\alpha} - \alpha\| = o_P(1)$ for $\alpha \in \{\gamma, \alpha, g\}$ for all $(w;z;x)$. Suppose $h < 1$, and $nh^d \to \infty$ as $n \to \infty$. Then

$$\hat{\theta}_h(y) - \theta(y) = O_P\left(1/\sqrt{nh^d} + R_2^k + 1/\sqrt{n}\right) + B_y; \tag{15}$$

where B_y is defined in Eq. (14), and R_2^k is defined in Eq. (13).

Thm. 1 implies that $\hat{\theta}_h(y)$ converges fast (see Corol. 1) to $\theta(y) + B_y$. A natural question is then how to choose the bandwidth h that minimizes the gap in Eq. (15). The following provides a guideline in choosing the bandwidth h:

Lemma 5 (Data-adaptive bandwidth selection). The bandwidth h that minimizes the error in Eq. (15) is $h = O(n^{-1/(d+4)})$. This choice of h satisfies the assumption in Lemma 3 (i.e., $nh^d \to \infty$).

Recall that Corol. 1 states the debiasedness property of $\hat{\psi}_h$ to ψ for any bandwidth h satisfying $nh^d \to \infty$. With the choice of h as in Lemma 5, $\hat{\psi}_h$ converges to ψ with the debiasedness property preserved.

Corollary 2 (Debiasedness property of $\hat{\psi}_h$ to ψ). Let $h = O(n^{-1/(d+4)})$. If nuisances $\{\hat{\eta}, \hat{\pi}, \hat{m}\}$ converge at $\{nh^d\}^{-1/4}$ rate for any (w, z, x, y), then the target estimator $\hat{\psi}_h(y)$ achieves $\sqrt{nh^d}$-rate convergence to ψ.

So far, we have analyzed the error $\hat{\psi}_h(y) - \psi(y)$ pointwise for the fixed $y \in \mathcal{Y}$. To analyze the difference between the two densities $\hat{p}_h(y)$ and $p(y)$ for all $y \in \mathcal{Y}$, we consider the following divergence function of two densities:

Definition 2 (f-Divergence D_f [20]). Let f denote a convex function with $f(1) = 0$. $D_f(p, q) \triangleq \int_{\mathcal{Y}} f(p(y), q(y))\, q(y)\, d[y]$, is a f-divergence function between two densities p and q.

f-divergence covers many well-known divergences. For example, D_f reduces to KL divergence with $f(p, q) = (p/q)\log(p/q)$. We will assume that the function $f(p, q)$ in D_f is differentiable w.r.t. p and q.

We now analyze the distance between $\hat{p}_h$ and p w.r.t. D_f. The following result provides an upper bound for D_f.

Lemma 6 (Upper bound of the divergence D_f). Suppose D_f is a f-divergence such that $f(p, q) = 0$ if $p = q$. Then,

$$D_f(p, \hat{p}_h) \leq \int_{\mathcal{Y}} w(y)\, |\hat{\psi}_h(y) - \psi(y)|\, d[y];$$

where $w(y) \triangleq |f_2'(\psi(y), \tilde{\psi}(y))|\, \hat{\pi}_h(y) + |f_2'(p(y), q(y))|$, $f_2'(p, q) \triangleq (\partial/\partial q) f(p, q)$, and $\tilde{\psi}_h(y) \triangleq t\,\hat{\psi}_h(y) + (1 - t)\,\psi(y)$ for some fixed $t \in [0, 1]$.

By invoking Thm. 1, we derive an upper bound for $D_f(p, \hat{p}_h)$ as follows:

Theorem 2 (Convergence rate of $\hat{p}_h$). Suppose the estimators for nuisances are consistent; i.e., $\|\hat{\eta} - \eta\| = o_P(1)$ for $\eta \in \{\pi, m\}$ for all (w, z, x, y). Suppose D_f is a f-divergence such that $f(p, q) = 0$ if $p = q$. Suppose $w(y)$ in Lemma 6 is finite. Then,

$$D_f(p, \hat{p}_h) \leq O_P\left(\sup_{y \in \mathcal{Y}} R_2^k + B_y\right) + 1/\sqrt{nh^d} + 1/\sqrt{n}; \tag{16}$$

where R_2^k is defined in Eq. (13) and B_y is defined in Eq. (14).

The following result asserts that the debiasedness property is exhibited by D_f w.r.t. ψ.

Corollary 3 (Debiasedness property of $\hat{p}_h$ w.r.t. D_f). Let $h = O(n^{-1/(d+4)})$. Suppose D_f satisfies $f(p, q) = 0$ if $p = q$. Suppose $w(y)$ in Lemma 6 is finite. If nuisances $\{\hat{\eta}, \hat{\pi}, \hat{m}\}$ converges at $\{nh^d\}^{-1/4}$ rate for any (w, z, x, y), then $D_f(p, \hat{p}_h)$ converges to 0 at $\sqrt{nh^d}$-rate.

4 Model-based approach

In this section, we develop a model-based approach for estimating the LTE density $p(y) = p(y_x|C)$. We will approximate p with a class of distributions or a density model $\mathcal{G} = \{g(y; \theta) : \theta \in \mathbb{R}^b\}$ where $g(y; \theta) \in \mathcal{G}$ is differentiable w.r.t. θ. Example density models include exponential family (e.g., Gaussian distribution), mixture of Gaussians, or more generally, mixture of exponential families. The choice of the density model may depend on domain knowledge. Alternatively, one may choose among a set of candidate density families using separate validation data or applying cross-validation. We adapt the model-based approach developed in [39] for estimating the causal density under the no unmeasured confounders assumption.

Given a density model $\mathcal{G}$, the best approximation for $p(y)$ is defined as $g(y; \theta_0) \in \mathcal{G}$ that achieves the minimum f-divergence to p:

$$\theta_0 \triangleq \arg\min_{\theta \in \mathbb{R}^b} D_f(p(y), g(y; \theta)); \tag{17}$$

where D_f is the f-divergence defined in Def. 2. Our goal is estimating θ_0.

Consider $m(\theta; \eta) := (\partial/\partial\theta) D_f(\pi(\theta), g(\theta))$. Definition of θ_0 given in Eq. (17) implies that $m(\theta_0; \eta) = 0$ at $\theta = \theta_0$. We note that $m(\theta; \eta)$ serves as a moment score function. The closed-form expression of the score is given by [39]:

$$m(\theta; \eta) := \int_{\mathcal{Y}} g_0'(y; \theta) \{ f_2'(\pi(\theta), g(\theta)) g(\theta) + f_1'(\pi(\theta), g(\theta)) \} d[y], \tag{18}$$

where $g_0'(y; \theta) = (\partial/\partial\theta) g(\theta)$ and $f_2'(p, q) := (\partial/\partial q) f(p, q)$.

To construct a DML estimator based on the score function $m(\theta; \eta)$, we first derive an influence function for the score:

Lemma 7 (Influence Function for $m(\theta; \eta)$). *An influence function for $m(\theta; \eta)$ in Eq. (18), denoted ψ_m, is given by*

$$\psi_m(\theta; \eta, g) := \psi(\theta; \eta) [R_f(Y; \theta)], \tag{19}$$

where $\psi(\theta; \eta)[\cdot]$ is defined in Eq. (9), and

$$R_f(Y; \theta) := g_0'(Y; \theta) \{ f_{21}''(\pi(Y), g(Y; \theta)) g(Y; \theta) + f_1'(\pi(Y), g(Y; \theta)) \},$$

where $g_0'(y; \theta) := (\partial/\partial\theta) g(y; \theta)$, $f_1'(p, q) := (\partial/\partial p) f(p, q)$ and $f_{21}''(p, q) := (\partial/\partial q) f_2'(p, q)$.

We derive a Neyman orthogonal score based on the moment score $m(\theta; \eta)$ and its influence function $\psi_m(\theta; \eta)$:

Lemma 8 (Neyman orthogonal score for θ). *A Neyman orthogonal score for estimating θ, denoted $\psi'(\theta, \eta(\cdot) = \{\pi(\cdot), g(\cdot)\})$, is given by*

$$\psi'(\theta, \eta(\cdot) = \{\pi(\cdot), g(\cdot)\}) := m(\theta; \pi) + \psi_m(\theta; \eta), \tag{20}$$

where $\psi_m(\theta; \eta)$ is defined in Eq. (19).

Given the orthogonal score $\psi'(\theta, \eta(\cdot))$ in Eq. (20), we propose the following estimator for θ, named `MLTE' (model-based estimator for LTE):

Definition 3 (MLTE estimator for θ). Let $\psi'(\theta, \eta(\cdot) = \{\pi(\cdot), g(\cdot)\})$ be the Neyman orthogonal score for θ given in Eq. (20). Let $\{D, D^0\}$ denote the randomly split halves of the samples, where $|D| = |D^0| = n$. Let $\hat{\eta} = \{\hat{\pi}, \hat{g}\}$ denote the estimators for the nuisances using D^0. Then, the MLTE estimator for θ, denoted $\hat{\theta}$, is given as a solution satisfying $\mathbb{E}_D[\psi'(\hat{\theta}; \hat{\eta})] = o_P(n^{-1/2})$.

To illustrate, we exemplify Eq. (18) and Lemma (7, 8) for the case where D_f is a KL-divergence and $g(y; \theta) = \{\theta, \sigma^2\}$ is a normal distribution. First, $m(\theta; \eta) = \{m_\mu(\theta; \eta), m_{\sigma^2}(\theta; \eta)\}$, where $m_\mu(\theta; \eta) = (1/\sigma^2)(\theta - \mathbb{E}[Y])$ and $m_{\sigma^2}(\theta; \eta) = (0.5/\sigma^4)\{\sigma^2 - \mathbb{E}[(Y - \theta)^2]\}$. We note that $\hat{\theta}_m = \hat{\mathbb{E}}[Y]$ and $\hat{\sigma}^2_m = \hat{\mathbb{E}}[(Y - \hat{\theta})^2]$ are estimators for $\theta_0 = \{\theta_0, \sigma_0^2\}$ for the score $m(\theta; \eta)$.

Also, $R_f(Y; \theta) := (\partial/\partial\theta) \log(g(Y; \theta)) = \{R_\mu(Y; \theta), R_{\sigma^2}(Y; \theta)\}$, where $R_\mu(Y; \theta) := (Y - \theta)/\sigma^2$ and $R_{\sigma^2}(Y; \theta) := 0.5\{\sigma^2 - (Y - \theta)^2\}/\sigma^4$. Then, the Neyman orthogonal score is given as $\psi(\theta; \sigma^2; \eta) = (1/\sigma^2)\{\theta - \mathbb{E}[Y] - \psi(\theta; \eta)[Y]\}$ and $\psi'(\sigma^2; \theta; \eta) = (0.5/\sigma^4)\{\sigma^2 - \mathbb{E}[(Y - \theta)^2] - \psi(\theta; \eta)[(Y - \theta)^2]\}$. Finally, solutions for $\psi'(\theta; \sigma^2; \eta)$ and $\psi'(\sigma^2; \theta; \eta)$ are given by $\{\hat{\theta}, \hat{\sigma}^2\}$, where, for $\psi(\theta; \eta)[\cdot]$ in Eq. (9), $\hat{\theta} = \hat{\mathbb{E}}[Y] + \mathbb{E}_D[\psi(\hat{\theta}; \hat{\eta})[Y]]$ and $\hat{\sigma}^2 = \hat{\mathbb{E}}[(Y - \hat{\theta})^2] + \mathbb{E}_D[\psi(\hat{\theta}; \hat{\eta})[(Y - \hat{\theta})^2]]$.

The MLTE estimator in Def. 3 is consistent provided that nuisances estimators are consistent [14, Thm.4]. Such $\hat{\theta}$ is known to achieve debiasedness [13], since it is a DML estimator. Specifically,

Theorem 3 (Convergence rate of $\hat{\theta}$). Let $\psi'(\theta, \eta(\cdot) = \{\pi(\cdot), g(\cdot)\})$ be given in Eq. (20). Let $\psi_m(\theta; \eta)$ be given in Eq. (19). Let $\theta_0, \sigma_0^2, \pi_0$ denote the true parameters. Let $\hat{\theta}$ be the MLTE estimator for θ defined in Def. 3. Suppose (1) $R_f(y; \theta)$ is bounded and $R_f^0(y; \theta) := (\partial/\partial\theta) R_f(y; \theta) < 1$; (2) There exists a function $H(y) < 1$ s.t. $\sup_\theta \max\{R_f(y; \theta), R_f^0(y; \theta)\} = O(H(y))$; (3) $\{\psi'(\theta, \eta(\theta))\}$ is Donsker[4] w.r.t. θ for

[4] A function class where complexities are restricted. See Def. S.1 in the Appendix for the definition. Donsker class include Sobolev, Bounded monotone, Lipschitz class, etc.

the xed ; (3) The estimators are consistent: $\hat{\ }_0 = o_P(1)$ and $k\ \hat{\ }k = o_P(1)$ for $2 f_z(w); \ _x(z;w); \ (x;z;w)[H(Y)]g$ for all $(w;z;x;y)$; and (4) $E_P[' (;(;))]$ is differentiable w.r.t. at $= _0$ with non-singular matrix $M(_0;(;))\ (@=@ j_{= _0} E_P[' (;(;))]$ for all $(;)$, where $M(_0;(\hat{\ };\hat{\ })) \to^P M\ \ M(_0;(_0; _0))$. Then,

$$b\ _0 = \ M^{-1} E_D[\ _m(_0;(_0; _0))] + o_P(n^{-1=2}) + O_P(R_2^m);$$

where

$$R_2^m = \sum_z \Big\{ k\hat{\ }_z\ \ _z k\ \hat{\ }_z\ \ _z + \hat{\ }_z\ \ _z + \hat{\ }_z\ \ _z{}^2 + \ _z\ \hat{\ }_z{}^2 + \hat{\ }_z\ \ _z\ \ _z\ \hat{\ }_z ;$$

where $_z\ \ _z(W)$, $_z\ \ _x(z;W)$, and $_z\ \ (x;z;W)[H(Y)]$.

Corollary 4 (Debiasedness property for $\hat{\ }$). If nuisances $f\hat{\ }; \hat{\ }; \hat{\ }g$ converges at $n^{-1=4}$ rate, then the target estimator $\hat{\ }$ converges to $_0$ at $\sqrt{n}$-rate.

For the above example where D_f is the KL divergence and $g(y;)$ is a normal distribution, $H(Y) = Y$ for $R_f(y; ;)$, and $H(Y) = Y^2$ for $R_f(y; {}^2;)$.

5 Empirical applications

In this section, we apply the proposed methods to synthetic and real datasets. For the kernel-smoothing based approach, we compare KLTE with a baseline plug-in estimator (`kernel-smoothing'), where estimates of nuisances $\hat{\ }= f\hat{\ }; \hat{\ }; \hat{\ }g$ are plugged in the estimand Eq. (2). We use the Gaussian kernel. The bandwidth is set to $= 0:5n^{-1=5}$. In estimating the density, we choose 200 equi-spaced points $f y_{(i)}g_{i=1}^{200}$ in Y and evaluate both estimators at $fh; y_{(i)}$ for $i = 1; \ ;200$. For the model-based approach, we compare MLTE (e.g., $\hat{\ }^2$) with a moment-score-based estimator (called `moment'), de ned as $\hat{\ }_m$ satisfying $m(\hat{\ }_m; \hat{\ }) = o_P(n^{-1=2})$ (e.g., $f\hat{\ }_m; \hat{\ }_m^2 g$). We use KL divergence for D_f and the normal distribution for $g(y;)$. For both approaches, nuisances are estimated through a gradient boosting model XGBoost [11], which is known to be exible.

5.1 Synthetic dataset

We applied the proposed estimators to estimate the $p(y_x jC)$ where the true densities are given as in the 4th plot in Fig. 2. As shown in the ground-truth in Fig. 3a, true densities $p(y_x jC); p(y_{x^1} jC)$ are given as a mixture of four Gaussians. Estimated densities for Moment and MLTE are given in Fig. (3b, 3c). We note that model-based approaches fail to capture important characteristics (such as the number of modes) of the true density (`ground-truth' in Fig. 3a) because the assumed density class is misspeci ed. The `kernel-smoothing' (Fig. 3d) captures only one of the modes having the highest densities, and this leads to misinterpretation of the true densities. KLTE (Fig. 3e) is able to capture the number, location, and scales of modes correctly.

(a) Ground-truth (b) Moment (c) MLTE (d) Kernel-smoothing (e) KLTE

Figure 3: LTE estimation with a synthetic dataset. The ground-truth density is in (a). Red and Green for x^0 and x^1, respectively.

5.2 Application to 401(k) data

We applied the proposed estimators (KLTE and MLTE) on 401(k) data, where the data generating processes corroborate with Fig. 1. Monotonicity assumption holds naturally, since ineligible units ($Z = 0$) cannot participate ($X = 1$) in 401(k). In our analysis, we used the dataset introduced

by [2] containing 9275 individuals, which has been studied in [2, 17, 5, 47, 58, 64], to cite a few. Model-based approaches (Moment in Fig. 4a and MLTE in Fig. 4b) and kernel-smoothing based approaches (kernel-smoothing in Fig. 4c and KLTE in Fig. 4d) are implemented to analyze the data.

The model-based (Fig. (4a,4b)) and kernel-smoothing based (Fig. (4c,4d)) estimates both capture important characteristics of the distribution, such as mode, location, and scale parameters. The results of proposed estimators (MLTE and KLTE in Fig. (4b,4d)) are consistent with findings from previous analyses [2, 17, 5, 58]: The effects of the 401(k) participation (i.e., $X_i = 1$) on net financial assets are positive over the whole range of asset distributions. To connect to CDF method, we provide in Fig. 4e the CDF estimate induced by KLTE density estimation (Fig. 4a). We note that the CDF in Fig. 4e captures the nonconstant impact trend of the 401(k) participation on the net financial assets, which has been also described in the previous analyses [2, 17, 5, 58].

(a) Moment (b) MLTE (c) Kernel-smoothing (d) KLTE (e) KLTE – CDF

Figure 4: LTE of 401(k) participation $X(x)$ on net financial asset $Y(y)$. Red and Green for x^0 and x^1, respectively.

6 Conclusion

In this paper, we develop kernel-smoothing-based and model-based approaches for estimating the LTE density in the presence of instruments. For each approach, we give Neyman orthogonal scores (Lemma (2,8)) and constructed corresponding DML estimators (KLTE in Def. 1 and MLTE in Def. 3), that exhibit debiasedness property (Corol. (3, 4)). We demonstrated our work through synthetic and real datasets. The performance of model-based estimators depends critically on the choice of the density class. Kernel-based estimators do not have to make assumptions about the true density class but will suffer from the curse of dimensionality. This work is limited to settings where the monotonicity assumption holds, i.e., there are no defiers. One could perform sensitivity analyses on the impact of potential defiers to the estimates as conducted in [65, 36].

Acknowledgements

We thank the reviewers for their feedback helping to improve this manuscript. Elias Bareinboim and Yonghan Jung were partially supported by grants from NSF IIS-1750807 (CAREER). Jin Tian was partially supported by ONR grant N000141712140.

References

[1] A. Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. Journal of the American statistical Association, 97(457):284–292, 2002.

[2] A. Abadie. Semiparametric instrumental variable estimation of treatment response models. Journal of econometrics, 113(2):231–263, 2003.

[3] J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. Journal of the American statistical Association, 91(434):444–455, 1996.

[4] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. On pearl's hierarchy and the foundations of causal inference. Technical Report R-60. Causal Artificial Intelligence Laboratory, Columbia University, 2020.

[5] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2):608–650, 2014.

[6] A. F. Bibaut and M. J. van der Laan. Data-adaptive smoothing for optimal-rate estimation of possibly non-regular parameters. arXiv preprint arXiv:1706.07406, 2017.

[7] P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. Efficient and adaptive estimation for semiparametric models, volume 4. Johns Hopkins University Press Baltimore, 1993.

[8] C. Brito. Instrumental sets. In R. Dechter, H. Geffner, and J. Y. Halpern, editors, Heuristics, Probability and Causality. A Tribute to Judea Pearl. College Publications, 2010.

[9] C. Brito and J. Pearl. Generalized instrumental variables. In Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, pages 85–93, 2002.

[10] M. D. Cattaneo, M. Jansson, and X. Ma. Local regression distribution estimators. Journal of Econometrics, 2021.

[11] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, 2016.

[12] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double/debiased/neyman machine learning of treatment effects. American Economic Review, 107(5):261–65, 2017.

[13] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. The Econometrics Journal, 21(1), 2018.

[14] V. Chernozhukov, J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins. Locally robust semiparametric estimation. arXiv preprint arXiv:1608.00033, 2016.

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

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

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

[18] V. Chernozhukov, C. Hansen, and K. Wuthrich. Instrumental variable quantile regression. arXiv preprint arXiv:2009.00436, 2020.

[19] K. Colangelo and Y.-Y. Lee. Double debiased machine learning nonparametric inference with continuous treatments. arXiv preprint arXiv:2004.03036, 2020.

[20] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. studia scientiarum Mathematicarum Hungarica, 2:229–318, 1967.

[21] I. Díaz and M. J. van der Laan. Targeted data adaptive estimation of the causal dose–response curve. Journal of Causal Inference, 1(2):171–192, 2013.

[22] J. DiNardo, N. M. Fortin, and T. Lemieux. Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. Econometrica: Journal of the Econometric Society, pages 1001–1044, 1996.

[23] M. Frölich. Nonparametric iv estimation of local average treatment effects with covariates. Journal of Econometrics, 139(1):35–75, 2007.

[24] M. Frölich and B. Melly. Unconditional quantile treatment effects under endogeneity. Journal of Business & Economic Statistics, 31(3):346–357, 2013.

[25] S. Ghosal et al. Convergence rates for density estimation with bernstein polynomials. The Annals of Statistics, 29(5):1264–1280, 2001.

[26] E. Giné, R. Nickl, et al. Con dence bands in density estimation. The Annals of Statistics, 38(2):1122–1170, 2010.

[27] L. Györ , M. Kohler, A. Krzyzak, and H. Walk. A distribution-free theory of nonparametric regression. Springer Science & Business Media, 2006.

[28] J. Hahn. On the role of the propensity score in ef cient semiparametric estimation of average treatment effects. Econometrica, pages 315–331, 1998.

[29] J. J. Heckman. Randomization as an instrumental variable, 1995.

[30] Y.-C. Hsu, T.-C. Lai, and R. P. Lieli. Estimation and inference for distribution functions and quantile functions in endogenous treatment effect models. Technical report, Institute of Economics, Academia Sinica, Taipei, Taiwan, 2015.

[31] G. W. Imbens and J. D. Angrist. Identi cation and estimation of local average treatment effects. Econometrica, 62(2):467–475, 1994.

[32] G. W. Imbens and D. B. Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.

[33] Y. Jung, J. Tian, and E. Bareinboim. Estimating identi able causal effects on markov equivalence class through double machine learning. In Proceedings of the 38th International Conference on Machine Learning (ICML), 2021.

[34] Y. Jung, J. Tian, and E. Bareinboim. Estimating identi able causal effects through double machine learning. In Proceedings of the 35th AAAI Conference on Arti cial Intelligence, 2021.

[35] N. Kallus and M. Uehara. Doubly robust off-policy value and gradient estimation for deterministic policies. Advances in Neural Information Processing Systems, 33, 2020.

[36] H. Kang, Y. Jiang, Q. Zhao, and D. S. Small. Ivmodel: an r package for inference and sensitivity analysis of instrumental variables models with one endogenous variable. Observational Studies, 7(2):1–24, 2021.

[37] E. H. Kennedy. Semiparametric theory and empirical processes in causal inference. In Statistical causal inferences and their applications in public health research, pages 141–167. Springer, 2016.

[38] E. H. Kennedy, S. Balakrishnan, M. G'Sell, et al. Sharp instruments for classifying compliers and generalizing causal effects. Annals of Statistics, 48(4):2008–2030, 2020.

[39] E. H. Kennedy, S. Balakrishnan, and L. Wasserman. Semiparametric counterfactual density estimation. arXiv preprint arXiv:2102.12034, 2021.

[40] E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81(1):121–143, 2019.

[41] E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 79(4):1229, 2017.

[42] K. Kim, J. Kim, and E. H. Kennedy. Causal effects based on distributional distances. arXiv preprint arXiv:1806.02935, 2018.

[43] C. A. Klaassen. Consistent estimation of the in uence function of locally asymptotically linear estimators. The Annals of Statistics, pages 1548–1562, 1987.

[44] T. Leonard. Density estimation, stochastic processes and prior information. Journal of the Royal Statistical Society: Series B (Methodological), 40(2):113–132, 1978.

[45] B. Melly and K. Wüthrich. Local quantile treatment effects. 2016.

[46] R. Neugebauer and M. van der Laan. Nonparametric causal effects based on marginal structural models. Journal of Statistical Planning and Inference, 137(2):419–434, 2007.

[47] E. L. Ogburn, A. Rotnitzky, and J. M. Robins. Doubly robust estimation of the local average treatment effect curve. Journal of the Royal Statistical Society. Series B, Statistical methodology, 77(2):373, 2015.

[48] R. Okui, D. S. Small, Z. Tan, and J. M. Robins. Doubly robust instrumental variable regression. Statistica Sinica, pages 173–205, 2012.

[49] J. Pearl. Causal diagrams for empirical research. Biometrika, 82(4):669–710, 1995.

[50] J. Pearl. Causality: Models, Reasoning, and Inference. Cambridge University Press, New York, 2000. 2nd edition, 2009.

[51] J. Pearl. Parameter identification: A new perspective. Technical Report R-276, 2001.

[52] J. Robins, L. Li, E. Tchetgen, A. van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. Probability and statistics: essays in honor of David A. Freedman, pages 335–421. Institute of Mathematical Statistics, 2008.

[53] J. Robins and A. Rotnitzky. Comment on "inference for semiparametric models: Some questions and an answer," by pj bickel and j. kwon. Statistica Sinica, 11:920–936, 2001.

[54] J. M. Robins and Y. Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. Statistics in medicine, 16(3):285–319, 1997.

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

[56] K. J. Rothman and S. Greenland. Causation and causal inference in epidemiology. American journal of public health, 95(S1):S144–S150, 2005.

[57] D. B. Rubin. Bayesian inference for causal effects: The role of randomization. The Annals of statistics, pages 34–58, 1978.

[58] R. Singh and L. Sun. De-biased machine learning for compliers. arXiv preprint arXiv:1909.05244, 2019.

[59] E. Smucler, A. Rotnitzky, and J. M. Robins. A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. arXiv preprint arXiv:1904.03737, 2019.

[60] Z. Tan. Regression and weighting methods for causal inference using instrumental variables. Journal of the American Statistical Association, 101(476):1607–1618, 2006.

[61] A. B. Tsybakov. Introduction to nonparametric estimation. Springer Science & Business Media, 2008.

[62] S. D. Uysal. Doubly robust iv estimation of the local average treatment effects, 2011.

[63] A. W. Van der Vaart. Asymptotic statistics, volume 3. Cambridge university press, 2000.

[64] L. Wang, Y. Zhang, T. S. Richardson, and J. M. Robins. Estimation of local treatment effects under the binary instrumental variable model. Biometrika, 2021.

[65] X. Wang, Y. Jiang, N. R. Zhang, and D. S. Small. Sensitivity analysis and power for instrumental variable studies. Biometrics, 74(4):1150–1160, 2018.

[66] L. Wasserman. All of nonparametric statistics. Springer Science & Business Media, 2006.

[67] J. M. Wooldridge. Econometric analysis of cross section and panel data. MIT press, 2010.

[68] P. G. Wright. Tariff on animal and vegetable oils. Macmillan Company, New York, 1928.

[69] K. Wüthrich. A comparison of two quantile models with endogeneity. Journal of Business & Economic Statistics, 38(2):443–456, 2020.

[70] W. Zheng and M. J. van der Laan. Cross-validated targeted minimum-loss-based estimation. In Targeted Learning, pages 459–474. Springer, 2011.

Appendix – Double Machine Learning Density Estimation for Local Treatment Effects with Instruments

A IV Settings and LTE

In this work, we consider the IV setting represented by the causal graph $\mathcal{G}$ in Fig. 1. It is common in the literature to define IV assumptions in terms of conditional independences among counterfactuals [2, 60, 47, 64], as given in the following:

Assumption A.1 (IV assumptions).

1. Exclusion restriction: $Y_{x;z} = Y_x$ almost surely for all z, x.
2. Independence: $Z \perp\!\!\!\perp (Y_x, X_z) | W$ for all z, x.
3. Instruments relevance: $P(X_{Z=1} = 1|W) \neq P(X_{Z=0} = 1|W)$ almost surely.

We show that the causal graph in Fig. 1 captures the set of IV assumptions in Assumption A.1.

Lemma A.1. *The causal graph $\mathcal{G}$ in Fig. 1 satisfies the set of IV assumptions in Assumption A.1.*

Proof. We will show the first item. We have $Y_{x;z} = Y_{x;z;W_x} = Y_{x;W_x} = Y_x$, where the first equality is due to the composition property [50, Property 1 (pp. 229)], the second due to exclusion restrictions [50, Eq.(7.25)], and the third by composition.

We will show the second. We have $(Z \perp\!\!\!\perp W; X_{z;w}, Y_{x;w})$ by independence restrictions [50, Eq.(7.26)]. Then by the weak union graphoid axiom (Rest. [50, pp.11]), $(Z \perp\!\!\!\perp X_{z;w}, Y_{x;w} | W)$, which leads to $(Z \perp\!\!\!\perp (Y_x, X_z) | W)$ by composition.

We will show the third. By $(Z \perp\!\!\!\perp X_z | W)$, $P(x_z|w) = P(x_z|w, z) = P(x|w, z)$, where the second equality is by composition. The third assumption is reflected by X that is not independent of Z given W in $\mathcal{G}$. □

Definition A.1 (Local treatment effect (LTE) density). *The local treatment effect (LTE) density is the density of outcome Y under treatment $X = x$ among compliers (i.e. $X_{Z=1} = 1$ and $X_{Z=0} = 0$) denoted by $p(y_x | X_{Z=1} = 1, X_{Z=0} = 0)$. We will use $C = (X_{Z=1} = 1 \wedge X_{Z=0} = 0)$ to denote the event that a unit is a complier and write the LTE density as $p(y_x|C)$.*

The LTE density $p(y_x|C)$ is known to be identifiable under monotonicity in the IV setting [2, 31]. In the notations of this paper, we present the identification results as follows, where for a given constant a and a variable X, x^a denotes the event $X = a$.

Lemma A.2. *In the causal graph $\mathcal{G}$ in Fig. 1, $p(y_x|w; C)$ is identifiable under monotonicity and is given by*

$$p(y_x|w; C) = \frac{p(y|x; z^x; w)P(x|z^x; w) - p(y|w; x; z^{1-x})P(x|z^{1-x}; w)}{P(x^1|z^1; w) - P(x^1|z^0; w)}.$$

Theorem A.1. *In the causal graph $\mathcal{G}$ in Fig. 1, the LTE density $p(y_x|C)$ is identifiable under monotonicity and is given by*

$$p(y_x|C) = \frac{\int_W [p(y|x; z^x; w)P(x|z^x; w) - p(y|w; x; z^{1-x})P(x|z^{1-x}; w)]P(w)\, d[w]}{\int_W [P(x^1|z^1; w) - P(x^1|z^0; w)]P(w)\, d[w]}$$
$$= \frac{E_P\left[p(y|x; z^x; W)P(x|z^x; W) - p(y|x; z^{1-x}; W)P(x|z^{1-x}; W)\right]}{E_P\left[P(x^1|z^1; W) - P(x^1|z^0; W)\right]}.$$

B Proofs

Notations We will use $P \rightsquigarrow P(1 + \epsilon g)$, where g is a mean zero bounded random function, to denote a parametric submodel for the probability measure P. Also, we note that the causal effect

$[f(Y)]$ in Eq. (3) can be written as $\Psi_{YX}[f(Y)] = \Psi_X$, where Ψ_X and $\Psi_{YX}[f(Y)]$ are defined in Eqs. (4,7).

We provide a formal definition of a function class called Donsker class, which is used throughout the proof.

Definition S.1 (Donsker Class [63, page. 269]). Let $G_n(f) \equiv \sqrt{n}(1/n)\sum_{i=1}^{n} f(v_{(i)}) - E_P[f(V)]$ denote the empirical process evaluated at a measurable function f. A class of measurable functions $\mathcal{F}$ is called (P-)Donsker class if the sequence of processes $\{G_n(f) : f \in \mathcal{F}\}$ converges in distribution to a limit process G in the space $l^1(\mathcal{F})$, where G is the process such that, for all $\epsilon > 0$, there is a compact set S such that $P(G \in S) > 1 - \epsilon$.

Lemma S.1 ([63, Thm.5.31], [39, Lemma 3]). Let $\psi(V; \theta, \eta)$ denote a vector estimating function for target parameter $\theta \in \mathbb{R}^p$ and nuisance functions $\eta \in H$ for some function space H. Suppose $E_P[\psi(V; \theta_0, \eta_0)] = 0$ (where θ_0, η_0 denote true parameters) and define the estimator $\hat{\theta}$ as a solution to $E_D[\psi(V; \hat{\theta}, \hat{\eta})] = o_P(n^{-1/2})$, where $\hat{\eta}$ is estimated on a separate independent sample. Assume

1. $\{\psi(V; \theta, \eta) : \theta \in \mathbb{R}^p\}$ is Donsker for any fixed η.
2. $\hat{\eta} - \eta_0 = o_P(1)$ and $\|\eta - \eta_0\|_2 = o_P(1)$.
3. The map $\eta \mapsto E_P[\psi(V; \theta, \eta)]$ is differentiable at η_0 uniformly in θ, with non-singular matrix $M(\theta, \eta) \equiv (\partial/\partial\theta')E_P[\psi(V; \theta, \eta)]$, where $M(\theta, \eta) \xrightarrow{P} M \equiv M(\theta_0, \eta_0)$.

Then,

$$\hat{\theta} - \theta_0 = -M^{-1}E_D[\psi(V; \theta_0, \eta_0)] - M^{-1}E_P[\psi(V; \theta_0, \hat{\eta})] + o_P(n^{-1/2}).$$

B.1 Proofs for Sec. 3

Lemma S.2 ([28, Proof of Thm. 1]). For a target estimand $E_P\{E_P[f(Y)|x^1, W] - E_P[f(Y)|x^0, W]\}$ for binary $X \in \{0, 1\}$ and $f(\cdot) < \infty$, an influence function is given by

$$\frac{1_{x^1}(X) - 1_{x^0}(X)}{P(X|W)}(f(Y) - E_P[f(Y)|X, W]) + E_P[f(Y)|x^1, W] - E_P[f(Y)|x^0, W].$$

Lemma S.3. An influence function for $[f(Y)]$ for $f(Y) < \infty$ is given by the mapping function in Eq. (9), which is

$$(\psi = \{\psi_1, \psi_2, \psi_3\})[f(Y)] \equiv \frac{1}{\psi_3}(V_{YX}(\psi_1, \psi_2)[f(Y)] - [f(Y)]V_X(\psi_1, \psi_2)).$$

Proof. We note that the estimand is given as $[f(Y)] = \Psi_{YX}[f(Y)] - \Psi_X$, where Ψ_X and $\Psi_{YX}[f(Y)]$ are defined in Eqs. (4,7).

By Lemma S.2, influence functions corresponding to Ψ_X and $\Psi_{YX}[f(Y)]$, denoted ψ_X and $\psi_{YX}[f(Y)]$ respectively, are given as

$$\psi_X \equiv \frac{1_{z^1}(Z) - 1_{z^0}(Z)}{\pi_Z(W)}(1_{x^1}(X) - \pi_{x_1}(Z; W)) + \pi_{x_1}(z^1; W) - \pi_{x_1}(z^0; W) - \Psi_X \tag{B.1}$$

$$\psi_{YX}[f(Y)] \equiv \frac{1_{z^x}(Z) - 1_{z^{1-x}}(Z)}{\pi_Z(W)}(f(Y)1_x(X) - \pi(x; Z; W)[f(Y)]) + \pi(x; z^x; W)[f(Y)] - \pi(x; z^{1-x}; W)[f(Y)] - \Psi_{YX}[f(Y)]. \tag{B.2}$$

Then, by applying the chain rule for the Gateaux derivative (since the influence function is a Gateaux derivative), an influence function for $[f(Y)] = {}_{YX}[f(Y)]=\ {}_X$ is given as

$$\begin{aligned}
&\frac{1}{X}\left({}_{YX}[f(Y)] \quad [f(Y)]\ {}_X \right) \\
&= \frac{1}{X}\ V_{YX}[f(Y)] \quad {}_{YX}[f(Y)] \quad [f(Y)]\ V_X \quad {}_X \\
&= \frac{1}{X}\ \{V_{YX}[f(Y)] \quad [f(Y)]V_X\} \quad [f(Y)] + \quad [f(Y)] \\
&= \frac{1}{X}\left(V_{YX}[f(Y)] \quad [f(Y)]V_X\right):
\end{aligned}$$

□

Lemma B.1 (Restated Lemma 1) Let $m(\ {}^0;\ {}_h)$ be the score defined in Eq. (6). Then, an influence function for $E_P[m(\ {}^0;\ {}_h)]$, denoted ${}_m$, is given by

$${}_m(\ = f\ ;\ ;\ g;\ ) \quad (\ ;\ )[K_{h;y}(Y)] \tag{B.3}$$

where $$ is given as

$$(\ = f\ ;\ ;\ g;\ )[f(Y)] \quad \frac{1}{X}\left(V_{YX}(f\ ;\ g)[f(Y)] \quad [f(Y)]V_X(f\ ;\ g)\right)$$

Proof. Let ${}_X$ denote the influence function corresponding to $\hat{}_X$, given in Eq. (B.1). This implies that $E_P[V_X] = {}_X$. Then, equipped with the true nuisance for V_X,

$$E_P[m(\ {}^0;\ {}_h)] = E_P\left[\frac{1}{X}(\ {}_h \quad {}^0)V_X\right] = \frac{1}{X}E_P[(\ {}_h \quad {}^0)V_X] = \ {}_h \quad {}^0.$$

Then, the influence function for $E_P[m(\ {}^0;\ {}_h)]$ coincides with the influence function for ${}_h$, which is given by Eq. (B.3) based on Lemma S.3. □

Lemma B.2 (Restated Lemma 2) Let $m(\ {}^0;\ {}_h)$ be the score function in Eq. (6), and ${}_m(\ = f\ ;\ ;\ g;\ {}_h)$ be the influence function for $E_P[m(\ {}^0;\ {}_h)]$ given in Eq. (10). Then, a Neyman orthogonal score for ${}_h$ is given as $'(\ {}^0;\ = f\ ;\ ;\ g) \quad m(\ {}^0;\ {}_h) + {}_m(\ ;\ )$; Specifically,

$$'(\ {}^0;\ = f\ ;\ ;\ g) = \frac{1}{X}\left(V_{YX}(f\ ;\ g)[K_{h;y}(Y)] \quad {}^0V_X(f\ ;\ g)\right): \tag{B.4}$$

Proof. For a score function for , denoted $m(\ )$, and the influence function of $E_P[m(\ )]$, denoted ${}_m(\ )$, a Neyman orthogonal score for is given as $m + {}_m$ [14, Thm. 1]. Applying this, $m(\ {}^0;\ {}_h) + {}_m(\ ;\ {}_h)$ is a Neyman orthogonal score. Specifically,

$$\begin{aligned}
&'(\ {}^0;\ = f\ ;\ ;\ g) \\
&= m(\ {}^0;\ {}_h) + {}_m(\ ;\ {}_h) \\
&= \frac{1}{X}(\ [K_{h;y}(Y)] \quad {}^0)V_X + \frac{1}{X}\left(V_{YX}(f\ ;\ g)[K_{h;y}(Y)] \quad [K_{h;y}(Y)]V_X(f\ ;\ g)\right) \\
&= \frac{1}{X}\left(V_{YX}(\ = f\ ;\ g)[K_{h;y}(Y)] \quad {}^0V_X(f\ ;\ g)\right):
\end{aligned}$$

□

Lemma B.3 (Restated Lemma 3) For any fixed $y \in \mathcal{Y}$, suppose the estimators for nuisances are consistent; i.e., $\|\hat{} \ - \ \|_2 = o_P(1)$ for $\in \{\ = f\ ;\ ;\ g\}$ for all $(w;z;x)$. Suppose $h < 1$, and $nh^d \to 1$ as $n \to 1$. Then,

$$\hat{}_h(y) \quad {}_h(y) = O_P\left(1=\sqrt{nh^d} + R_2^k + 1=\sqrt{n}\right);$$

where

$$R_2^k \quad \sum_z \left\{\|\hat{}_z \quad {}_z\|_2\|\hat{}_z \quad {}_z\| + \|\hat{}_z \quad {}_z\|_2\right\}; \tag{B.5}$$

where ${}_z \quad {}_z(W)$, ${}_z \quad {}_x(z;W)$ and ${}_z \quad (x;z;W)[K_{h;y}(Y)]$.

Proof. We note that the condition $nh^d \to 1$ means that $h = O(n^{-\gamma})$ for some $\gamma < 1/d$. $h < 1$ implies that such h is either constant or decreasing function over n. Combining, the condition implies $h = O(n^{-\gamma})$ for $\gamma \in [0, 1/d)$.

We recall that μ_X, μ_{YX} are defined in Eq. (4.7) and V_X, V_{YX} are defined in Eq. (5,8).

Now, we will prove this Lemma through the master result in Lemma S.1. The KLTE estimator $\hat{\tau}_h$ for Eq. (12) satisfies $E_D[\psi(\hat{\tau}_h, b)] = o_P(n^{-1/2})$, because

$$\begin{aligned}
E_D[\psi(\hat{\tau}_h, b)] &= \frac{1}{\mu_X(b)} E_D[V_{YX}(f_b, b_g)[K_{h,y}(Y)]] - \hat{\tau}_h E_D[V_X(f_b, b_g)] \\
&= \frac{1}{\mu_X(b)} \left(E_D[V_{YX}(f_b, b_g)[K_{h,y}(Y)]] - \frac{E_D[V_{YX}(f_b, b_g)[K_{h,y}(Y)]]}{E_D[V_X(f_b, b_g)]} E_D[V_X(f_b, b_g)] \right) \\
&= 0.
\end{aligned}$$

The Neyman orthogonal score function in Lemma 2 satisfies the assumptions in Lemma S.1, since ψ is a linear function of τ when nuisances are fixed. Also M in Lemma S.1 is given as 1, which can be witnessed by the following:

$$M(\eta_0, \cdot) = (\partial/\partial\eta) \Big|_{\eta_0} \frac{1}{\mu_X} E_P[f V_{YX} - \tau_0 V_X g] = \frac{1}{\mu_X} E_P[V_X] = 1,$$

and, with the true nuisances $M_0 = M(\eta_0, \cdot) = 1$ since $E_P[V_X] = \mu_X$.

Then, by the result of Lemma S.1,

$$\hat{\tau}_h - \tau_h = E_D[\psi_m(\tau_h, \cdot)] + E_P[\psi_m(\tau_h, \hat{\cdot})] + o_P(n^{-1/2}).$$

We will first study the convergence behavior of $E_D[\psi_m(\tau_h, \cdot)]$. We will show that $E_P[E_D[\psi_m(\tau_h, \cdot)]] = O(1/\sqrt{nh^d})$. Then, the $\sqrt{nh^d}$-consistency of $E_D[\psi_m(\tau_h, \cdot)]$ (i.e., $E_D[\psi_m(\tau_h, \cdot)] = O_P(1/\sqrt{nh^d})$) can be shown immediately by the Markov inequality. This implies that $E_D[\psi_m(\tau_h, \cdot)]$ converges if $nh^d \to 1$.

Let $\psi_m(V_i, \cdot, \cdot)$ denote the influence function evaluated at $V_i \in D$.

Consider the following:

$$\begin{aligned}
E_P[|E_D[\psi_m(\tau_h, \cdot)]|] &\le \sqrt{E_P[(E_D[\psi_m(\tau_h, \cdot)])^2]} \\
&= \sqrt{\mathrm{var}_P(E_D[\psi_m(\tau_h, \cdot)])} \\
&= \sqrt{(1/n) E_P[\psi_m^2(\tau_h, \cdot)]},
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality, the second and third equality are from the iid assumption and $E_P[\psi_m] = 0$.

We note that

$$\begin{aligned}
\psi_m &= \frac{1}{\mu_X}(V_{YX}[K_{h,y}(Y)] - \tau_h V_X) \\
&= \frac{1}{\mu_X}(V_{YX}[K_{h,y}(Y)] - \tau_h V_X) + \underbrace{\frac{\mu_{YX}[K_{h,y}(Y)]}{\mu_X} - \frac{\mu_X}{\mu_X}\tau_h}_{=0} \\
&= \frac{1}{\mu_X} V_{YX}[K_{h,y}(Y)] - \frac{\mu_{YX}[K_{h,y}(Y)]}{\mu_X} \tau_h - \tau_h V_X + \tau_h \mu_X \\
&= \frac{1}{\mu_X}(\psi_{YX}[K_{h,y}(Y)] - \tau_h \psi_X).
\end{aligned}$$

Next,

$$\begin{aligned}
E_P\left[\sigma^2_m(\hat{\theta}_h;\theta)\right] &= E_P\left[\frac{1}{\sigma^2_X}\left\{\mu_{XY}[K_{h;y}(Y)] - \theta_h \mu_X\right\}^2\right] \\
&= \frac{1}{\sigma^2_X} E_P\left[\left\{\mu_{XY}[K_{h;y}(Y)] - \theta_h \mu_X\right\}^2\right] \\
&= \frac{1}{\sigma^2_X} E_P\left[\sigma^2_{XY}[K_{h;y}(Y)] + \theta_h^2 \sigma^2_X - 2\,\sigma_{XY}[K_{h;y}(Y)]\,\mu_X\,\theta_h\right].
\end{aligned}$$

We first analyze $E_P\left[\sigma^2_{XY}[K_{h;y}(Y)]\right] = \mathrm{var}_P[\mu_{XY}[K_{h;y}(Y)]]$. By [28, Thm. 1],

$$\begin{aligned}
\mathrm{var}_P[\mu_{XY}[K_{h;y}(Y)]] = E_P\Bigg[&\frac{\mathrm{Var}_P(K_{h;y}(Y)1_x(X)|z^x;W)}{z^x(W)} + \frac{\mathrm{Var}_P\left(K_{h;y}(Y)1_x(X)|z^{1-x};W\right)}{z^{1-x}(W)}\Bigg] \\
&+ E_P\left[E_P[K_{h;y}(Y)1_x(X)|z^x;W] - E_P\left[K_{h;y}(Y)1_x(X)|z^{1-x};W\right]\right]^2 - \left[\sigma_{YX}[K_{h;y}]\right]^2.
\end{aligned}$$

First,

$$\begin{aligned}
E_P[\mathrm{Var}_P(K_{h;y}(Y)1_x(X)|z^x;W)] &\le \mathrm{var}_P(K_{h;y}(Y)1_x(X)|z^x) \\
&\le E_P\left[K^2_{h;y}(Y)1_x(X)|z^x\right] \\
&\le E_P\left[K^2_{h;y}(Y)|x;z^x\right] \\
&= \int_{\mathcal{Y}} K^2_{h;y}(y')p(y|x;z^x)\,d[y'] \\
&\le \int_{\mathcal{Y}} K^2_{h;y}(y')\,d[y'] \\
&= \frac{1}{h^{2d}}\int_{\mathcal{Y}} K^2\left(\frac{y'-y}{h}\right)d[y'] \\
&= \frac{1}{h^d}\int_{\mathcal{Y}} K^2(u)\,d[u] \\
&= O\left(1/h^d\right). \qquad \text{(B.6)}
\end{aligned}$$

The 1st equality holds by Law of total variance, the 2nd and 3rd by the standard algebra, the 5th by the assumption that $p(y|x;z^x)$ is bounded, and the remaining parts from the change of a variable in the integral computation.

Also,

$$\begin{aligned}
&E_P\left[E_P[K_{h;y}(Y)1_x(X)|z^x;W] - E_P\left[K_{h;y}(Y)1_x(X)|z^{1-x};W\right]\right]^2 - \left[\sigma_{YX}[K_{h;y}]\right]^2 \\
&= \mathrm{var}_P\left[E_P[K_{h;y}(Y)1_x(X)|z^x;W] - E_P\left[K_{h;y}(Y)1_x(X)|z^{1-x};W\right]\right] \\
&\quad 2\sup_{z\in\{0,1\}}\mathrm{var}_P(E_P[K_{h;y}(Y)1_x(X)|z;W]) \\
&= O(1/h^d);
\end{aligned}$$

where the first (in)equality is by the definition of the variance, the second by the linear combination of the variance, and the last by Eq. (B.6). Therefore, $\mathrm{var}_P[\sigma^2_{XY}[K_{h;y}(Y)]] = O(1/h^d)$.

Next, we will study $E_P[\hat{\sigma}_h^2 \hat{\sigma}_X^2]$. We first note that $E_P[\hat{\sigma}_h^2 \hat{\sigma}_X^2] = \hat{\sigma}_h^2 E_P[\hat{\sigma}_X^2] = O(\hat{\sigma}_h^2)$. Therefore, it suffices to analyze $O(\hat{\sigma}_h^2)$.

$$\begin{aligned}
\hat{\sigma}_h^2 &= \int_Y \left[\int_Y K_{h;y}(y') (y') \, d[y'] \right]^2 \\
&\quad \int_Y K_{h;y}^2(y') \, {}^2(y') \, d[y'] \\
&\quad \int_Y K_{h;y}^2(y') \, d[y'] \\
&= \int_Y \frac{1}{h^{2d}} K^2 \left(\frac{y' - y}{h} \right) d[y'] \\
&= \int_U \frac{1}{h^d} K^2(u) \, d[u] \\
&= O(1/h^d);
\end{aligned}$$

where the 2nd line inequality by the Cauchy-Schwarz inequality, the 3rd by the assumption that (y) is bounded, the fifth by the change of variables.

Finally, consider the term $2E_P[{}_{Y|X}[K_{h;y}(Y)] \, {}_X \, {}_h]$. Note, $E_P[{}_{Y|X}[K_{h;y}(Y)] \, {}_X \, {}_h] = {}_h \, E_P[{}_{Y|X}[K_{h;y}(Y)] \, {}_X]$. We first consider $E_P[{}_{Y|X}[K_{h;y}(Y)] \, {}_X]$:

$$\begin{aligned}
E_P[{}_{Y|X}[K_{h;y}(Y)] \, {}_X] &= E_P[{}_{Y|X}[K_{h;y}(Y)] \, {}_X] \\
&\quad \sqrt{\frac{}{E_P[{}^2_{Y|X}[K_{h;y}(Y)]] \, E_P[{}^2_X]}} \\
&= O\left(\sqrt{E_P[{}^2_{Y|X}[K_{h;y}(Y)]]} \right) = O\left(h^{-d/2} \right);
\end{aligned}$$

where the last equality holds by Eq. (B.6).

Next, consider ${}_h$:

$$\begin{aligned}
{}_h &= \int_Y K_{h;y}(y') \, (y') \, d[y'] \\
&= \int_Y \frac{1}{h} K\left(\frac{(y' - y)}{h} \right) (y') \, d[y'] \\
&= \int_U K(u) \, (hu + y) \, d[u] \\
&= \int_U K(u) \left[(y) + hu \, {}^{(1)}(y) + h^2 u^2 \, {}^{(2)}(y) + O(h^2 u^2) \right] d[u] \\
&= C + O(h^2);
\end{aligned}$$

for some constant C, The 4th line equality holds by the differentiability assumption of (y), and the last equality holds since (y) is bounded and twice differentiable. Combining, we can rewrite the term $2E_P[{}_{Y|X}[K_{h;y}(Y)] \, {}_X \, {}_h]$ as $O(h^{-d/2} + h^{-d/2} h^2)$.

Therefore,

$$E_P[{}^2_m({}_h; \,)] = O(h^{-d} + h^{-d/2} + h^{-d/2} h^2).$$

With $h = O(n^{-})$ with $2 [0; 1/d)$, we can rewrite

$$E_P[{}^2_m({}_h; \,)] = O(h^{-d} + h^{-d/2} + h^{-d/2} h^2) = O\left(n^{d} \right) = O(h^{-d}).$$

This shows that

$$E_P[E_D[{}_m({}_h; \,)]] \quad \sqrt{(1/n) E_P[{}^2_m({}_h; \,)]} = O\left(1/\sqrt{nh^d} \right).$$

We now consider $E_P[\ _m(\ _h;\hat{}\,)]$.

$$
\begin{aligned}
&E_P[\ _m(\ _h;\hat{}\,)] \\
&= E_P\left[\frac{1}{\hat{}_X}\ \hat{V}_{YX}[K_{h;y}(Y)]\quad[K_{h;y}(Y)]\hat{V}_X\right] \\
&= E_P\left[\frac{1}{\ _X}\ \hat{V}_{YX}[K_{h;y}(Y)]\quad[K_{h;y}(Y)]\hat{V}_X\right] + E_P\left[\frac{1}{\hat{}_X}\quad\frac{1}{\ _X}\ \hat{V}_{YX}[K_{h;y}(Y)]\quad[K_{h;y}(Y)]\hat{V}_X\right]: \qquad \text{(B.7)}
\end{aligned}
$$

For further analysis, we consider $E_P\left[\hat{V}_{YX}[K_{h;y}(Y)]\ \ V_{YX}[K_{h;y}(Y)]\right]$. First, define

$$
V_{YX;(x;z)}(\ ;\ )[f(Y)] \quad \frac{1_z(Z)}{\ _Z(W)}(f(Y)1_x(X)\quad\ (x;Z;W)[f(Y)]) + \ (x;z;W)[f(Y)]:
$$

Then, $V_{YX}[f(Y)] = V_{YX;(x;z^x)}[f(Y)]\quad V_{YX;(x;z^{1-x})}[f(Y)]$. Now, consider $E_P\left[\hat{V}_{YX;(x;z)}[K_{h;y}(Y)]\ \ V_{YX;(x;z)}[K_{h;y}(Y)]\right]$. We have

$$
\begin{aligned}
&E_P\left[V_{YX;(x;z)}(\hat{};\hat{}\,)[f(Y)]\ \ V_{YX;(x;z)}(\ ;\ )[f(Y)]\right] \\
&= E_P\left[\frac{1_z(Z)}{\hat{}_Z(W)}\ f(Y)1_x(X)\quad\hat{}(x;Z;W)[f(Y)] + \hat{}(x;z;W)[f(Y)]\quad\ (x;z;W)[f(Y)]\right] \\
&= E_P\left[\frac{1_z(Z)}{\hat{}_Z(W)}\quad\ (x;Z;W)[f(Y)]\quad\hat{}(x;Z;W)[f(Y)] + \hat{}(x;z;W)[f(Y)]\quad\ (x;z;W)[f(Y)]\right] \\
&= E_P\left[\frac{\ _z(W)}{\hat{}_z(W)}\quad(x;z;W)[f(Y)]\quad\hat{}(x;z;W)[f(Y)] + \hat{}(x;z;W)[f(Y)]\quad\ (x;z;W)[f(Y)]\right] \\
&= E_P\left[\quad(x;z;W)[f(Y)]\quad\hat{}(x;z;W)[f(Y)]\quad 1\quad\frac{\ _z(W)}{\hat{}_z(W)}\right] \\
&= E_P\left[\quad(x;z;W)[f(Y)]\quad\hat{}(x;z;W)[f(Y)]\quad\frac{\hat{}_z(W)\quad\ _z(W)}{\hat{}_z(W)}\right] \\
&= O_P\left[\quad(x;z;W)[f(Y)]\quad\hat{}(x;z;W)[f(Y)]\quad k\hat{}_z(W)\quad\ _z(W)k\ ;\right.
\end{aligned}
$$

where the first and the second are by the fact that $E_P[f(Y)1_x(X)|W;Z;X] = \ (x;Z;W)[f(Y)]$, the third is by taking an expectation over Z conditioned on W, the fourth and the fifth by rearrangement, and the sixth by Cauchy-Schwarz inequality and Positivity. Then,

$$
\begin{aligned}
R_{YX} &\quad E_P\left[V_{YX}(\hat{};\hat{}\,)[f(Y)]\ \ V_{YX}(\ ;\ )[f(Y)]\right] \\
&= \sum_{z2\{0;1\}} O_P\left[\quad(x;z;W)[f(Y)]\quad\hat{}(x;z;W)[f(Y)]\quad k\hat{}_z(W)\quad\ _z(W)k\ :\right.
\end{aligned}
$$

Also, let

$$
V_{X;x}(\ ;\ )\quad\frac{1_z(Z)}{\ _Z(W)}(1_x(X)\quad\ _x(Z;W)) + \ _x(z;W):
$$

Then, with the similar proof as above, we have

$$
E_P\left[V_{X;x}(\hat{};\hat{}\,)\ \ V_{X;x}(\ ;\ )\right] = O_P\left[\quad\ _x(z;W)\quad\hat{}_x(z;W)\quad k\hat{}_z(W)\quad\ _z(W)k\ ;\right.
$$

and

$$
E_P\left[V_X(\hat{};\hat{}\,)\ \ V_X(\ ;\ )\right] = \sum_{z2\{0;1\}} O_P\left[\quad\ _x(z;W)\quad\hat{}_x(z;W)\quad k\hat{}_z(W)\quad\ _z(W)k\ :\right.
$$

Recall $R_{YX} = E_P\left[\hat{V}_{YX} - V_{YX}\right]$ and let $R_X \quad E_P\left[\hat{V}_X - V_X\right]$. Then, continuing from Eq. (B.7),

$$\begin{aligned}
\text{Eq. (B.7)} &= E_P\left[\frac{1}{x} \quad {}_{YX} + R_{YX} \quad \frac{{}_{YX}}{x}\left({}_X + R_X\right) + \frac{1}{\hat{}_X} \quad \frac{1}{x} \quad {}_{YX} + R_{YX} \quad \frac{{}_{YX}}{x}\left({}_X + R_X\right)\right] \\
&= E_P\left[\frac{1}{x}\left(R_{YX} \quad R_X\right) + \frac{1}{\hat{}_X} \quad \frac{1}{x}\left(R_{YX} \quad R_X\right)\right] \\
&= O_P\left(R_{YX} + R_X\right) \\
&= O_P\left(R_2^k\right);
\end{aligned}$$

where

$$R_2^k = \sum_{z2\{0;1\}} O_P\left(k\hat{}_z \quad {}_z k\left\|\hat{}_z \quad {}_z\right\| + \left\|\hat{}_z \quad {}_z\right\|\right):$$

Note the first equality is by $E_P\left[\hat{V}_{YX}\right] = R_{YX} + E_P\left[V_{YX}\right]$ and $E_P\left[\hat{V}_X\right] = R_X + E_P\left[V_X\right]$, the second by rearrangement, the third by Positivity, the fourth by the definition of R_{YX} and R_X.

Summing up, we have shown that $E_P\left[{}_m({}_h;)\right] = O(1=\sqrt{nh^d})$ and $E_P\left[{}_m({}_h; \hat{})\right] = O_P\left(R_2^k\right)$.

□

Corollary 1 ((Restated Corol. 1)) If all nuisances $\{\hat{}; \hat{}; \hat{}\}$ for any given $(w; z; x; y)$ converge at rate $\{nh^d\}^{1=4}$, then the target estimator $\hat{}_h(y)$ achieves $\sqrt{nh^d}$-rate convergence to ${}_h$.

Proof. This result follows immediately from Lemma 3. □

Theorem B.1 (Restated Thm. 1) For any fixed $y \in Y$, suppose the estimators for nuisances are consistent; i.e., $\|\hat{} \quad \| = o_P(1)$ for $= \{ ; ; \}$ for all $(w; z; x)$. Suppose $h < 1$, and $nh^d \to 1$ as $n \to 1$. Then

$$\hat{}_h(y) \quad (y) = O_P\left(1=\sqrt{nh^d} + R_2^k + 1=\sqrt{n}\right) + B_y; \tag{B.8}$$

where B_y is defined in Eq. (14), and R_2^k is defined in Eq. (13).

Proof. This result follows immediately from Lemmas 3 and 4. □

Lemma B.5 (Restated Lemma 5) The bandwidth h that minimizes the error in Eq. (15) is $h = O(n^{1=(d+4)})$. This choice of h satisfies the assumption in Lemma. 3 that $nh^d \to 1$.

Proof. We note that the error in Eq. (15) w.r.t. h is $O_P\left(1=\sqrt{nh^d} + h^2\right)$. Since the function $1=\sqrt{nh^d} + h^2$ is convex w.r.t. h and the global minimum is at $h = n^{1=(d+4)}$, the optimal h minimizing the error is $h = O(n^{1=(d+4)})$. Then, $O(nh^d) = O(n^{4=(d+4)})$, implying that $nh^d \to 1$. □

Corollary 2 (Restated Corol. 2) Let $h = O(n^{1=(d+4)})$. If nuisances $\{\hat{}; \hat{}; \hat{}\}$ converge at $\{nh^d\}^{1=4}$ rate for any $(w; z; x; y)$, then the target estimator $\hat{}_h(y)$ achieves $\sqrt{nh^d}$-rate convergence to .

Proof. It suffices to show that B_y converges at $\sqrt{nh^d}$-rate with the choice of h as in Lemma 5, since the rest is guaranteed by Corol. 1. We first note that $B_y = O(h^2)$. Since $O(nh^d) = O(n^{4=(d+4)})$, we have $O(1=\sqrt{nh^d}) = O(n^{2=(d+4)}) = O(h^2)$. □

Lemma B.6 (Restated Lemma 6) Suppose D_f is a f-divergence such that $f(p; q) = 0$ if $p = q$. Then,

$$D_f\left(; \hat{}_h\right) \quad \int_Y w(y) \quad \left\|\hat{}_h(y) \quad (y)\right\| \, d[y];$$

where $w(y) \quad f_2'\left((y); \tilde{}(y)\right)\left\|\hat{}_h(y) \quad (y)\right\|$, $f_2'(p; q) \quad (@=@q)f(p; q)$, and $\tilde{}_h(y) \quad t\,\hat{}_h(y) + (1 \quad t)\,(y)$ for some fixed $t \in [0; 1]$.

Proof. For $f(p,q)$, by applying Taylor's expansion, we have

$$f(p,q) = f(p,p) + f_2^0(p,p_\xi)(q-p),$$

for some $p_\xi \in [p,q]$. Applying this idea,

$$\begin{aligned} D_f(\theta, \hat{\theta}_h) &= \int_Y f(\pi(y), \hat{\pi}_h(y))\, \hat{\pi}_h(y)\, d[y] \\ &= \int_Y \Big\{ \underbrace{f(\pi(y), \pi(y))}_{=0} + f_2^0(\pi(y), \tilde{\pi}(y))\,(\hat{\pi}_h(y) - \pi(y)) \Big\}\, \hat{\pi}_h(y)\, d[y]; \\ &= \int_Y w(y)\,(\hat{\pi}_h(y) - \pi(y))\, d[y]; \end{aligned}$$

were the second equality holds by Taylor's expansion for $\tilde{\pi}$ and the third equality by the given assumption that $f(p,q) = 0$ whenever $p = q$.

□

Theorem B.2 (Restated Thm. 2) *Suppose the estimators for nuisances are consistent, i.e., $\hat{\pi}_k - \pi_k = o_P(1)$ for $k = 2 \in \{f, g, \ldots\}$ for all (w,z,x,y). Suppose D_f is a f-divergence such that $f(p,q) = 0$ if $p = q$. Suppose $w(y)$ in Lemma 6 is finite. Then,*

$$D_f(\theta, \hat{\theta}_h) - O_P\Big(\sup_{y \in Y} R_2^k + B_y\Big) + 1/\sqrt{nh^d} + 1/\sqrt{n}\Big), \tag{B.9}$$

where R_2^k is defined in Eq. (13) and B_y is defined in Eq. (14).

Proof. Under the given conditions, with Thm. 1,

$$\begin{aligned} D_f(\theta, \hat{\theta}_h) &= \int_Y w(y)\,(\hat{\pi}_h(y) - \pi(y))\, d[y] \\ &= \int_Y w(y)\, O_P\big(1/\sqrt{nh^d} + R_2^k + 1/\sqrt{n}\big) + B_y\big)\, d[y] \\ &= O_P(1/\sqrt{nh^d} + 1/\sqrt{n}) + \int_Y (w(y) O_P(R_2^k) + B_y)\, d[y] \\ &= O_P(1/\sqrt{nh^d} + 1/\sqrt{n}) + O_P\Big(\sup_{y \in Y} R_2^k + B_y\Big). \end{aligned}$$

□

Corollary 3 (Restated Corol. 3) *Let $h = O(n^{-1/(d+4)})$. Suppose D_f satisfies $f(p,q) = 0$ if $p = q$. Suppose $w(y)$ in Lemma 6 is finite. If nuisances $\{\hat{\pi}, \hat{\mu}, \hat{\pi}\}$ converges at $o_P(\{nh^d\}^{-1/4})$ rate for any (w,z,x,y), then $D_f(\theta, \hat{\theta}_h)$ converges to 0 at $\sqrt{nh^d}$-rate.*

Proof. This result follows immediately from Thm. 2. □

B.2 Proofs for Sec. 4

We will use θ_p to denote θ as a functional of p. Let p_τ denote a parametric submodel. We will use S to denote a score function for p_τ.

Lemma B.7 (Restated Lemma 7) *An influence function for $m(\theta, \pi, \mu)$ in Eq. (18), denoted ϕ_m, is given by*

$$\phi_m(\theta, \pi, \mu, \{w, z, x, y\}) = \psi(\theta, \pi, \mu)[R_f(Y, \pi, \mu)], \tag{B.10}$$

where $\psi(\theta, \pi, \mu)[\cdot]$ is defined in Eq. (9), and

$$R_f(Y, \pi, \mu) = g^0(Y, \mu)\{f_{21}^{00}(\pi(Y), g(Y, \mu))\, g(Y, \mu) + f_1^0(\pi(Y), g(Y, \mu))\},$$

where $g^0(y, \mu) = (\partial/\partial g)\, g(y, \mu)$, $f_1^0(p,q) = (\partial/\partial p) f(p,q)$ and $f_{21}^{00}(p,q) = (\partial/\partial p) f_2^0(p,q)$.

Proof. Let θ denote the estimand written w.r.t. the parametric submodel $p_t = p(1 + tg)$ where g is a bounded mean-zero random function. Set $\frac{d}{dt}|_{t=0} \log p_t$.

Let

$$\bar{m}(y; \theta, \eta) \equiv g'(y; \eta) f f_2'(\eta(y); g(y; \eta)) g(y; \eta) + f(\eta(y); g(y; \eta)) g. \tag{B.11}$$

Note $m(\theta, \eta) = \int_Y \bar{m}(y; \theta, \eta) d[y]$. Also, we note that $\frac{d}{dt}|_{t=0} \bar{m}(y; \theta, \eta) = R_f(y; \theta, \eta)$.

Also, recall that an influence function for $\theta[f(Y)]$ (for $f(Y) < 1$) is given as $\phi(\theta, \eta)[f(Y)]$ in Lemma S.3. Then, by the definition of the influence function, $\theta[f(Y)]$ satisfies the following,

$$\frac{d}{dt}|_{t=0} \theta[f(Y)] = E_P[\phi(\theta, \eta)[f(Y)] S].$$

Now, we will prove that $\psi_m(\theta, \eta = \{f, g\}) \equiv \phi(\theta, \eta)[R_f(Y; \theta, \eta)]$ is a functional satisfying

$$\frac{d}{dt}|_{t=0} m(\theta, \eta) = E_P[\phi(\theta, \eta)[R_f(Y; \theta, \eta)] S];$$

then this equation implies that $\phi(\theta, \eta)[R_f(Y; \theta, \eta)]$ is an influence function for the score $m(\theta, \eta)$.

This can be shown as follows:

$$\begin{aligned}
&\frac{d}{dt}|_{t=0} m(\theta, \eta) \\
&= \frac{d}{dt}|_{t=0} \int_Z \int_Y \bar{m}(y; \theta, \eta) d[y] \\
&= \int_Z \frac{d}{dt}|_{t=0} \int_Y \bar{m}(y; \theta, \eta) d[y] \\
&= \int_Y \frac{d}{dt}|_{t=0} \phi(y)(\partial/\partial \theta(y))|_{\theta=\theta_0} \bar{m}(y; \theta, \eta) d[y] \\
&= \frac{d}{dt}|_{t=0} \int_Y \phi(y) R_f(y; \theta, \eta) d[y] \\
&= \frac{d}{dt}|_{t=0} [R_f(Y; \theta, \eta)] \\
&= E_P[\phi(\theta, \eta)[R_f(Y; \theta, \eta)] S];
\end{aligned}$$

where the first equality is by the definition of m, the second by the exchange of derivation/integration, the third by the chain rule, the fourth by the fact that $\frac{d}{dt}|_{t=0} \bar{m}(y; \theta, \eta) = R_f(y; \theta, \eta)$ and the exchange of derivation/integration, the fifth by the definition of $\theta[f(Y)]$ in Eq. (9), the sixth by the definition of the influence function (i.e., the influence function for $\theta[f(Y)]$ is a function $\phi[f(Y)]$ satisfying $\frac{d}{dt}|_{t=0} \theta[f(Y)] = E_P[\phi[f(Y)] S]$.

□

Lemma B.8 ((Restated Lemma 8)) A Neyman orthogonal score for estimating θ denoted $\psi(\theta_0, \eta(= \{f, g\}))$, is given by

$$\psi(\theta_0, \eta(= \{f, g\})) \equiv m(\theta_0, \eta) + \psi_m(\theta_0, \eta); \tag{B.12}$$

where $\psi_m(\theta, \eta)$ is defined in Eq. (19).

Proof. We first note that $E_P[m(\theta_0, \eta)] = m(\theta_0, \eta)$, because this is not a random function. Then, the influence function for $E_P[m(\theta_0, \eta)]$ is given by Lemma 7. For any score function which expectation is zero at the true parameter, its addition with the influence function is a Neyman orthogonal score [14, Thm.1]. That is, $m(\theta_0, \eta) + \psi_m(\theta, \eta)$ is a Neyman orthogonal score. □

Theorem B.3 ((Restated Thm. 3)) Let $\psi(\theta_0, \eta(= \{f, g\}))$ be given in Eq. (20). Let $\psi_m(\theta, \eta)$ be given in Eq. (19). Let $\theta_0, \eta_0 \equiv \{f_0, g_0\}$ denote the true parameters. Let $\hat{\theta}$ be the MLTE estimator for θ defined in Def. 3. Suppose (1) $R_f(y; \theta, \eta)$ is bounded and $R_f^0(y; \theta, \eta) \equiv \frac{d}{dt} R_f(y; \theta, \eta) < 1$; (2) There exists a function $H(y) < 1$ s.t. $\sup_{\theta, \eta} \max\{R_f(y; \theta, \eta); R_f^0(y; \theta, \eta)\} = O(H(y))$; (3) $\{\psi(\theta, \eta(\theta, \eta))\}$ is Donsker w.r.t. θ for the fixed η; (3) The estimators are consistent: $\hat{\eta} - \eta_0 = o_P(1)$ and $\|\hat{\eta} - \eta\| = o_P(1)$ for $\eta \in \{\eta_z(w); \eta_x(z; w); \eta(x; z; w)[H(Y)]\}$ for all $(w; z; x; y)$; and (4)

$E_P[\psi(\cdot;(\cdot;\cdot))]$ is differentiable w.r.t. $\cdot$ at $\cdot = \cdot_0$ with non-singular matrix $M(\cdot_0;(\cdot;\cdot))$ $(\partial/\partial\cdot)|_{\cdot=\cdot_0} E_P[\psi(\cdot;(\cdot;\cdot))]$ for all $(\cdot;\cdot)$, where $M(\cdot_0;(\hat{\cdot};\hat{\cdot})) \xrightarrow{P} M$ $M(\cdot_0;(\cdot_0;\cdot_0))$. Then,

$$\hat{b}_0 = M^{-1} E_D[m(\cdot_0;(\cdot_0;\cdot_0))] + o_P(n^{-1/2}) + O_P(R_2^m);$$

where

$$R_2^m = \sum_z k\hat{z}_z \; {}_z k \left\{ \hat{z}_z \; {}_z + \hat{z}_z \; {}_z + \hat{z}_z \; {}_z{}^2 + {}_z \; \hat{z}_z{}^2 + \hat{z}_z \; {}_z \; {}_z \; \hat{z}_z \right\};$$

where ${}_z \; {}_z(W)$, ${}_z \; {}_x(z;W)$, and ${}_z \; (x;z;W)[H(Y)]$.

Proof. We follow the proof strategy used in [39, Lemma 1, Thm.3]. First,

$$\begin{aligned}\hat{b}_0 &= M^{-1} E_D[\psi(\cdot_0;(\cdot_0;\cdot_0))] \quad M^{-1} E_P\left[\psi(\cdot_0;(\hat{\cdot};\hat{b}))\right] + o_P(n^{-1/2}) \\ &= M^{-1} E_D[m(\cdot_0;f_{\cdot_0}(\cdot_0;\cdot_0))] \quad M^{-1} E_P\left[\psi(\cdot_0;(\hat{\cdot};\hat{b}))\right] + o_P(n^{-1/2}); \qquad \text{(B.13)}\end{aligned}$$

where the first equality holds by Lemma S.1, and the second holds since $m(\cdot_0;\cdot_0) = 0$ by the moment condition in Eq. (18). Since $E_D[m(\cdot_0;\cdot_0;\cdot_0)]$ converges to $N(0;\mathrm{var}(\cdot_m^2))$ in distribution at $\sqrt{n}$-rate, the only remaining term to analyze is

$$E_P\left[\psi(\cdot_0;(\hat{\cdot};\hat{b}))\right] = m(\cdot_0;\hat{\cdot}) + E_P\left[(\cdot_0;(\hat{\cdot};\hat{b}))[R_f(Y;\cdot_0;\hat{\cdot})]\right]; \qquad \text{(B.14)}$$

which can be analyzed as

$$\begin{aligned}&E_P\left[(\cdot_0;(\hat{\cdot};\hat{b}))[R_f(Y;\cdot_0;\hat{\cdot})]\right] \\ &= E_P\left[\frac{1}{\hat{\cdot}_X}\left\{\hat{\cdot}_{YX}[R_f(Y;\cdot_0;\hat{\cdot})] \quad \hat{\cdot}[R_f(Y;\cdot_0;\hat{\cdot})]\hat{\cdot}_X\right\}\right] \\ &= E_P\left[\frac{1}{\hat{\cdot}_X}\hat{\cdot}_{YX}[R_f(Y;\cdot_0;\hat{\cdot})]\right] \quad E_P\left[\frac{1}{\hat{\cdot}_X}\hat{\cdot}[R_f(Y;\cdot_0;\hat{\cdot})]\hat{\cdot}_X\right] \\ &= E_P\left[\frac{1}{\hat{\cdot}_X}\frac{\hat{\cdot}_{z^x}(W)}{{}_{z^x}(W)}\left\{(x;z^x;W)[R_f(Y;\cdot_0;\hat{\cdot})] \quad \hat{\cdot}(x;z^x;W)[R_f(Y;\cdot_0;\hat{\cdot})]\right\} + \hat{\cdot}(x;z^x;W)R_f(Y;\cdot_0;\hat{\cdot})\right] \qquad \text{(B.15)} \\ &\quad E_P\left[\frac{1}{\hat{\cdot}_X}\frac{\hat{\cdot}_{z^{1\ x}}(W)}{{}_{z^{1\ x}}(W)}\left\{(x;z^{1\ x};W)[R_f(Y;\cdot_0;\hat{\cdot})] \quad \hat{\cdot}(x;z^{1\ x};W)[R_f(Y;\cdot_0;\hat{\cdot})]\right\} + \hat{\cdot}(x;z^{1\ x};W)[R_f(Y;\cdot_0;\hat{\cdot})]\right] \qquad \text{(B.16)} \\ &\quad E_P\left[\frac{1}{\hat{\cdot}_X}\hat{\cdot}[R_f(Y;\cdot_0;\hat{\cdot})]\frac{{}_{z^x}(W)}{\hat{\cdot}_{z^x}(W)}\left\{{}_x(z^x;W) \quad \hat{\cdot}_x(z^x;W)\right\} + \hat{\cdot}_x(z^x;W)\right] \qquad \text{(B.17)} \\ &+ E_P\left[\frac{1}{\hat{\cdot}_X}\hat{\cdot}[R_f(Y;\cdot_0;\hat{\cdot})]\frac{{}_{z^{1\ x}}(W)}{\hat{\cdot}_{z^{1\ x}}(W)}\left\{{}_x(z^{1\ x};W) \quad \hat{\cdot}_x(z^{1\ x};W)\right\} + \hat{\cdot}_x(z^{1\ x};W)\right]; \qquad \text{(B.18)}\end{aligned}$$

where

$$\text{(B.15)} = E_P\left[\frac{1}{\hat{\cdot}_X}\frac{\hat{\cdot}_{z^x}(W)}{{}_{z^x}(W)}\left\{1 \quad (x;z^x;W)[R_f(Y;\cdot_0;\hat{\cdot})] \quad \hat{\cdot}(x;z^x;W)[R_f(Y;\cdot_0;\hat{\cdot})]\right\}\right] \qquad \text{(B.19)}$$

$$+ E_P\left[\frac{1}{\hat{\cdot}_X}(x;z^x;W)[R_f(Y;\cdot_0;\hat{\cdot})]\right] \qquad \text{(B.20)}$$

$$\text{(B.16)} = E_P\left[\frac{1}{\hat{\cdot}_X}\frac{\hat{\cdot}_{z^{1\ x}}(W)}{{}_{z^{1\ x}}(W)}\left\{1 \quad (x;z^{1\ x};W)[R_f(Y;\cdot_0;\hat{\cdot})] \quad \hat{\cdot}(x;z^{1\ x};W)[R_f(Y;\cdot_0;\hat{\cdot})]\right\}\right] \qquad \text{(B.21)}$$

$$E_P\left[\frac{1}{\hat{\cdot}_X}(x;z^{1\ x};W)[R_f(Y;\cdot_0;\hat{\cdot})]\right] \qquad \text{(B.22)}$$

$$(B.17) = E_P\left[\frac{1}{\hat{\pi}_X}\hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\left(\frac{\pi_{z^x}(W)}{\hat{\pi}_{z^x}(W)}\right)\left(1-\frac{\pi_x(z^x;W)}{\hat{\pi}_x(z^x;W)}\right)\right] \quad (B.23)$$

$$E_P\left[\frac{1}{\hat{\pi}_X}\hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\frac{\pi_x(z^x;W)}{\pi_x(z^x;W)}\right] \quad (B.24)$$

$$(B.18) = E_P\left[\frac{1}{\hat{\pi}_X}\hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\left(\frac{\pi_{z^{1-x}}(W)}{\hat{\pi}_{z^{1-x}}(W)}\right)\left(1-\frac{\pi_x(z^{1-x};W)}{\hat{\pi}_x(z^{1-x};W)}\right)\right] \quad (B.25)$$

$$+ E_P\left[\frac{1}{\hat{\pi}_X}\hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\frac{\pi_x(z^{1-x};W)}{\pi_x(z^{1-x};W)}\right] \quad (B.26)$$

First, consider the summation of (B.20,B.22,B.24,B.26):

Eq. (B.20) + Eq. (B.22) + Eq. (B.24) + Eq. (B.26)

$$= E_P\left[\frac{1}{\hat{\pi}_X}\left(\frac{\pi(x;z^x;W)}{\pi(x;z^x;W)}[R_f(Y;\theta_0;\hat{\eta})] - \frac{\pi(x;z^{1-x};W)}{\pi(x;z^{1-x};W)}[R_f(Y;\theta_0;\hat{\eta})]\right)\right]$$

$$E_P\left[\frac{1}{\hat{\pi}_X}\hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\frac{\pi_x(z^x;W)}{\pi_x(z^x;W)} - \frac{\pi_x(z^{1-x};W)}{\pi_x(z^{1-x};W)}\right]$$

$$= E_P\left[\frac{1}{\hat{\pi}_X}\left(\pi_{YX}[R_f(Y;\theta_0;\hat{\eta})] - \hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\frac{\pi_X}{\hat{\pi}_X}\right)\right]$$

$$= E_P\left[\frac{1}{\hat{\pi}_X}\left(\pi_{YX}[R_f(Y;\theta_0;\hat{\eta})] - \frac{\hat{\pi}_{YX}[[R_f(Y;\theta_0;\hat{\eta})]]}{\hat{\pi}_X}\frac{\pi_X}{\hat{\pi}_X}\right)\right]$$

$$= E_P\left[\frac{\pi_X}{\hat{\pi}_X}\left(\frac{\pi_{YX}[R_f(Y;\theta_0;\hat{\eta})]}{\pi_X} - \frac{\hat{\pi}_{YX}[R_f(Y;\theta_0;\hat{\eta})]}{\hat{\pi}_X}\right)\right]$$

$$= E_P\left[\frac{\pi_X}{\hat{\pi}_X}\left[R_f(Y;\theta_0;\hat{\eta})] - \hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\right]\right]:$$

$$= E_P\left[\frac{\pi_X}{\hat{\pi}_X}\left(1-\frac{\hat{\pi}_X}{\pi_X}\right)[R_f(Y;\theta_0;\hat{\eta})] - \hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\right] + E_P\left[[R_f(Y;\theta_0;\hat{\eta})] - \hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\right]:$$

(B.27)

Then,

Eq. (B.14)= $m(\theta_0;\hat{\eta})$ + Sum of (B.20, B.22, B.24, B.26) + Sum of (B.19, B.21, B.23, B.25)

$$= m(\theta_0;\hat{\eta}) + E_P\left[[R_f(Y;\theta_0;\hat{\eta})] - \hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\right] \quad (B.28)$$

$$+ E_P\left[\frac{\pi_X}{\hat{\pi}_X}\left(1-\frac{\hat{\pi}_X}{\pi_X}\right)[R_f(Y;\theta_0;\hat{\eta})] - \hat{\Delta}[R_f(Y;\theta_0;\hat{\eta})]\right] + \text{Sum of (B.19,B.21,B.23,B.25)} \quad (B.29)$$

To analyze (B.28), we recall that $\frac{\partial}{\partial\theta}m(\theta_0;\theta) = \int_Y R_f(y;\theta_0;\theta)\,d[y]$ and $m(\theta_0;\theta_0) = 0$. Also, by Taylor's expansion of $\bar{m}(y;\theta;\eta)$ defined in Eq. (B.11),

$$\bar{m}(y;\theta_0;\eta) = \bar{m}(y;\theta_0;\hat{\eta}) + R_f(y;\theta;\hat{\eta})(\eta(y) - \hat{\eta}(y)) + R_f^{(1)}(y;\theta;\tilde{\eta})(\eta(y) - \hat{\eta}(y))^2;$$

where $R_f^{(1)}$ is a first derivative of R_f w.r.t. η. This implies that

$$0 = m(\theta_0;\theta_0) = m(\theta_0;\theta_0) + \int_Y R_f(y;\theta;\hat{\eta})(\eta(y) - \hat{\eta}(y))\,d[y] + \int_Y R_f^{(1)}(y;\theta;\tilde{\eta})(\eta(y) - \hat{\eta}(y))^2\,d[y];$$

where $\tilde{\eta}$ is some unknown estimand within the interval $[\eta, \hat{\eta}]$. We obtain

$$\int_Y R_f^{(1)}(y;\theta;\tilde{\eta})(\eta(y) - \hat{\eta}(y))^2\,d[y] = m(\theta_0;\theta_0) + \int_Y R_f(y;\theta;\hat{\eta})(\eta(y) - \hat{\eta}(y))\,d[y]:$$

By taking expectations for both sides,

$$E_P \int_Y R_f^{(1)}(y;\ ;\ \tilde{}) \quad (y) \quad \hat{}(y)^{\ 2} d[y] \ = \ m(\ _0; b) + E_P \int_Y R_f(y;\ ;\ \hat{}) \quad (y) \quad \hat{}(y) \quad d[y] \ : \tag{B.30}$$

We have

$$\begin{aligned}
\int_Y R_f^{(1)}(y;\ ;\ \tilde{}) \quad (y) \quad \hat{}(y)^{\ 2} d[y] &= O \int_{Y} R_f^{(1)}(y;\ ;\ \tilde{}) \quad (y) \quad \hat{}(y)^{\ 2} d[y] \\
&= O \int_{Y} H(y) \quad (y) \quad \hat{}(y)^{\ 2} d[y] \\
&= O \int_{Y} H^2(y) \quad (y) \quad \hat{}(y)^{\ 2} d[y] \\
&= O \quad [H(Y)] \quad \hat{}[H(Y)]^{\ 2} \ ;
\end{aligned}$$

where the second equality is by the de nition of $H(y)$, the third by $H(y) < 1$, and the fourth by the de nition of L_2 norm.

This implies that

$$(B.28) = \quad E_P \int_Y R_f^{(1)}(y;\ ;\ \tilde{}) \quad \hat{}^{\ 2} d[y] \ = \ O \quad [H(Y)] \quad \hat{}[H(Y)]^{\ 2} \ ;$$

where the first equality is by Eq. (B.30) and the second equality is by the above.

Also, Sum of (B.19,B.21,B.23,B.25) in (B.29) can be written as follows:

Sum of (B.19,B.21,B.23,B.25)

$$\begin{aligned}
&= \sum_{z2f\ 0;1g} O_P \quad k\hat{}_z(W) \quad _z(W)k^n \quad \hat{}(x;z;W)[R_f(Y;\ _0;\hat{})] \quad (x;z;W)[R_f(Y;\ _0;\hat{})] \ + \ \hat{}_x(z;W) \quad _x(z;W)^o \\
&= \sum_{z2f\ 0;1g} O_P \quad k\hat{}_z(W) \quad _z(W)k^n \quad \hat{}(x;z;W)[H(Y)] \quad (x;z;W)[H(Y)] \ + \ \hat{}_x(z;W) \quad _x(z;W)^o \ :
\end{aligned}$$

For simplicity, we assume, for any x, z,

$$O_P \quad ^n \ _x(z;W) \quad \hat{}_x(z;W)^o \quad ^n \ _x(1 \quad z;W) \quad \hat{}_x(1 \quad z;W)^o \ = \sum_{z2f\ 0;1g} O_P \quad _x(z;W) \quad \hat{}_x(z;W)^{\ 2} \ ; \text{ and}$$

$$\begin{aligned}
&O_P \quad _x(z;W) \quad \hat{}_x(z;W) \quad (x;1 \quad z;W)[H(Y)] \quad \hat{}(x;1 \quad z;W)[H(Y)] \\
&= \sum_{z2f\ 0;1g} O_P \quad _x(z;W) \quad \hat{}_x(z;W) \quad (x;z;W)[H(Y)] \quad \hat{}(x;z;W)[H(Y)] \ :
\end{aligned}$$

The other part of Eq. (B.29) is given as

$$
\begin{aligned}
&E_P\left[\frac{X}{\hat{\sigma}_X}\left(1\{R_f(Y;\gamma_0;\hat{\eta})\}-\hat{m}[R_f(Y;\gamma_0;\hat{\eta})]\right)\right]\\
&=O_P\left[\frac{X}{\hat{\sigma}_X}\left(m[R_f(Y;\gamma_0;\hat{\eta})]-\hat{m}[R_f(Y;\gamma_0;\hat{\eta})]\right)\right]\\
&=O_P\left[\frac{X}{\hat{\sigma}_X}\left(\frac{\sigma_{YX}[R_f(Y;\gamma_0;\hat{\eta})]}{\sigma_X}-\frac{\hat{\sigma}_{YX}[R_f(Y;\gamma_0;\hat{\eta})]}{\sigma_X}+\frac{\hat{\sigma}_{YX}[R_f(Y;\gamma_0;\hat{\eta})]}{\sigma_X}-\frac{\hat{\sigma}_{YX}[R_f(Y;\gamma_0;\hat{\eta})]}{\hat{\sigma}_X}\right)\right]\\
&=O_P\left[\frac{X}{\sigma_X}\left(\sigma_{YX}[R_f(Y;\gamma_0;\hat{\eta})]-\hat{\sigma}_{YX}[R_f(Y;\gamma_0;\hat{\eta})]\right)\left(\frac{1}{\sigma_X}-\frac{1}{\hat{\sigma}_X}\right)\right]\\
&=O_P\left[\frac{X}{\sigma_X}\left(\sigma_{YX}[R_f(Y;\gamma_0;\hat{\eta})]-\hat{\sigma}_{YX}[R_f(Y;\gamma_0;\hat{\eta})]\right)\left(\frac{\hat{\sigma}_X}{\sigma_X}-1\right)\right]\\
&=O_P\left[\frac{X}{\sigma_X}\left(\frac{\hat{\sigma}_X}{\sigma_X}-1\right)^2\right]+O_P\left[\frac{X}{\sigma_X}\left(\sigma_{YX}[R_f(Y;\gamma_0;\hat{\eta})]-\hat{\sigma}_{YX}[R_f(Y;\gamma_0;\hat{\eta})]\right)\right]\\
&=O_P\left[\frac{X}{\sigma_X}\left(\frac{\hat{\sigma}_X}{\sigma_X}-1\right)^2\right]+O_P\left[\frac{X}{\sigma_X}\left(\sigma_{YX}[H(Y)]-\hat{\sigma}_{YX}[H(Y)]\right)\right]\\
&=\sum_z O_P\left[\hat{\sigma}_x(z;W)\left(\sigma_x(z;W)^2+\hat{\sigma}_x(z;W)\left(\sigma_x(z;W)-\hat{\sigma}_x(z;W)\right)\right)\left(\sigma(x;z;W)[H(Y)]-\hat{\sigma}(x;z;W)[H(Y)]\right)\right];
\end{aligned}
$$

593 where the equalities can be shown using the standard computation and the positivity assumption.

Similarly we assume, for any x, y, z,

$$
\begin{aligned}
&O_P\left[\sigma(x;z;W)[H(Y)]-\hat{\sigma}(x;z;W)[H(Y)]\right]\left[\sigma(x;1-z;W)[H(Y)]-\hat{\sigma}(x;1-z;W)[H(Y)]\right]\\
&=\sum_{z\in\{0,1\}} O_P\left[\left(\sigma(x;z;W)[H(Y)]-\hat{\sigma}(x;z;W)[H(Y)]\right)^2\right].
\end{aligned}
$$

We have

$$
\begin{aligned}
&O_P\left[\hat{\sigma}[H(Y)]\left(\sigma[H(Y)]\right)^2\right]\\
&=O_P\left[\hat{\sigma}_X[H(Y)]\left(\sigma_{YX}[H(Y)]+\hat{\sigma}_X\left(\sigma_X\right)^2\right)\right]\\
&=O_P\left[\hat{\sigma}_X[H(Y)]\left(\sigma_{YX}[H(Y)]\right)^2\right]+\hat{\sigma}_X\left(\sigma_X\right)^2+\left[\hat{\sigma}_X[H(Y)]\left(\sigma_{YX}[H(Y)]-\hat{\sigma}_X\sigma_X\right)\right]\\
&=\sum_{z\in\{0,1\}} O_P\left[\left(\sigma(x;z;W)[H(Y)]-\hat{\sigma}(x;z;W)[H(Y)]\right)^2\right]+\sum_{z\in\{0,1\}} O_P\left[\left(\sigma_x(z;W)-\hat{\sigma}_x(z;W)\right)^2\right]\\
&+\sum_{z\in\{0,1\}} O_P\left[\left(\sigma(x;z;W)[H(Y)]-\hat{\sigma}(x;z;W)[H(Y)]\right)\left(\sigma_x(z;W)-\hat{\sigma}_x(z;W)\right)\right].
\end{aligned}
$$

Finally

$$
\begin{aligned}
\text{Eq. (B.14)}=&\sum_z O_P\left[\left(\sigma_z(W)-\hat{\sigma}_z(W)\right)\left\{\left(\sigma(x;z;W)[H(Y)]-\hat{\sigma}(x;z;W)[H(Y)]\right)+\left(\sigma_x(z;W)-\hat{\sigma}_x(z;W)\right)\sigma_x(z;W)\right\}\right]\\
&+\sum_z O_P\left[\hat{\sigma}_x(z;W)\left(\sigma_x(z;W)-\hat{\sigma}_x(z;W)\right)^2+\left(\sigma(x;z;W)-\hat{\sigma}(x;z;W)\right)^2\right]\\
&+\sum_z O_P\left[\hat{\sigma}_x(z;W)\left(\sigma_x(z;W)-\hat{\sigma}_x(z;W)\right)\left(\sigma(x;z;W)-\hat{\sigma}(x;z;W)\right)\right]. \qquad \text{(B.31)}
\end{aligned}
$$

Therefore, with Eq. (B.13), the following holds

$$
\hat{b}-b_0=-M^{-1}E_D\left[\psi_m(V;\gamma_0;\sigma_0;\rho_0)\right]+\text{Eq. (B.31)}+o_P(n^{-1/2});
$$

594 where Eq. (B.31) $=R_2^m$.

595 □

596 Corollary 4 (Restated Corol. 4) If nuisances $\{\hat{\gamma},\hat{\sigma},\hat{\rho}\}$ converges at $n^{-1/4}$ rate, then the target
597 estimator $\hat{\theta}$ converges to θ_0 at $\sqrt{n}$-rate.

Proof. If all nuisances converge at $1/4$ rate, then the R_2^m term in Thm. 3 converges at $1/2$ rate. Also, $E_D[\ m(\ 0;(\ 0;\ 0))]$ converges in distribution to $N(0, \mathrm{var}(\ m(\ 0;(\ 0;\ 0))))$ at $\sqrt{n}$-rate. So $\hat{}$ converges to $_0$ at $\sqrt{n}$-rate by Thm. 3. □

C Details of empirical applications

C.1 Data generating processes for synthetic datasets

The following structural equations are used for all four data generating processes in Fig. 2:

$$
\begin{aligned}
&U \quad N(0,1)\\
&f_W(U) = 2\,U \quad 1 + \ _W;\ \text{where}\ _W \quad N(0,1)\\
&f_Z(W) = \ 1(0.25W + \ _Z > 0);\ \text{where}\ _Z \quad N(0,1)\\
&f_X(W,Z,U) = \ 1(Z + 0.25 \quad W + 0.25 \quad U + \ _X > 0.5) \quad (1 \quad Z) + Z;\ \text{where}\ _X \quad N(0,1).
\end{aligned}
$$

With such data generating process $X_{Z=1} \quad X_{Z=0}$ is satisfied. We will denote four figures in Fig. 2 as Fig. 2(a,b,c,d). For Fig. 2a,

$$f_Y(W,X,U) = 0.6501(W \quad (2X \quad 1) + 2\,U + 0.374).$$

For Fig. 2b,

$$f_Y(W,X,U) = 0.9515(2X \quad 1 + W) + 0.8(\ 2X + 1 + U) + WU + 0.082$$

For Fig. 2c,

$$
\begin{aligned}
f_Y(W,X,U) = {} & 1.0854\ 1(W < 0)(2X \quad 1 + 0.1U) + \ 1(0 \quad W < 1)(\ 2X + 1 + 0.1U)\\
& + 1.0854 \quad 0.9163(1(W \quad 1)(\ 3(2X \quad 1) + 0.2U + 0.3) \quad 0.122)
\end{aligned}
$$

For Fig. 2d,

$$
\begin{aligned}
f_Y(W,X,U) = {} & 0.7865 \quad 1.0628 \ 1(W < \quad 1)(\ 0.8(2X \quad 1) + 0.1U) + \ 1(\ 1 \quad W < 0)(\ 2(2X \quad 1) + 0.1U)\\
& + 0.7865 \quad 1.0628 \ (1(0 \quad W < 1)(2(2X \quad 1) + 0.2U) + \ 1(W > 1)(0.5(2X \quad 1) + 0.2U) + 0.0525)\\
& + 1.0628 \quad 0.104
\end{aligned}
$$

C.2 Application to 401(k) data

We use the 401(k) dataset that is initially introduced by [21]. Specifically, we used the version of the data named `The Woodridge Data Set' [37] [originally entitled `401ksu.dta' in STATA format (available in https://www.stata.com/texts/eacsap/). In the dataset, we used nettfa (net financial asset in \$1000) as Y, p401k (participation in 401(k), participation = 1) as X, e401k (eligibility for 401(k), eligible = 1) as Z, and $W = \{W_1, W_2, W_3, W_4, W_5\} = \{$agesq, fsize, male, marr, incsq$\}$, where agesq means the square of the age, fsize the family size, male the gender (male = 1), marr the marital status (married = 1) and incsq the square of the income.

C.3 Density plots illustrating uncertainty

In this section, we present the density plots corresponding to Figs. (3,4) illustrating uncertainty of the results. The same data generating processes as used for Figs. (3,4) are leveraged.

Synthetic dataset. To represent the uncertainty, we generate 100 synthetic datasets $\{D_k\}_{k=1}^{100}$, each of which has $N = 50000$ samples (i.e., $|D_k| = 50000$), from the same data generating process used for the simulation for Fig. 3. After learning the density estimation with D_k, we obtain a vector of density values $(p_1^k, p_2^k, \ldots, p_m^k)$ at m equi-spaced points for each method (`Moment', 'MLTE', `Kernel-smoothing', 'KLTE'). For the model-based approach (Moment, MLTE), m is set to 1000. For the kernel-based approach (Kernel-smoothing, KLTE), m is set to 25. For each {model, kernel}-based approach, we have estimates of the density in the form of a matrix $(p_1^k, \ldots, p_m^k)_{k=1}^{100}$. Let $p_{avg,i}$ denote the average of $\{p_i^k\}_{k=1}^{100}$. Let $_i$ denote the standard deviation of $\{p_i^k\}_{k=1}^{100}$. Then, we take

$$
\begin{aligned}
p_{\text{avg}} \quad & \{p_{\text{avg},i}\}_{i=1}^m\\
p_{\text{upper}} \quad & \{p_{\text{avg},i} + \ _i\}_{i=1}^m\\
p_{\text{lower}} \quad & \{p_{\text{avg},i} \quad \ _i\}_{i=1}^m.
\end{aligned}
$$

A set of density plots corresponding to Fig. 3 with uncertainty information in Fig. C.5. For each density estimate in Fig. C.5, the middle dark-colored dotted line shows p_{avg}, and the above,below light-colored solid line shows $\mathsf{p}_{\text{upper}}$, $\mathsf{p}_{\text{lower}}$ respectively.

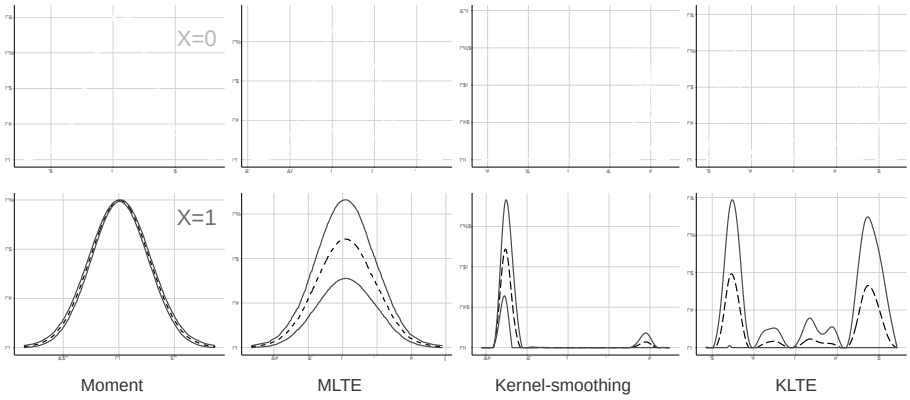

Figure C.5: LTE estimation with a synthetic dataset. The middle dark-colored dotted line denotes p_{avg}, and the upper and lower light-colored solid lines represents $\mathsf{p}_{\text{upper}}$, $\mathsf{p}_{\text{lower}}$, respectively.

Application to 401(k) data. To represent the uncertainty, we randomly resample (with replacement) the dataset from the original dataset D, where the kth regenerated dataset is denoted D_k. We conducted this data regeneration process for 100 times and have $fD_k g_{k=1}^{100}$. After learning the density estimation with D_k, we obtain a vector of density values $(p_1^k, p_2^k, \quad, p_m^k)$ at m equi-spaced points for each method ('Moment', 'MLTE', 'Kernel-smoothing', 'KLTE'). For the model-based approach (Moment, MLTE), m is set to 1000. For the kernel-based approach (Kernel-smoothing, KLTE), m is set to 25. For each {model, kernel}-based approach, we have estimates of the density in the form of a matrix $f(p_1^k, \quad, p_m^k) g_{k=1}^{100}$. Let $p_{\text{avg};i}$ denote the average of $fp_i^k g_{k=1}^{100}$. Let σ_i denote the standard deviation of $fp_i^k g_{k=1}^{100}$. Then, we take $\mathsf{p}_{\text{avg}} \quad fp_{\text{avg};i} g_{i=1}^m$, $\mathsf{p}_{\text{upper}} \quad fp_{\text{avg};i} + \sigma_i g_{i=1}^m$ and $\mathsf{p}_{\text{lower}} \quad fp_{\text{avg};i} \quad \sigma_i g_{i=1}^m$.

A set of density plots corresponding to Fig. 3 with uncertainty information in Fig. C.6. For each density estimate in Fig. C.6, the middle dark-colored dotted line shows p_{avg}, and the above,below light-colored solid line shows $\mathsf{p}_{\text{upper}}$, $\mathsf{p}_{\text{lower}}$ respectively.

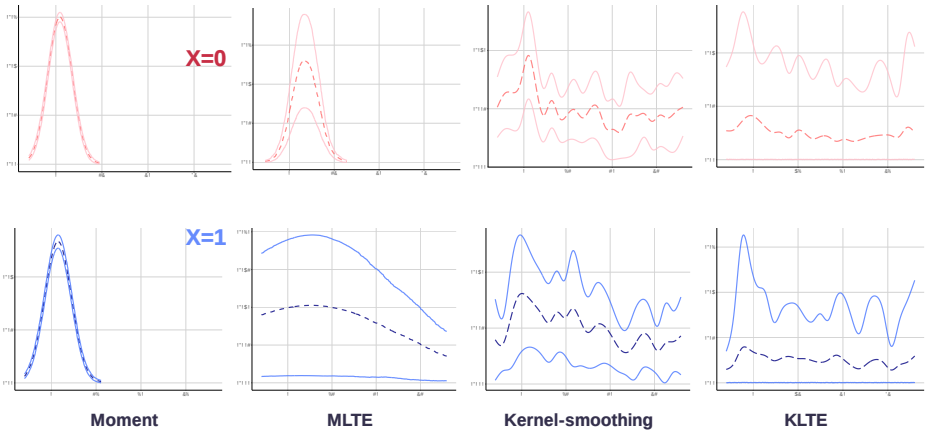

Figure C.6: LTE of 401(k) participation (X) on net financial asset (Y). The middle dark-colored dotted line denotes p_{avg}, and the upper and lower light-colored solid lines represents $\mathsf{p}_{\text{upper}}$, $\mathsf{p}_{\text{lower}}$, respectively.