cemmap
centre for microdata methods and practice

Double debiased machine learning nonparametric inference with continuous treatments

Kyle Colangelo
Ying-Ying Lee

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP54/19

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments*

Kyle Colangelo Ying-Ying Lee†

University of California Irvine

October 2019

Abstract

We propose a nonparametric inference method for causal effects of continuous treatment variables, under unconfoundedness and in the presence of high-dimensional or nonparametric nuisance parameters. Our simple kernel-based double debiased machine learning (DML) estimators for the average dose-response function (or the average structural function) and the partial effects are asymptotically normal with a nonparametric convergence rate. The nuisance estimators for the conditional expectation function and the generalized propensity score can be nonparametric kernel or series estimators or ML methods. Using doubly robust influence function and cross-fitting, we give tractable primitive conditions under which the nuisance estimators do not affect the first-order large sample distribution of the DML estimators.

Keywords: Average structural function, continuous treatment, cross-fitting, dose-response function, double debiased machine learning, doubly robust, high dimension, nonseparable models, partial mean, post-selection inference.

JEL Classification: C14, C21, C55

*The first version was circulated as "Double machine learning nonparametric inference on continuous treatment effects" (February 2019). We are grateful to Max Farrell, Whitney Newey, and Takuya Ura for valuable discussion. We also thank conference participants in 2019: Barcelona Summer Forum workshop on Machine Learning for Economics, North American Summer Meeting of the Econometric Society, Vanderbilt/CeMMAP/UCL conference on Advances in Econometrics, and the Midwest Econometrics Group.

†Department of economics, 3151 Social Science Plaza, University of California Irvine, Irvine, CA 92697. E-mail: yingying.lee@uci.edu

1 Introduction

We propose a nonparametric inference method for *continuous* treatment effects on the outcome Y, under the unconfoundedness assumption[1] and in the presence of high-dimensional or nonparametric nuisance parameters. We focus on the heterogenous effect with respect to the continuous treatment or policy variables T. To identify the causal effects, it is plausible to allow the number of the control variables X to be large relative to the sample size n. To obtain precise estimation, it is useful to include control variables that account for residual variation. To achieve valid inference and to estimate nuisance parameters by machine learning (ML) methods, we employ a double debiased ML approach using doubly robust influence function and cross-fitting. Our work builds upon the results for semiparametric problems in Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018) and Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) and extends the literature to nonparametric continuous treatment effects.

We show that the proposed estimator is asymptotically normal and converges at a nonparametric rate. Such asymptotic theory is fundamental for inference, such as constructing confidence intervals and testing hypotheses. We provide tractable conditions under which the nuisance estimators do not affect the first-order asymptotic distribution of the final double debiased ML estimator. Thus the nuisance estimators of the conditional expectation function $\mathbb{E}[Y|T,X]$ and the conditional density (or generalized propensity score) $f_{T|X}$ can be conventional nonparametric estimators, such as kernels or series, as well as modern ML methods, such as lasso or deep neural nets (see, e.g., Athey and Imbens (2019) for potential methods, such as ridge, boosted trees, random forest, and various ensembles of these methods).

We consider the outcome equation to be fully nonparametric $Y = g(T,X,\varepsilon)$. No functional form assumption is imposed on the general disturbances ε, like monotonicity, dimensionality, or separability. The potential outcome is $Y(t) = g(t,X,\varepsilon)$ indexed by the hypothetical treatment value t. The causal object of interest is the *average dose-response function* as a function of t, defined by the mean of the potential outcome across observations with the observed and unobserved heterogeneity (X,ε), i.e., $\beta_t = \mathbb{E}[Y(t)] = \int\int g(t,X,\varepsilon)dF_{X\varepsilon}$. It is also known as the *average structural function* in nonseparable models in Blundell and Powell (2003). We further define the marginal or partial effect of the first element of the continuous treatment T at $t = (t_1, ...t_{d_t})'$ to be $\theta_t \equiv \partial\beta_t/\partial t_1$. In an example of demand analysis when T contains price and income, β_t can be the Engel curve. The partial effect θ_t can reveal the average price elasticity at given values of price

[1]This commonly used identifying assumption, also known as conditional independence or selection on observables, assumes that conditional on a set of observables X, the treatment T is independent of the unobservable disturbances in the outcome equation. In other words, T is conditionally exogenous, or as good as randomly assigned.

and income and hence captures the unrestricted heterogenous effects.

We introduce a *doubly robust* estimator for continuous treatments, defined by

$$\hat{\beta}_t^{DR} = \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}(t, X_i) + \frac{K_h(T_i - t)}{\hat{f}_{T|X}(t|X_i)}\big(Y_i - \hat{\gamma}(t, X_i)\big), \tag{1}$$

where $\hat{\gamma}(t,x)$ is an estimator of $\gamma(t,x) \equiv \mathbb{E}[Y|T=t, X=x]$, $\hat{f}_{T|X}(t|x)$ is an estimator of $f_{T|X}(t|x)$, $K_h(T_i - t) = \Pi_{j=1}^{d_t} k((T_{ji} - t_j)/h)/h$ is a product kernel with a standard second-order kernel function $k()$ and bandwidth h that is a positive sequence vanishing as n grows. Based on $\hat{\beta}_t^{DR}$, we propose a *double debiased machine learning* (DML) estimator where $\hat{\gamma}$ and $\hat{\lambda}$ use cross-fitting and ML or traditional nonparametric methods. Then we estimate the partial effect θ_t by a numerical differentiation. We also propose an estimator for the conditional density $f_{T|X}(t|x)$ for the low-dimensional T and high-dimensional X, which could be of independent interest.

Our estimators use doubly robust influence function and cross-fitting, inspired by the DML method in semiparametric problems in Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) (CCDDHNR, hereafter). The doubly robust influence function can reduce sensitivity in estimating β_t with respect to nuisance parameters. In particular, Neyman orthogonality holds for the moment function in the estimator $\hat{\beta}_t^{DR}$ in (1) as $h \to 0$. Using cross-fitting via sample-splitting can remove bias induced by overfitting and achieve stochastic equicontinuity without strong entropy condition. In particular CCDDHNR point out that the commonly used results in empirical process theory, such as Donsker properties, can break down in the high-dimensional setting. For example, the nuisance lasso estimator may require strong sparsity assumption.

It is useful to note that the doubly robust estimator for a binary/multivalued treatment replaces the kernel $K_h(T_i - t)$ with the indicator function $\mathbf{1}\{T_i = t\}$ in equation (1) and has been widely-studied, especially in the recent growing ML literature.[2] We show that the advantageous properties of the DML estimator for the binary treatment carry over to the continuous treatments case. Moreover, our primitive condition on the convergence rates of the nuisance parameters can be weaker due to the bandwidth h in our nonparametric DML estimator. We further make novel observations of unique features of continuous treatments: First we motivate the kernel-based

[2]Our estimator is doubly robust in the sense that the causal effect remains identified and consistently estimated if either one of the nuisance functions $\mathbb{E}[Y|T,X]$ or $f_{T|X}$ is misspecified. The recent ML literature has been utilizing this doubly robust property to reduce regularization and modeling biases in estimating the nuisance parameters by ML or nonparametric methods; for example, Belloni, Chernozhukov, and Hansen (2014), Farrell (2015), Belloni, Chernozhukov, Fernández-Val, and Hansen (2017), Farrell, Liang, and Misra (2018), Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018), CCDDHNR, Rothe and Firpo (2018), and references therein. The benefit of cross-fitting is further investigated by Wager and Athey (2018) for heterogeneous causal effects, Newey and Robins (2018) for double cross-fitting, and Cattaneo and Jansson (2019) for cross-fitting bootstrap.

moment function in $\hat{\beta}_t^{DR}$ by the limit of the Gateaux derivative, as in Ichimura and Newey (2017) and Carone, Luedtke, and van der Laan (2018). The kernel function in $\hat{\beta}_t^{DR}$ is a natural choice to approximate the distribution of a point mass. Further the kernel function provides a simple moment function for the *partial mean* structure of β_t that fixes T at t and averages out X (Newey, 1994; Lee, 2018). A second motivation of the moment function is adding to the influence function of the regression estimator $n^{-1}\sum_{i=1}^{n}\hat{\gamma}(t, X_i)$ the adjustment term from a kernel-based estimator $\hat{\gamma}$. A series estimator $\hat{\gamma}$ would yield a different adjustment. These two distinct features of continuous treatments are in contrast to the binary treatments case, where different nonparametric estimators of γ result in the same efficient influence function.

The main contribution of this paper is a formal inference theory for the fully nonparametric causal effects of continuous variables, allowing for high-dimensional nuisance parameters. To uncover the causal effect of the continuous variable T on Y, our nonparametric model $Y = g(T, X, \varepsilon)$ is compared to the partial linear model $Y = \theta T + g(X) + \varepsilon$ in Robinson (1988) that specifies the homogenous effect by θ and hence is a semiparametric problem. The important partial linear model has many applications and is one of the leading examples in Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018), CCDDHNR, and references therein, where the nuisance function $g(X)$ allows for high-dimensional X and can be estimated by a ML method. Demirer, Syrgkanis, Lewis, and Chernozhukov (2019) and Oprescu, Syrgkanis, and Wu (2019) extend to more general functional forms. In contrast, our average structural function β_t and the partial effect θ_t capture the fully nonparametric heterogenous effect of T. Our simple estimator utilizes the kernel function $K_h(T_i - t)$ for the low-dimensional continuous treatments T and averages out the high-dimensional covariates X, so we can maintain the nonparametric feature and circumvent the complexity of the nuisance parameter space.

To the best of our knowledge, we are among the first to apply the double debiased ML approach for inference on the average structural function and the partial effect of continuous treatments. There is a small yet growing literature on employing the DML approach for objects that cannot be estimated at the regular root-n rate. For example, the conditional average binary treatment effect $\mathbb{E}[Y(1) - Y(0)|X_1]$ for a low-dimensional subset $X_1 \subset X$ is studied in Chernozhukov, Newey, Robins, and Singh (2019), Chernozhukov and Semenova (2019), Fan, Hsu, Lieli, and Zhang (2019), and Zimmert and Lechner (2019). Their setups do not cover our average structural function and the partial effect of continuous treatments. The causal objects of interest are different.[3]

[3]In particular, Chernozhukov, Newey, Robins, and Singh (2019) provide sparse regression methods for non-regular linear functionals of the conditional expectation function, such as $\mathbb{E}[m(Z, \gamma(T, X))|T = t]$ where $\gamma \mapsto m(z, \gamma)$ is a linear operator for each $z = (y, t, x)$. For a simple example that $m(z, \gamma) = \gamma$, their perfectly localized functional $\lim_{h\to 0}\int\int \gamma(T, X)K_h(T - t)/\mathbb{E}[K_h(T - t)]dF_{TX}(T, X) = \int \gamma(t, X)dF_{X|T}(X|t) = \mathbb{E}[\gamma(t, X)|T = t] = \mathbb{E}[Y(t)|T = t]$, while we identify the average structural function $\beta_t = \mathbb{E}[Y(t)]$ by $\lim_{h\to 0}\int\int \gamma(T, X)K_h(T -$

We also contribute to the literature of continuous treatment effects estimation. In high-dimensional settings, Su, Ura, and Zhang (2019) propose a doubly robust estimator $\hat{\beta}_t^{DR}$ as in equation (1). Assuming approximate sparsity, they use lasso-type estimators $\hat{\gamma}(t,\cdot)$ and $\hat{f}_{T|X}(t|\cdot)$ to select the high-dimensional covariates X via a localized method of L_1-penalization at each t. In contrast, we estimate $\gamma(t,x)$ and $f_{T|X}(t|x)$ with cross-fitting and provide high-level conditions that allow us to use a variety of nonparametric and ML methods with weaker assumptions. Kennedy, Ma, McHugh, and Small (2017) and Kallus and Zhou (2018) propose some versions of the doubly robust estimators.[4] In low-dimensional settings, see Hirano and Imbens (2004), Flores (2007), and Lee (2018) for examples of a class of regression estimators $n^{-1}\sum_{i=1}^n \hat{\gamma}(t,X_i)$. Galvao and Wang (2015) and Hsu, Huber, Lee, and Pipoz (2018) study a class of inverse probability weighting estimators. For examples of empirical applications, see Flores, Flores-Lagunes, Gonzalez, and Neumann (2012) and Kluve, Schneider, Uhlendorff, and Zhao (2012). We extend this literature to high-dimensional settings enabling ML methods for inference.

The paper proceeds as follows. We introduce the framework and estimation procedure in Section 2. Section 3 presents the asymptotic theory. All the proofs are in the Appendix.

2 Setup and estimation

We give identification assumptions and introduce the double debiased ML (DML) estimator.

Assumption 1 *Let* $\{Y_i, T_i', X_i'\}_{i=1}^n$ *be an i.i.d. sample from* $Z = \{Y, T', X'\}' \in \mathcal{Z} = \mathcal{Y} \times \mathcal{T} \times \mathcal{X} \subseteq \mathcal{R}^{1+d_t+d_x}$ *and* $Y = g(T, X, \varepsilon)$. *(i)* **(Conditional independence)** *Conditional on* X, T *and* ε *are independent, or equivalently* T *and* $Y(t) = g(t, X, \varepsilon)$ *are independent for any* $t \in \mathcal{T}$. *(ii)* **(Common support)** *For any* $t \in \mathcal{T}$ *and* $x \in \mathcal{X}$, $f_{T|X}(t|x)$ *is bounded away from zero.*

Assumption 2 (Kernel) *The second-order symmetric kernel function* $k()$ *is bounded differentiable and has a convex bounded support.*

By Assumptions 1-2 and the same reasoning for the binary treatment, it is straightforward to show the identification of $\beta_t = \mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y|T=t,X]] = \lim_{h\to 0} \mathbb{E}\left[K_h(T-t)Y/f_{T|X}(t|X)\right]$.

Estimation procedure

$t)/f_{T|X}(t|X)dF_{TX}(T,X) = \int \gamma(t,X)dF_X(X) = \mathbb{E}[\gamma(t,X)]$.

[4] Kallus and Zhou (2018) use a known $f_{T|X}$. Kennedy, Ma, McHugh, and Small (2017) construct a "pseudo-outcome" based on the doubly robust mapping, by plugging in the nuisance estimators under high-level assumptions. In particular, their efficient influence function uses $f_T(T)$ rather than a kernel function $K_h(T-t)$. Then they regress the pseudo-outcome on the treatment variable using ML or nonparametric methods.

Step 1. (Cross-fitting) For some $L \in \{2, ..., n\}$, partition the observation indices into L groups I_l, $l = 1, ..., L$. For each $l = 1, ..., L$, the estimators $\hat{\gamma}_l(t, x)$ for $\gamma(t, x)$ and $\hat{f}_l(t|x)$ for $f_{T|X}(t|x)$ use observations not in I_l and satisfy Assumption 3 below.

Step 2. (Doubly robust) The double debiased ML (DML) estimator is defined as

$$\hat{\beta}_t = \frac{1}{n} \sum_{l=1}^{L} \sum_{i \in I_l} \hat{\gamma}_l(t, X_i) + \frac{K_h(T_i - t)}{\hat{f}_l(t|X_i)} \left(Y_i - \hat{\gamma}_l(t, X_i)\right). \tag{2}$$

Step 3. (Partial effect) Let $t^+ = (t_1 + \eta/2, t_2, ..., t_{d_t})'$ and $t^- = (t_1 - \eta/2, t_2, ..., t_{d_t})'$, where η is a positive sequence converging to zero as $n \to \infty$. Define the estimator to be $\hat{\theta}_t = (\hat{\beta}_{t^+} - \hat{\beta}_{t^-})/\eta$.

When $L = n$, $\hat{\beta}_t$ is known as the leave-one-out estimator. When there is no sample splitting $L = 1$, $\hat{\gamma}_1$ and $\hat{\lambda}_1$ use all observations in the full sample. Then the DML estimator $\hat{\beta}_t$ in (2) is the doubly robust estimator $\hat{\beta}_t^{DR}$ in equation (1). When the dependent variable Y is continuous, we can further estimate the distributional effects by replacing Y with $\mathbf{1}\{Y \leq y\}$ in our procedure.

Denote the L_2-norm $\|\hat{f}_l - f_{T|X}\|_{L_2} \equiv \left(\int_{\mathcal{T}} \int_{\mathcal{X}} \left(\hat{f}_l(t|x) - f_{T|X}(t|x)\right)^2 f_{TX}(t, x) dx dt \right)^{1/2}$ and $\|\hat{\gamma}_l - \gamma\|_{L_2} \equiv \left(\int_{\mathcal{T}} \int_{\mathcal{X}} \left(\hat{\gamma}_l(t, x) - \gamma(t, x)\right)^2 f_{TX}(t, x) dx dt \right)^{1/2}$ for each $l = 1, ..., L$.

Assumption 3 (Nuisance estimators) *For each $l = 1, ..., L$, (i) $\|\hat{\gamma}_l - \gamma\|_{L_2} \xrightarrow{p} 0$, $\|\hat{f}_l - f_{T|X}\|_{L_2} \xrightarrow{p} 0$; (ii) $\sqrt{nh^{d_t}} \|\hat{\gamma}_l - \gamma\|_{L_2} \|\hat{f}_l - f_{T|X}\|_{L_2} \xrightarrow{p} 0$.*

In Assumption 3, (i) requires mean square consistency of the first step estimator $\hat{\gamma}$ and $\hat{f}_{T|X}$. The only convergence rate condition is in (ii) that requires the product of estimation errors for the two estimators to vanish fast than $1/\sqrt{nh^{d_t}}$ that is slower than $1/\sqrt{n}$ in the semiparametric problem. The convergence rates in Assumption 3 are available for kernel or series estimators, the deep neural network in Farrell, Liang, and Misra (2018), the lasso in Belloni, Chernozhukov, and Hansen (2014), Farrell (2015), Su, Ura, and Zhang (2019), for example. The numerical differentiation estimator $\hat{\theta}_t$ is simple and circumvents estimating the derivatives of the nuisance parameters.

2.1 Conditional density estimation

We can estimate the generalized propensity score (GPS) $f_{T|X}$ by a kernel density estimator. The ML method for estimating the conditional density is less developed comparing with estimating the conditional mean. We propose an estimator that allows us to use the ML methods designed

for the conditional mean. Let $\hat{\mathbb{E}}[W|X]$ be an estimator of the conditional mean $\mathbb{E}[W|X]$ for a bounded random variable W. Suppose the convergence rate is available, $\left\|\hat{\mathbb{E}}[W|X]-\mathbb{E}[W|X]\right\|_\infty \equiv \sup_{x\in\mathcal{X}}\left|\hat{\mathbb{E}}[W|X=x]-\mathbb{E}[W|X=x]\right| = O_p(R_1)$; for example, the deep neural network in Farrell, Liang, and Misra (2018). Let $G(u)=\int_{-\infty}^{u} g(v)dv$ with a standard second-order kernel function $g()$. When $d_T=1$, the CDF estimator $\hat{F}_{T|X}(t|x)=\hat{\mathbb{E}}[G((t-T)/h_1)]$, where h_1 and ϵ are positive sequences vanishing as n grows. Then we estimate the GPS by the numerical derivative estimator $\hat{f}_{T|X}(t|x)=(2\epsilon)^{-1}\big(\hat{F}_{T|X}(t+\epsilon|x)-\hat{F}_{T|X}(t-\epsilon|x)\big)$. Lemma 2.1 shows that the convergence rate $\left\|\hat{f}_{T|X}-f_{T|X}\right\|_\infty = O_p(R_1\epsilon^{-1}+h_1^2\epsilon^{-1}+\epsilon^2)$. Similar numerical derivative approaches have been used in Belloni, Chernozhukov, Fernández-Val, and Hansen (2017) and Su, Ura, and Zhang (2019).[5]

When T is multi-dimensional, let $G((t-T)/h_1)=\Pi_{j=1}^{d_t}G((t_j-T_j)/h_1)$. We illustrate the GPS estimator for $d_T=2$. The general GPS estimator for $d_T>2$ can be implemented by the same procedure. The estimator of the partial derivative of $F_{T|X}(t_1,t_2|x)$ with respect to t_1 is $\widehat{\frac{\partial F_{T|X}}{\partial t_1}}(t_1,t_2|x)=\big(\hat{F}_{T|X}(t_1+\epsilon,t_2|x)-\hat{F}_{T|X}(t_1-\epsilon,t_2|x)\big)/(2\epsilon)$. Then the GPS estimator for $d_T=2$ is

$$\begin{aligned}\hat{f}_{T|X}(t|x)=\widehat{\frac{\partial^2 F_{T|X}}{\partial t_2\partial t_1}}(t|x) &= \left(\widehat{\frac{\partial F_{T|X}}{\partial t_1}}(t_1,t_2+\epsilon|x)-\widehat{\frac{\partial F_{T|X}}{\partial t_1}}(t_1,t_2-\epsilon|x)\right)\frac{1}{2\epsilon}\\ &= \Big(\hat{F}_{T|X}(t_1+\epsilon,t_2+\epsilon|x)-\hat{F}_{T|X}(t_1-\epsilon,t_2+\epsilon|x)\\ &\quad -\hat{F}_{T|X}(t_1+\epsilon,t_2-\epsilon|x)+\hat{F}_{T|X}(t_1-\epsilon,t_2-\epsilon|x)\Big)\frac{1}{4\epsilon^2}.\end{aligned}$$

Lemma 2.1 (GPS) *Let $f_{T|X}(t|x)$ be (d_T+1)-times differentiable with respect to t for any $x\in\mathcal{X}$. Then $\left\|\hat{f}_{T|X}-f_{T|X}\right\|_\infty = O_p(R_1\epsilon^{-d_T}+h_1^2\epsilon^{-d_T}+\epsilon^2)$.*

Alternatively, we may use a mixture density network in Hartford, Lewis, Leyton-Brown, and Taddy (2017) and Bishop (2006) that model $f_{T|X}$ as a mixture of Gaussian distributions. There is some recent development for random forest, e.g., Athey and Wager (2019), Pospisil and Lee (2018) and Criminisi, Shotton, and Konukoglu (2012).

3 Asymptotic theory

We first derive the asymptotic linear representation and normality for $\hat{\beta}_t$, showing that the nuisance estimators have no first-order effect. Then we discuss the construction of the doubly robust moment function by Gateaux derivative in Section 3.1. In Section 3.2, we discuss the adjustment

[5] Su, Ura, and Zhang (2019) use Lasso to approximate $F_{T|X}$ by a Logistic CDF. In contrast, we suggest a kernel g to smooth CDF estimates (such as a Gaussian kernel). And we allow for various nonparametric and ML methods.

for the first-step kernel estimators in the influence functions of the regression estimator and inverse probability weighting estimator that do not use the doubly robust moment function and cross-fitting. We illustrate how the DML estimator assumes weaker conditions. Section 3.3 gives a heuristic overview of deriving the asymptotic linear representation.

Theorem 1 (Asymptotic normality) *Let Assumptions 1-3 hold. Let* $h \to 0$, $nh^{d_t} \to \infty$, *and* $nh^{d_t+4} \to C \in [0, \infty)$. *Assume that for* $(y, t', x')' \in \mathcal{Z}$, $f_{YTX}(y, t, x)$ *is three-times differentiable with respect to* t, *and* $var(Y|T = t, X = x) f_{T|X}(t|x)$ *is bounded above uniformly over* $x \in \mathcal{X}$. *Then for any* t *in the interior of* $\mathcal{T}$,

$$\sqrt{nh^{d_t}} \left(\hat{\beta}_t - \beta_t \right) = \sqrt{\frac{h^{d_t}}{n}} \sum_{i=1}^{n} \left\{ \frac{K_h(T_i - t)}{f_{T|X}(t|X_i)} (Y_i - \mathbb{E}[Y|T = t, X = X_i]) \right. \\ \left. + \mathbb{E}[Y|T = t, X = X_i] - \beta_t \right\} + o_p(1) \tag{3}$$

and $\sqrt{nh^{d_t}} \left(\hat{\beta}_t - \beta_t - h^2 \mathsf{B}_t \right) \xrightarrow{d} \mathcal{N}(0, \mathsf{V}_t)$, *where* $\mathsf{V}_t \equiv \mathbb{E}\left[var[Y|T = t, X]/f_{T|X}(t|X) \right] \int_{-\infty}^{\infty} k(u)^2 du$ *and* $\mathsf{B}_t \equiv \sum_{j=1}^{d_t} \mathbb{E}\left[\frac{1}{2} \frac{\partial^2}{\partial t_j^2} \mathbb{E}\left[Y|T = t, X \right] + \frac{\partial}{\partial t_j} \mathbb{E}\left[Y|T = t, X \right] \frac{\partial}{\partial t_j} f_{T|X}(t|X)/f_{T|X}(t|X) \right] \int_{-\infty}^{\infty} u^2 k(u) du$.

Theorem 1 is fundamental for inference, such as constructing confidence intervals and the optimal bandwidth h that minimizes the asymptotic mean squared error. Theorem 1 can be generalized to provide inference that is uniformly valid over the classes of data-generating processes for which our assumptions hold uniformly, following Theorem 3.1 in CCDDHNR, for example, at the cost of notational complication.

Note that the second part in the influence function in (3) $n^{-1} \sum_{i=1}^{n} \mathbb{E}[Y|T = t, X = X_i] - \beta_t = O_p(1/\sqrt{n}) = o_p(1/\sqrt{nh^{d_t}})$ and hence does not contribute to the first-order asymptotic variance V_t. We keep these terms to show that the nuisance estimators do not affect the first-order asymptotic linear representation. This is in contrast to the binary treatment case, where $K_h(T_i - t)$ is replaced by $\mathbf{1}\{T_i - t\}$ and $\hat{\beta}_t$ converges at a root-n rate, so this second part in (3) is of first-order. Then we obtain the well-studied efficient influence function in estimating the binary treatment effect in Hahn (1998). For the continuous treatment case here, it is crucial to include this adjustment term in the moment function in $\hat{\beta}_t$ to achieve double robustness.

There is a bias term arising from the kernel function $K_h(T - t)$. We may estimate the leading bias $h^2 \mathsf{B}_t$ by the sample analogue. We can estimate the asymptotic variance V_t by the sample variance of the influence function (3). Specifically $\hat{\mathsf{V}}_t = h^{d_t} n^{-1} \sum_{l=1}^{L} \sum_{i \in I_l} \hat{\psi}_{li}^2$, where the estimated influence function $\hat{\psi}_{li} = K_h(T_i - t)(Y_i - \hat{\gamma}_l(t, X_i))/\hat{f}_l(t|X_i) + \hat{\gamma}_l(t, X_i) - \hat{\beta}_t$. Then we can estimate

the optimal bandwidth that minimizes the asymptotic mean squared error given in the following corollary.

Corollary 1 (AMSE optimal bandwidth) *Let the conditions in Theorem 1 hold. If* B_t *is non-zero, then the bandwidth that minimizes the asymptotic mean squared error is* $h^* = \big(d_t \mathsf{V}_t / \big(4\mathsf{B}_t^2\big)\big)^{1/(d_t+4)} n^{-1/(d_t+4)}$.

By choosing an undersmoothing bandwidth h smaller than h^*, the bias is first-order asymptotically negligible, i.e., $h^2\sqrt{nh^{d_t}} \to 0$. Then we can construct the usual $(1-\alpha)\times 100\%$ confidence interval $\Big[\hat{\beta}_t \pm \Phi(1-\alpha/2)\sqrt{\hat{\mathsf{V}}_t/\big(nh^{d_t}\big)}\Big]$. Alternatively, we may consider a further bias correction following Calonico, Cattaneo, and Farrell (2018) to allow for a wider range of bandwidth choice. Such robust bias-corrected inference is beyond the scope of the current paper and left for future research.

Next we present the asymptotic theory for $\hat{\theta}_t$. We consider two conditions for the tuning parameter η via $\eta/h \to \rho$ for (i) $\rho = 0$ and (ii) $\rho \in (0,\infty]$. Let $\partial_t^\nu \equiv \partial^\nu g(t,\cdot)/\partial t^\nu$ denote the ν-th order partial derivative of a generic function g with respect to t.

Theorem 2 (Asymptotic normality - Partial effect) *Let the conditions in Theorem 1 hold. Assume that for* $(y,t',x')' \in \mathcal{Z}$, $f_{YTX}(y,t,x)$ *is four-times differentiable with respect to* t, *and* β_t *is twice differentiable.*

(i) Let $\eta/h \to 0$, $nh^{d_t+2} \to \infty$, *and* $nh^{d_t+2}\eta^2 \to 0$. *Assume (a)* $\eta^{-1}h\|\hat{\gamma}_l - \gamma\|_{L_2} \xrightarrow{p} 0$, $\eta^{-1}h\|\hat{f}_l - f_{T|X}\|_{L_2} \xrightarrow{p} 0$; *(b)* $\eta^{-1}h\sqrt{nh^{d_t}}\|\hat{f}_l - f_{T|X}\|_{L_2}\|\hat{\gamma}_l - \gamma\|_{L_2} \xrightarrow{p} 0$. *Then for any* $t \in \mathcal{T}$,

$$\sqrt{nh^{d_t+2}}(\hat{\theta}_t - \theta_t) = \sqrt{\frac{h^{d_t+2}}{n}} \sum_{i=1}^n \frac{\partial}{\partial t_1} K_h(T_i - t) \frac{Y_i - \gamma(t, X_i)}{f_{T|X}(t|X_i)} + o_p(1)$$

and $\sqrt{nh^{d_t+2}}(\hat{\theta}_t - \theta_t - h^2\mathsf{B}_t^\theta) \xrightarrow{d} \mathcal{N}(0, \mathsf{V}_t^\theta)$, *where* $\mathsf{B}_t^\theta \equiv \sum_{j=1}^{d_t} \mathbb{E}\Big[\Big(\frac{1}{2}\partial_{t_j}^2\partial_{t_1}\gamma(t,X) f_{T|X}(t|X) + \partial_{t_j}\partial_{t_1}\gamma(t,X)\partial_{t_j} f_{T|X}(t|X) + \partial_{t_j}\gamma(t,X)\big(\partial_{t_j}\partial_{t_1} f_{T|X}(t|X) - \partial_{t_j} f_{T|X}(t|X)\partial_{t_1} f_{T|X}(t|X) f_{T|X}(t|X)^{-1}\big)\Big) f_{T|X}(t|X)^{-1}\Big] \int u^2 k(u)du$ *and* $\mathsf{V}_t^\theta \equiv \mathbb{E}\left[var(Y|T=t,X)/f_{T|X}(t|X)\right] \int k'(u)^2 du$.

(ii) Let $\eta/h \to \rho \in (0,\infty]$, $nh^{d_t}\eta^2 \to \infty$, *and* $nh^{d_t}\eta^4 \to 0$. *Then* $\sqrt{nh^{d_t}\eta^2}(\hat{\theta}_t - \theta_t - h^2\mathsf{B}_t^\theta) \xrightarrow{d} \mathcal{N}(0, \mathsf{V}_t^\theta)$, *where* $\mathsf{V}_t^\theta \equiv 2\mathbb{E}\big[var[Y|T=t,X]/f_{T|X}(t|X)\big]\big(\int_{-\infty}^{\infty} k(u)^2 du - \bar{k}(\rho)\big)$ *with the convolution kernel* $\bar{k}(\rho) = \int_{-\infty}^{\infty} k(u)k(u-\rho)du$ *and* $\mathsf{B}_t^\theta \equiv \partial \mathsf{B}_t/\partial t_1$ *given in Theorem 1.*

Theorem 2(i) is for the case when η is chosen to be of smaller order than h. The conditions (a) and (b) imply that η cannot be too small and depends on the precision of the nuisance estimators. In Theorem 2(ii) when $\eta/h \to \infty$, $\bar{k}(\eta/h) = 0$ and hence $\mathsf{V}_t^\theta = 2\mathsf{V}_t$. This is consistent with the special case of a fixed η implied by the result in Theorem 1.

3.1 Gateaux derivative limit

One way to obtain the influence function is to calculate the limit of the Gateaux derivative with respect to a smooth deviation, as the deviation approaches a point mass, following Ichimura and Newey (2017) for semiparametric estimators. The partial mean β_t is a marginal integration over (Y, X), fixing the value of T at t. As a result, the Gateaux derivative depends on the choice of the distribution f_T^h that belongs to a family of distributions approaching a point mass at T as $h \to 0$. We construct the locally robust estimator based on the influence function derived by the Gateaux derivative, so the asymptotic distribution of $\hat{\beta}_t$ depends on the choice of f_T^h that is the kernel function $K_h(T - t)$.

More specifically, for any $t \in \mathcal{T}$, let $\beta_t(\cdot) : \mathcal{F} \to \mathcal{R}$, where $\mathcal{F}$ is a set of CDFs of $Z \equiv (Y, T', X')'$ that is unrestricted except for regularity conditions. The estimator converges to $\beta_t(F)$ for some $F \in \mathcal{F}$, which describes how the limit of the estimator varies as the distribution of a data observation varies. Let F^0 be the true distribution of Z. Let F_Z^h approach a point mass at Z as $h \to 0$. Consider $F^{\tau h} = (1 - \tau)F^0 + \tau F_Z^h$ for $\tau \in [0, 1]$ such that for all small enough τ, $F^{\tau h} \in \mathcal{F}$ and the corresponding pdf $f^{\tau h} = f^0 + \tau(f_Z^h - f^0)$. We calculate the Gateaux derivative of the functional $\beta_t(F^{\tau h})$ with respect to a deviation $F_Z^h - F^0$ from the true distribution F^0.

In the Appendix, we show that the Gateaux derivative for the direction $f_Z^h - f^0$ is

$$\begin{aligned} \lim_{h\to 0} \frac{d}{d\tau} \beta_t(F^{\tau h})\Big|_{\tau=0} &= \gamma(t, X) - \beta_t + \lim_{h\to 0} \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{y - \gamma(t, x)}{f_{T|X}(t|x)} f_{YTX}^h(y, t, x) dy dx \\ &= \gamma(t, X) - \beta_t + \frac{Y - \gamma(t, X)}{f_{T|X}(t|X)} \lim_{h\to 0} f_T^h(t). \end{aligned} \tag{4}$$

Note that the last term in (4) is a partial mean that is a marginal integration over $\mathcal{Y} \times \mathcal{X}$, fixing the value of T at t (Newey, 1994). Thus the Gateaux derivative depends on the choice of f_T^h. In particular we choose $f_Z^h(z) = K_h(Z - z)\mathbf{1}\{f^0(z) > h\}$, following Ichimura and Newey (2017). Then

$$\frac{Y - \gamma(t, X)}{f_{T|X}(t|X)} \lim_{h\to 0} f_T^h(t) = \frac{Y - \gamma(t, X)}{f_{T|X}(t|X)} \lim_{h\to 0} K_h(T - t).$$

Theorem 1 in Ichimura and Newey (2017) shows that if a semiparametric estimator is asymp-

totic linear and locally regular, then the influence function is $\lim_{h\to 0} d\beta_t(F^{\tau h})/d\tau|_{\tau=0}$. Here, we use the Gateaux derivative limit calculation to motivate our estimator that depends on F_T^h. Then we show that the estimator is asymptotically linear with such influence function. This is the estimator-based approach in Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018) (CEINR, hereafter) in the sense that the influence function is determined by the limit $\beta_t(F^{\tau h})$ of an estimator $\hat{\beta}_t$ or the adjustment from the first step estimator $\hat{\gamma}$ discussed in the next section.

3.2 Adjustment for first-step kernel estimation

We discuss another way to motivate our moment function. We consider two alternative estimators for the dose response function, or the average structural function, β_t: the regression estimator

$$\hat{\beta}_t^{REG} = \frac{1}{n}\sum_{i=1}^{n} \hat{\gamma}(t, X_i)$$

and the inverse probability weighting (IPW) estimator

$$\hat{\beta}_t^{IPW} = \frac{1}{n}\sum_{i=1}^{n} \frac{K_h(T_i - t)Y_i}{\hat{f}_{T|X}(t|X_i)}.$$

Adding the influence function that accounts for the first-step estimation partials out the first-order effect of the first-step estimation on the final estimator, as in Section 2.2.5 in CCDDHNR.

For $\hat{\beta}_t^{REG}$, when $\hat{\gamma}(t, x)$ is a local constant or local polynomial estimator with bandwidth h, Newey (1994) and Lee (2018) have derived the asymptotic linear representation of $\hat{\beta}_t^{REG}$ that is first-order equivalent to that of our DML estimator given in Theorem 1. Specifically we can obtain the adjustment term by the influence function of the partial mean $\int_{\mathcal{X}} \hat{\gamma}(t, x)f(x)dx = n^{-1}\sum_{i=1}^{n} K_h(T_i - t)(Y_i - \gamma(t, X_i))/f_{T|X}(t|X_i) + o_p((nh^{d_t})^{-1/2})$ with a suitably chosen h and regularity conditions. Thus the moment function can be constructed by adding the influence function adjustment for estimating the nuisance function $\gamma(t, X)$ to the original score function $\gamma(t, X)$.

Similarly for $\hat{\beta}_t^{IPW}$, when $\hat{f}_{T|X}$ is a standard kernel density estimator with bandwidth h, Hsu, Huber, Lee, and Pipoz (2018) derive the asymptotic linear representation of $\hat{\beta}_t^{IPW}$ that is first-order equivalent to our DML estimator. We can show that the partial mean $\int_{\mathcal{Z}} K_h(T - t)Y/\hat{f}_{T|X}(t|X)dF_{YTX} = n^{-1}\sum_{i=1}^{n} \gamma(t, X_i)\left(1 - K_h(T_i - t)/f_{T|X}(t|X_i)\right) + o_p((nh^{d_t})^{-1/2})$ with a suitably chosen h and regularity conditions. Thus the moment function can be constructed by adding the influence function adjustment for estimating the nuisance function $f_{T|X}$ to the original score function $K_h(T - t)Y/f_{T|X}(t|X)$.

Remark 1 (First-step bias reduction) In general, nonparametric estimation of an infinite-dimensional nuisance parameter contributes a finite-sample bias to the final estimator. It is noteworthy that although the kernel function in the DML estimator $\hat{\beta}_t$ introduces the first-order bias $h^2\mathsf{B}_t$, $\hat{\beta}_t$ requires a weaker bandwidth condition for controlling the bias of the first-step estimator than the regression estimator $\hat{\beta}_t^{REG}$ and the IPW estimator $\hat{\beta}_t^{IPW}$. Our DML estimator for continuous treatments inherits this advantageous property from the DML estimator for a binary treatment. Therefore the DML estimator can be less sensitive to variation in tuning parameters of the first-step estimators. To illustrate with a simple example of $\hat{\beta}_t^{REG}$, consider the first-step $\hat{\gamma}$ to be a local constant estimator with bandwidth h_1 and a kernel of order r. To control the bias of $\hat{\gamma}$ to be asymptotically negligible for $\hat{\beta}_t^{REG}$, we assume $h_1^r\sqrt{nh_1^{d_t}} \to 0$. In contrast, when $\hat{\gamma}$ and $\hat{f}_{T|X}$ in the DML estimator $\hat{\beta}_t$ are local constant estimators with bandwidth h_1 and a kernel of order r, Assumption 3(ii) requires $h_1^{2r}\sqrt{nh^{d_t}} \to 0$. Moreover we observe that the condition is weaker than $h_1^r\sqrt{n} \to 0$ for the binary treatment that has a regular root-n convergence rate.

Remark 2 (First-step series estimation) When $\hat{\gamma}(t,x)$ is a series estimator in $\hat{\beta}_t^{REG}$, computing the partial mean $\int_{\mathcal{X}} \hat{\gamma}(t,x)f(x)dx$ for the influence function results in a different adjustment term than the kernel estimation discussed above.[6] Heuristically, let us consider a basis function $R(T,X)$ that contains (T,X) as well as interactions and other transformations of these regressors, such as power series or splines. Computing $\int \hat{\gamma}(t,x)f(x)dx$ implies the adjustment term of the form $\mathbb{E}[R(t,X)]\left(n^{-1}\sum_{i=1}^n R(T_i,X_i)R(T_i,X_i)'\right)^{-1} n^{-1}\sum_{i=1}^n R(T_i,X_i)'\big(y_i-\gamma(T_i,X_i)\big) = n^{-1}\sum_{i=1}^n \lambda_{ti}\big(y_i-\gamma(T_i,X_i)\big)$, resulting in a form of an average weighted residuals in estimation or projection of the residual on the space generated by the basis functions. Notice that the generalized propensity score $f_{T|X}(t|X)$ is not explicit in the weight λ_{ti}. Such adjustment term may motivate different estimators of β_t, such as the minimum distance estimator with first-step series estimation. That is beyond the scope of this paper and left for future research. For similar debiased estimation for binary treatment effects, see Athey, Imbens, and Wager (2018) for the approximate residual balancing estimator and Chernozhukov, Newey, Robins, and Singh (2019) for Riesz representers. See also CEINR and Demirer, Syrgkanis, Lewis, and Chernozhukov (2019).

3.3 Asymptotic linear representation

We give an outline of deriving the asymptotic linear representation in Theorem 1, following CEINR. Let $\gamma(t,x) \equiv \mathbb{E}[Y|T=t,X=x]$ and $\lambda(t,x) \equiv 1/f_{T|X}(t|x)$. The moment function for identification is $m(Z_i,\beta_t,\gamma) = \gamma(t,X_i) - \beta_t$, i.e., $\mathbb{E}[m(Z_i,\beta_t,\gamma(t,X_i))] = 0$ uniquely defines β_t. The adjustment

[6]For example, Lee and Li (2018) derive the asymptotic theory of a partial mean of a series estimator, in estimating the average structural function with a special regressor.

term is $\phi(Z_i, \beta_t, \gamma, \lambda) = K_h(T_i - t)\lambda(t, X_i)\,(Y_i - \gamma(t, X_i))$. The doubly robust moment function is $\psi(Z_i, \beta_t, \gamma, \lambda) = m(Z_i, \beta_t, \gamma(t, X_i)) + \phi(Z_i, \beta_t, \gamma(t, X_i), \lambda(t, X_i))$.

Let $\gamma_i = \gamma(t, X_i)$ and $\lambda_i = \lambda(t, X_i)$ for notational ease. We decompose the remainder term

$$
\begin{aligned}
&\sqrt{nh^{d_t}}\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\psi}(Z_i, \beta_t, \hat{\gamma}_i, \hat{\lambda}_i) - \psi(Z_i, \beta_t, \gamma_i, \lambda_i)\Big) \\
&= \sqrt{\frac{h^{d_t}}{n}}\sum_{i=1}^{n}\Big\{\hat{\gamma}_i - \gamma_i - \mathbb{E}[\hat{\gamma}_i - \gamma_i] + K_h(T_i - t)\lambda_i(\gamma_i - \hat{\gamma}_i) - \mathbb{E}\big[K_h(T_i - t)\lambda_i(\gamma_i - \hat{\gamma}_i)\big]\Big\} \qquad \text{(R1-1)} \\
&\quad + \sqrt{\frac{h^{d_t}}{n}}\sum_{i=1}^{n}\Big\{K_h(T_i - t)(\hat{\lambda}_i - \lambda_i)(Y_i - \gamma_i) - \mathbb{E}\big[K_h(T_i - t)(\hat{\lambda}_i - \lambda_i)(Y_i - \gamma_i)\big]\Big\} \qquad \text{(R1-2)} \\
&\quad + \sqrt{\frac{h^{d_t}}{n}}\sum_{i=1}^{n}\Big\{\mathbb{E}[(\hat{\gamma}_i - \gamma_i)(1 - K_h(T_i - t)\lambda_i)] + \mathbb{E}[(\hat{\lambda}_i - \lambda_i)K_h(T_i - t)(Y_i - \gamma_i)]\Big\} \qquad \text{(R1-DR)} \\
&\quad - \sqrt{\frac{h^{d_t}}{n}}\sum_{i=1}^{n}K_h(T_i - t)\big(\hat{\lambda}_i - \lambda_i\big)\big(\hat{\gamma}_i - \gamma_i\big). \qquad \text{(R2)}
\end{aligned}
$$

The remainder terms (R1-1) and (R1-2) are stochastic equicontinuous terms that are controlled to be $o_p(1)$ by the mean square consistency conditions Assumption 3(i) and cross-fitting.

The remainder term (R1-DR) is controlled by the doubly robust property. Note that in the binary treatment case when $K_h(T_i - t)$ is replaced by $\mathbf{1}\{T_i = t\}$, the term (R1-DR) is zero because ψ is the Neyman-orthogonal score. In our continuous treatment case, the Neyman orthogonality holds as $h \to 0$. Under the conditions in Theorem 1, (R1-DR) is $O_p\big((\|\hat{\gamma} - \gamma\|_{L_2} + \|\hat{\lambda} - \lambda\|_{L_2})\sqrt{nh^{4+d_t}}\big) = o_p(1)$.

The second-order remainder term (R2) is controlled by Assumption 3(ii).[7] To control these remainder terms to be of smaller order, we follow the proofs of Theorem 13 in CEINR.

4 Conclusion and outlook

This paper provides a nonparametric inference method for continuous treatments effects under unconfoundedness and in the presence of high-dimensional or nonparametric nuisance parameters. The proposed double debiased machine learning estimator uses a doubly robust moment function and cross-fitting. We provide tractable primitive conditions for the nuisance estimators and asymptotic theory for inference on the average dose-response function (or the average structural function) and the partial effect. Numerical examples of Monte Carlo simulations and empirical

[7] We can follow Newey and Robins (2018) and Rothe and Firpo (2018) to further investigate weaker conditions for variance terms, but still require that the product of biases of $\hat{\gamma}$ and $\hat{\lambda}$ converge to zero faster than $1/\sqrt{nh^{d_t}}$.

illustration using various ML methods are in progress. For a future extension, our DML estimator serves as the preliminary element for policy learning and optimization with a continuous decision, following Kitagawa and Tetenov (2018), Kallus and Zhou (2018), Demirer, Syrgkanis, Lewis, and Chernozhukov (2019), Athey and Wager (2019), for example.

When unconfoundedness is violated, we can use the control function approach in triangular simultaneous equations models by including in the covariates some estimated control variables using instrumental variables. For example, Imbens and Newey (2009) show that the conditional independence assumption holds when the covariates X include the additional control variable $V = F_{T|Z}(T|Z)$, the conditional distribution function of the endogenous variable given the instrumental variables Z. The influence function that accounts for estimating the control variables as generated regressors has derived in Corollary 2 in Lee (2015). Lee (2015) shows that the adjustment terms for the estimated control variables are of smaller order in the influence function of the final estimator, but it may be important to include them to achieve local robustness. This is a distinct feature of the dose-response function/the average structural function of continuous treatments, as discussed in Section 3. Using such influence function to construct the corresponding double debiased ML estimator is left for future research.

Appendix

Proof of Lemma 2.1 Denote $\tilde{f}_{T|X}(t|X)$ to be the infeasible estimator using the true $F_{T|X}$; for example of $d_T = 1$, $\tilde{f}_{T|X}(t|x) = (2\epsilon)^{-1}\left(F_{T|X}(t+\epsilon|x) - F_{T|X}(t-\epsilon|x)\right)$. By the triangular inequality, it suffices to show that

$$\begin{aligned}\left\|\hat{f}_{T|X}(t|X) - f_{T|X}(t|X)\right\|_\infty \leq & \left\|\hat{\mathbb{E}}\left[G((T-t)/h_1)|X=x\right] - \mathbb{E}[G((T-t)/h_1)|X]\right\|_\infty \epsilon^{-d_T} \quad (5)\\ & + \left\|\mathbb{E}[G((T-t)/h_1)|X] - F_{T|X}(t|X)\right\|_\infty \epsilon^{-d_T} \quad (6)\\ & + \|(\tilde{f}_{T|X}(t|X) - f_{T|X}(t|X)\|_\infty \quad (7)\\ = & \, O_p(R_1\epsilon^{-d_T} + h_1^2\epsilon^{-d_T} + \epsilon^2).\end{aligned}$$

For (5), we give a crude bound by exploiting the convergence rate of the ML or nonparametric estimators. For (6), we follow the standard algebra for kernel, using integration by parts and change of variables. We analyze (7) below.

We first prove the results for $d_T = 1$. By a Taylor expansion, $F_{T|X}(t \pm \epsilon|x) = F_{T|X}(t|x) \pm \epsilon f_{T|X}(t|x) + \frac{\epsilon^2}{2}\frac{d}{dt}f_{T|X}(t|x) \pm \frac{\epsilon^3}{3!}\frac{d^2}{dt^2}f_{T|X}(t_\pm|x)$ for some $t_+ \in (t, t+\epsilon)$ and $t_- \in (t-\epsilon, t)$. Thus, $\|(2\epsilon)^{-1}(F_{T|X}(t+\epsilon|X) - F_{T|X}(t-\epsilon|X)) - f_{T|X}(t|X)\|_\infty = O(\epsilon^2)$.

Next we prove the results for $d_T = 2$. The general $d_T > 2$ can be derived by induction. Consider any $x \in \mathcal{X}$ and $t = (t_1, t_2)' \in \mathcal{T}$. Let $F \equiv F_{T|X}(t_1, t_2|x)$. For any positive sequences $\epsilon = (\epsilon_1, \epsilon_2)' \to 0$, let $F_{++} \equiv F_{T|X}(t_1+\epsilon_1, t_2+\epsilon_2|x)$, $F_{+-} \equiv F_{T|X}(t_1+\epsilon_1, t_2-\epsilon_2|x)$, $F_{-+} \equiv F_{T|X}(t_1-\epsilon_1, t_2+\epsilon_2|x)$, $F_{--} \equiv F_{T|X}(t_1-\epsilon_1, t_2-\epsilon_2|x)$, and $\partial_j^\nu F = \frac{\partial^l}{\partial t_j^\nu}F_{T|X}(t|x)$ that is the ν^{th} partial derivative of F with respect to t_j.

By a Taylor expansion,

$$\begin{aligned}F_{++} = & F + \epsilon_1\partial_1 F + \epsilon_2\partial_2 F + \frac{\epsilon_1^2}{2}\partial_1^2 F + \frac{\epsilon_1^2}{2}\partial_2^2 F + \epsilon_1\epsilon_2\partial_1\partial_2 F\\ & + \frac{\epsilon_1^3}{3!}\partial_1^3 F + \frac{\epsilon_1^2\epsilon_2}{2}\partial_1^2\partial_2 F + \frac{\epsilon_1\epsilon_2^2}{2}\partial_1\partial_2^2 F + \frac{\epsilon_2^3}{3!}\partial_2^3 F\\ & + \frac{\epsilon_1^4}{4!}\partial_1^4\bar{F}_{++} + \frac{4\epsilon_1^3\epsilon_2}{4!}\partial_1^3\partial_2\bar{F}_{++} + \frac{6\epsilon_1^2\epsilon_2^2}{4!}\partial_1^2\partial_2^2\bar{F}_{++} + \frac{4\epsilon_1\epsilon_2^3}{4!}\partial_1\partial_2^3\bar{F}_{++} + \frac{\epsilon_2^4}{4!}\partial_2^4\bar{F}_{++},\end{aligned}$$

where $\bar{F}_{++} = F_{T|X}(\bar{t}|x)$ with the mean value $\bar{t} \in (t, t+\epsilon)$. Similarly,

$$\begin{aligned}F_{+-} = & F + \epsilon_1\partial_1 F - \epsilon_2\partial_2 F + \frac{\epsilon_1^2}{2}\partial_1^2 F + \frac{\epsilon_1^2}{2}\partial_2^2 F - \epsilon_1\epsilon_2\partial_1\partial_2 F\\ & + \frac{\epsilon_1^3}{3!}\partial_1^3 F - \frac{\epsilon_1^2\epsilon_2}{2}\partial_1^2\partial_2 F + \frac{\epsilon_1\epsilon_2^2}{2}\partial_1\partial_2^2 F - \frac{\epsilon_2^3}{3!}\partial_2^3 F\\ & + \frac{\epsilon_1^4}{4!}\partial_1^4\bar{F}_{+-} - \frac{4\epsilon_1^3\epsilon_2}{4!}\partial_1^3\partial_2\bar{F}_{+-} + \frac{6\epsilon_1^2\epsilon_2^2}{4!}\partial_1^2\partial_2^2\bar{F}_{+-} - \frac{4\epsilon_1\epsilon_2^3}{4!}\partial_1\partial_2^3\bar{F}_{+-} + \frac{\epsilon_2^4}{4!}\partial_2^4\bar{F}_{+-},\end{aligned}$$

where $\bar{F}_{+-} = F_{T|X}(\bar{t}|x)$ with the mean values $\bar{t}_1 \in (t_1, t_1+\epsilon_1)$ and $\bar{t}_2 \in (t_2-\epsilon_2, t_2)$. We implement

the same Taylor expansions on F_{-+} and F_{--}. Then

$$\tilde{f}_{T|X} = (F_{++} - F_{+-} - F_{-+} + F_{--})/(4\epsilon_1\epsilon_2) = \partial_1\partial_2 F + \frac{\epsilon_2^2}{3!}\partial_2^3\partial_1 F + \frac{\epsilon_1^2}{3!}\partial_2\partial_1^3 F + o((\epsilon_1+\epsilon_2)^4/(\epsilon_1\epsilon_2)),$$

assuming $(\epsilon_1+\epsilon_2)^4/(\epsilon_1\epsilon_2) = O(1)$ that holds for $\epsilon_1 = \epsilon_2 = \epsilon$.

For a general d_T, by induction, we can obtain $\tilde{f}_{T|X} = f_{T|X} + O(\epsilon^2)$, where the error $O(\epsilon^2)$ is from the $(d_T+2)^{th}$ derivatives of F. We can allow ϵ_j to be different for $j = 1, ..., d_T$ by assuming $(\sum_{j=1}^{d_T}\epsilon_j)^{d_T+2}/\Pi_{j=1}^{d_T}\epsilon_j = O(1)$. □

We present more primitive conditions on estimating the nuisance parameters in Assumption 4 that is implied by Assumption 3.

Assumption 4 *For each $l = 1, ..., L$ and for any $t \in \mathcal{T}$,*

(i) $\int_{\mathcal{X}} \left(\hat{\gamma}_l(t,x) - \gamma(t,x)\right)^2 f_X(x)dx \xrightarrow{p} 0$ *and* $\int_{\mathcal{X}} \left(\hat{f}_l(t|x) - f_{T|X}(t|x)\right)^2 f_X(x)dx \xrightarrow{p} 0$.

(ii) Either (a) $\sqrt{nh^{d_t}}n^{-1}\sum_{i=1}^n K_h(T_i - t)\big(1/\hat{f}_l(t|X_i) - 1/f_{T|X}(t|X_i)\big)\big(\hat{\gamma}_l(t,X_i) - \gamma(t,X_i)\big) \xrightarrow{p} 0$,
or (b) $\sqrt{nh^{d_t}}\int_{\mathcal{X}} \big|(\hat{f}_l(t|x) - f_{T|X}(t|x))(\hat{\gamma}_l(t,x) - \gamma(t,x))\big| f_{TX}(t,x)dx \xrightarrow{p} 0$, *or*
(c) $\sqrt{nh^{d_t}}\Big(\int_{\mathcal{X}} \left(\hat{f}_l(t|x) - f_{T|X}(t|x)\right)^2 f_{TX}(t,x)dx\Big)^{1/2}\Big(\int_{\mathcal{X}} \left(\hat{\gamma}_l(t|x) - \gamma(t,x)\right)^2 f_{TX}(t,x)dx\Big)^{1/2} \xrightarrow{p} 0$.

Under Assumption 1(ii), Assumption 4 is implied by Assumption 3.[8] Moreover, a weaker condition on the first step estimators is possible by the choice of h. In the proof of Theorem 1, we note that in Assumption 4(ii), the condition (c) implies (b), which then implies (a).

Proof of Theorem 1 The proof modifies Assumptions 4 and 5 and extends Lemma A1, Lemma 12, and Theorem 13 in CEINR. Let Z_l^c denote the observations z_i for $i \neq I_l$. Let $\hat{\gamma}_{il} = \hat{r}_l(t, X_i)$ using Z_l^c for $i \in I_l$. Following the proof of Lemma A1 in CEINR, define $\hat{\Delta}_{il} = \hat{\gamma}_{il} - \gamma_i - \mathbb{E}\left[\hat{\gamma}_{il} - \gamma_i\right]$ for $i \in I_l$. By construction and independence of Z_l^c and $z_i, i \in I_l$, $\mathbb{E}[\hat{\Delta}_{il}|Z_l^c] = 0$ and $\mathbb{E}[\hat{\Delta}_{il}\hat{\Delta}_{jl}|Z_l^c] = 0$ for $i, j \in I_l$. For $i \in I_l$ and for all t, $h\mathbb{E}[\hat{\Delta}_{il}^2|Z_l^c] = h\int(\hat{\gamma}_{il} - \gamma_i)^2 f_X(X_i)dX_i \xrightarrow{p} 0$ by Assumption 4(i). Then $\mathbb{E}\left[\left(\sqrt{h^{d_t}/n}\sum_{i\in I_l}\hat{\Delta}_{il}\right)^2 \middle| Z_l^c\right] = (h/n)\sum_{i\in I_l}\mathbb{E}\left[\hat{\Delta}_{il}^2\middle|Z_l^c\right] \leq h\int(\hat{\gamma}_{il} - \gamma_i)^2 f_X(X_i)dX_i \xrightarrow{p} 0$. The conditional Markov inequality implies that $\sqrt{h^{d_t}/n}\sum_{i\in I_l}\hat{\Delta}_{il} \xrightarrow{p} 0$.

The analogous results also hold for $\hat{\Delta}_{il} = K_h(T_i - t)\lambda_i(\gamma_i - \hat{\gamma}_{il}) - \mathbb{E}\left[K_h(T_i - t)\lambda_i(\gamma_i - \hat{\gamma}_{il})\right]$ in (R1-1) and $\hat{\Delta}_{il} = K_h(T_i - t)(\hat{\lambda}_{il} - \lambda_i)(Y_i - \gamma_i) - \mathbb{E}\left[K_h(T_i - t)(\hat{\lambda}_{il} - \lambda_i)(Y_i - \gamma_i)\right]$ in (R1-2). In particular, for (R1-2), $h\mathbb{E}[\hat{\Delta}_{il}^2|Z_l^c] = O_p\left(\int k(u)^2du\int_{\mathcal{X}}\left(\hat{\lambda}_{il} - \lambda_i\right)^2 f_X(X_i)dX_i\right) \xrightarrow{p} 0$ by the smoothness condition and Assumption 4(i). So (R1-1)$\xrightarrow{p} 0$ and (R1-2)$\xrightarrow{p} 0$.

[8] We claim that Assumption 3(i) is implied by Assumption 4(i). Other conditions can be shown by analogous arguments. Denote $\hat{A}(t) \equiv \int\left(\hat{\gamma}_l(t,x) - \gamma(t,x)\right)^2 f_{TX}(t,x)dx \geq 0$. The following shows $\int_{\mathcal{T}}\hat{A}(t)dt = o_p(1)$ implies $\hat{A}(t) = o_p(1)$ for ant $t \in \mathcal{T}$. For any positive C and ϵ, there exists a positive integer N such that $Pr(\int_{\mathcal{T}}\hat{A}(t)dt \geq C) \leq \epsilon$ for $n \geq N$. Under Assumption 1(ii), $\hat{A}(t) \geq C$ for all $t \in \mathcal{T}$ implies $\int_{\mathcal{T}}\hat{A}(t)dt \geq C$. So $Pr(\hat{A}(t) \geq C, \forall t \in \mathcal{T}) \leq Pr(\int_{\mathcal{T}}\hat{A}(t)dt \geq C) \leq \epsilon$ for $n \geq N$.

For (R2),

$$
\begin{aligned}
&\mathbb{E}\bigg[\sqrt{h^{d_t}/n}\sum_{i\in I_l} K_h(T_i-t)(\hat{\lambda}_{il}-\lambda_i)(\gamma_i-\hat{\gamma}_{il})\bigg|Z_l^c\bigg] \\
&\leq \sqrt{nh^{d_t}}\int_{\mathcal{X}}\int_{\mathcal{T}}\Big|K_h(T_i-t)(\hat{\lambda}_{il}-\lambda_i)(\gamma_i-\hat{\gamma}_{il})\Big|f_{TX}(T_i,X_i)dT_idX_i \\
&\leq \sqrt{nh^{d_t}}\int_{\mathcal{X}} f_{T|X}(t|X_i)\Big|(\hat{\lambda}_{il}-\lambda_i)(\gamma_i-\hat{\gamma}_{il})\Big|f_X(X_i)dX_i + o_p(\sqrt{nh^{d_t}}h^2) \\
&\leq \sqrt{nh^{d_t}}\Big(\int_{\mathcal{X}} f_{T|X}(t|X_i)(\hat{\lambda}_{li}-\lambda_i)^2 f_X(X_i)dX_i\Big)^{1/2}\Big(\int_{\mathcal{X}} f_{T|X}(t|X_i)(\hat{\gamma}_{li}-\gamma_i)^2 f_X(X_i)dX_i\Big)^{1/2} + o_p(1) \\
&\xrightarrow{p} 0
\end{aligned}
$$

by Cauchy-Schwartz inequality, Assumption 4(ii)(c), and $nh^{d_t+4}\to C$. So (R2)$\xrightarrow{p}0$ follows by the conditional Markov and triangle inequalities.

For (R1-DR), in the first part $\mathbb{E}\big[1-K_h(T_i-t)\lambda_i\big|X_i\big] = \mathbb{E}\big[f_{T|X}(t|X_i)-K_h(T_i-t)\big|X_i\big]\lambda_i = h^2 f''_{T|X}(t|X_i)\lambda_i\int u^2K(u)du/2 + O_p(h^3)$. A similar argument yields (R1-DR)$= O_p((\|\hat{\gamma}-\gamma\|_{L_2} + \|\hat{\lambda}-\lambda\|_{L_2})\sqrt{nh^{d_t}}h^2) = o_p(1)$.

By the triangle inequality, we obtain the asymptotic linear representation $\sqrt{nh^{d_t}}n^{-1}\sum_{i=1}^n\big(\hat{\psi}(Z_i,\beta_t,\hat{\gamma}_t,\hat{\lambda}_t)-\psi(Z_i,\beta_t,\gamma_t,\lambda_t)\big) = o_p(1)$.

For B_t, $\mathbb{E}\left[\frac{K_h(T-t)}{f_{T|X}(t|X)}(Y-\gamma(t,X))\right] = \mathbb{E}\left[\frac{1}{f_{T|X}(t|X)}\mathbb{E}\left[K_h(T-t)\left(\gamma(T,X)-\gamma(t,X)\right)|X\right]\right]$. Let $\partial_t \equiv \partial/\partial t$ and $\partial_t^2 \equiv \partial^2/\partial t^2$. A standard algebra for kernel yields

$$
\begin{aligned}
&\mathbb{E}\left[K_h(T-t)\left(\gamma(T,X)-\gamma(t,X)\right)|X\right] \\
&= \int_{\mathcal{T}} K_h(T-t)\left(\gamma(T,X)-\gamma(t,X)\right)f_{T|X}(T|X)dT \\
&= \int k(u)\left(\gamma(t+uh,X)-\gamma(t,X)\right)f_{T|X}(t+uh|X)du \\
&= \int k(u_1)\cdots k(u_{d_t})\left(\sum_{j=1}^{d_t} u_j h\partial_{t_j}\gamma(t,X) + \frac{u_j^2h^2}{2}\partial_{t_j}^2\gamma(t,X)\right) \\
&\quad\times\left(f_{T|X}(t|X) + \sum_{j=1}^{d_t} u_j h\partial_{t_j}f_{T|X}(t|X) + \frac{u_j^2h^2}{2}\partial_{t_j}^2 f_{T|X}(t|X)\right)du_1\cdots du_{d_t} + O(h^3) \\
&= h^2\int u^2k(u)du\sum_{j=1}^{d_t}\left(\partial_{t_j}\gamma(t,X)\partial_{t_j}f_{T|X}(t|X) + \frac{1}{2}\partial_{t_j}^2\gamma(t,X)f_{T|X}(t|X)\right) + O(h^3)
\end{aligned}
$$

for all $X \in \mathcal{X}$. Thus

$$\mathbb{E}\left[\frac{1}{f_{T|X}(t|X)}\mathbb{E}\left[K_h(T-t)\left(\gamma(T,X)-\gamma(t,X)\right)|X\right]\right]$$
$$= h^2 \int u^2 k(u)du \sum_{j=1}^{d_t} \mathbb{E}\left[\partial_{t_j}\gamma(t,X)\frac{\partial_{t_j} f_{T|X}(t|X)}{f_{T|X}(t|X)} + \frac{1}{2}\partial_{t_j}^2\gamma(t,X)\right] + O(h^3).$$

The asymptotic variance is determined by $h\mathbb{E}\left[\left((Y-\gamma(t,X))K_h(T_i-t)/f_{T|X}(t|X)\right)^2\right]$. A standard algebra for kernel as above yields V_t. Asymptotic normality follows directly from the central limit theorem. □

Proof of Corollary 1 By Theorem 1, the asymptotic mean squared error is $h^4\mathsf{B}^2 + \mathsf{V}_t/(nh^{d_t})$. The result follows. □

Proof of Theorem 2 We decompose $\hat{\theta}_t - \theta_t = (\hat{\theta}_t - \theta_{t\eta}) + (\theta_{t\eta} - \theta_t)$, where $\theta_{t\eta} \equiv (\beta_{t^+} - \beta_{t^-})/\eta$. By a Taylor expansion, the second part $\theta_{t\eta} - \theta_t = O(\eta)$ if $\partial^2\beta_t/\partial t_1^2$ exists.

Let $\hat{\beta}_t = n^{-1}\sum_{i=1}^n \hat{\psi}_{ti} = n^{-1}\sum_{i=1}^n \left(\psi_{ti}+R_{ti}\right)$, where $\psi_{ti} = \psi(Z_i, \beta_t, \gamma_i, \lambda_i)$, $\hat{\psi}_{ti} = \psi(Z_i, \beta_t, \hat{\gamma}_i, \hat{\lambda}_i)$, and the remainder terms R_{ti} are defined in Section 3.3. Thus $\hat{\theta}_t - \theta_{t\eta} = \eta^{-1}n^{-1}\sum_{i=1}^n \left(\psi_{t^+i} - \psi_{t^-i} + R_{t^+i} - R_{t^-i}\right)$.

(i) By $\eta/h \to 0$ and a Taylor expansion, the variance of $\eta^{-1}n^{-1}\sum_{i=1}^n \left(\psi_{t^+i} - \psi_{t^-i}\right)$ is dominated by the variance of $n^{-1}\sum_{i=1}^n \partial_{t_1}\psi_{ti}$, where

$$\partial_{t_1}\psi_{ti} = \partial_{t_1}K_h(T_i-t)\frac{Y_i-\gamma(t,X_i)}{f_{t|X_i}} + K_h(T_i-t)\partial_{t_1}\left(\frac{Y_i-\gamma(t,X_i)}{f_{t|X_i}}\right) + \partial_{t_1}\gamma(t,X_i) - \theta_t.$$

Thus the leading term of the variance of $\eta^{-1}n^{-1}\sum_{i=1}^n \left(\psi_{t^+i} - \psi_{t^-i}\right)$ is $\int \left(\partial_{t_1}K_h(T-t)\right)^2\mathbb{E}\big[(Y - \gamma(t,X))^2|T,X\big]\frac{f_{T|X}}{f_{t|X}^2}dT = h^{-(d_t+2)}\mathbb{E}\big[var(Y|T=t,X)/f_{T|X}(t|X)\big]\int k'^2(u)du + o(h^{-(d_t+2)}) = O\big(h^{-(d_t+2)}\big)$.

To control $\sqrt{nh^{d_t+2}}\eta^{-1}n^{-1}\sum_{i=1}^n \left(R_{t^+i} - R_{t^-i}\right) = o_p(1)$, the conditions (a) and (b) give a coarse bound $\sqrt{h^{d_t}/n}\sum_{i=1}^n R_{ti}h\eta^{-1} = o_p(1)$ following the proof of Theorem 1.

For the bias B_t^θ,

$$\begin{aligned}
&\int \left\{ \partial_{t_1} K_h(T_i - t) \frac{\gamma(T_i, X_i) - \gamma(t, X_i)}{f_{t|X_i}} + K_h(T_i - t) \partial_{t_1} \left(\frac{\gamma(T_i, X_i) - \gamma(t, X_i)}{f_{t|X_i}} \right) \right\} f_{T_i|X_i} dT_i \\
= &\int K_h(T_i - t) \Bigg\{ \frac{\partial_{t_1} \gamma(T_i, X_i) f_{T_i|X_i}}{f_{t|X_i}} + (\gamma(T_i, X_i) - \gamma(t, X_i)) \frac{\partial_{t_1} f_{T_i|X_i}}{f_{t|X_i}} \\
&- \frac{\partial_{t_1} \gamma(t, X_i) f_{T_i|X_i}}{f_{t|X_i}} - (\gamma(T_i, X_i) - \gamma(t, X_i)) \frac{\partial_{t_1} f_{t|X_i}}{f_{t|X_i}^2} f_{T_i|X_i} \Bigg\} dT_i \\
= &\int \Bigg\{ \left(f_{t|X_i} + \sum_{j=1}^{d_t} \partial_{t_j} f_{t|X_i} u_j h + \partial_{t_j}^2 f_{t|X_i} \frac{u_j^2 h^2}{2} \right) \left(\sum_{j=1}^{d_t} \partial_{t_j} \partial_{t_1} \gamma(t, X_i) u_j h + \partial_{t_j}^2 \partial_{t_1} \gamma(t, X_i) \frac{u_j^2 h^2}{2} \right) \\
&+ \Big(\sum_{j=1}^{d_t} \partial_{t_j} \gamma(t, X_i) u_j h + \partial_{t_j}^2 \gamma(t, X_i) \frac{u_j^2 h^2}{2} \Big) \Bigg(\partial_{t_1} f_{t|X_i} + \sum_{j=1}^{d_t} \partial_{t_j} \partial_{t_1} f_{t|X_i} u_j h + \partial_{t_j}^2 \partial_{t_1} f_{t|X_i} \frac{u_j^2 h^2}{2} \\
&- \Big(f_{t|X_i} + \sum_{j=1}^{d_t} \partial_{t_j} f_{t|X_i} u_j h + \partial_{t_j}^2 f_{t|X_i} \frac{u_j^2 h^2}{2} \Big) \frac{\partial_{t_1} f_{t|X_i}}{f_{t|X_i}} \Bigg) \Bigg\} \frac{1}{f_{t|X_i}} k(u_1) \cdots k(u_{d_t}) du_1 \cdots du_{d_t} + O(h^3) \\
= &\, h^2 \sum_{j=1}^{d_t} \left(\frac{1}{2} \partial_{t_j}^2 \partial_{t_1} \gamma(t, X_i) + \partial_{t_j} \partial_{t_1} \gamma(t, X_i) \frac{\partial_{t_j} f_{t|X_i}}{f_{t|X_i}} + \frac{\partial_{t_j} \gamma(t, X_i)}{f_{t|X_i}} \left(\partial_{t_j} \partial_{t_1} f_{t|X_i} - \partial_{t_j} f_{t|X_i} \frac{\partial_{t_1} f_{t|X_i}}{f_{t|X_i}} \right) \right) \\
&\int u^2 k(u) du + O(h^3),
\end{aligned}$$

where the first equality is by integration by parts. □

(ii) $\sqrt{nh^{d_t}\eta^2}(\hat{\theta}_t - \theta_{t\eta}) = \sqrt{nh^{d_t}}\big(\hat{\beta}_{t^+} - \hat{\beta}_{t^-} - (\beta_{t^+} - \beta_{t^-})\big) = \sqrt{nh^{d_t}} n^{-1} \sum_{i=1}^n \big(\psi_{t^+i} - \psi_{t^-i} + R_{t^+i} - R_{t^-i}\big) = \sqrt{nh^{d_t}} n^{-1} \sum_{i=1}^n \big(\psi_{t^+i} - \psi_{t^-i}\big) + o_p(1)$ by Theorem 1.

For V_t^θ, the term involved the convolution kernel comes from the covariance of ψ_{t^+i} and ψ_{t^-i} in the following. $\mathbb{E}\big[\psi_{t^+i}\psi_{t^-i}\big]$ is bounded by the order of

$$\begin{aligned}
&\mathbb{E}\left[\int \int K_h(T - t^+) K_h(T - t^-) (Y - \gamma(t^+, X))(Y - \gamma(t^-, X)) \frac{f_{Y|TX}(Y|T, X) f_{T|X}(T|X)}{f_{t^+|X} f_{t^-|X}} dY dT \right] \\
&= \frac{1}{h} \mathbb{E}\Bigg[\int \big(\mathbb{E}[Y^2 | T = t^+ + uh, X] - \gamma(t^+ + uh, X)(\gamma(t^+, X) + \gamma(t^-, X)) + \gamma(t^+, X)\gamma(t^-, X) \big) \\
&\quad k(u) k\left(u - \frac{\eta}{h}\right) \frac{f_{T|X}(t^+ + uh|X)}{f_{t^+|X} f_{t^-|X}} du \Bigg] \\
&= \frac{1}{h} \bar{k}\left(\frac{\eta}{h}\right) \mathbb{E}\left[\frac{var(Y|T = t, X)}{f_{T|X}(t|X)} \right] + O(h).
\end{aligned}$$

□

Gateaux derivative Let the Dirac delta function $\delta_t(T) = \infty$ for $T = t$, $\delta_t(T) = 0$ for $T \neq t$, and $\int g(s)\delta_t(s)ds = 1$, for any continuous compactly supported function g.[9] For any $F \in \mathcal{F}$,

$$\begin{aligned}\beta_t(F) &= \int_{\mathcal{X}} \mathbb{E}[Y|T=t, X=x] f_X(x) dx \\ &= \int_{\mathcal{X}} \int_{\mathcal{T}} \mathbb{E}[Y|T=s, X=x] \delta_t(s) ds f_X(x) dx \\ &= \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} y \delta_t(s) \frac{f_{YTX}(y,s,x) f_X(x)}{f_{TX}(s,x)} dy ds dx.\end{aligned}$$

$$\begin{aligned}\frac{d}{d\tau}\beta_t(F^{\tau h}) = & \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} y\delta_t(s) \frac{d}{d\tau} \left(\frac{f_{YTX}(y,s,x) f_X(x)}{f_{TX}(s,x)} \right) dy ds dx \\ = & \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} \frac{y\delta_t(s)}{f_{TX}(s,x)} \big(\left(-f^0_{YTX}(y,s,x) + f^h_{YTX}(y,s,x) \right) f_X(x) \\ & + f_{YTX}(y,s,x) \left(-f^0_X(x) + f^h_X(x) \right) \big) dy ds dx \\ & - \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} y\delta_t(s) \frac{f_{YTX}(y,s,x) f_X(x)}{f_{TX}(s,x)^2} \left(-f^0_{TX}(s,x) + f^h_{TX}(s,x) \right) dy ds dx.\end{aligned}$$

The influence function can be calculated as

$$\lim_{h\to 0} \frac{d}{d\tau}\beta_t(F^{\tau h})\Big|_{\tau=0} = \gamma(t,X) - \beta_t + \lim_{h\to 0} \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{y - \gamma(t,x)}{f_{T|X}(t|x)} f^h_{YTX}(y,t,x) dy dx.$$

In particular, we specify F^h_Z following equation (3.1) in Ichimura and Newey (2017). Let $K_h(Z) = \Pi_{l=1}^{d_z} k(Z_l/h)/h$, where $Z = (Z_1, ..., Z_{d_z})'$ and k satisfies Assumption 2 and is continuously differentiable of all orders with bounded derivatives. Let $F^{\tau h} = (1-\tau)F^0 + \tau F^h_Z$ with pdf with respect to a product measure given by $f^{\tau h}(z) = (1-\tau)f^0(z) + \tau f^0(z)\delta^h_Z(z)$, where $\delta^h_Z(z) = K_h(Z - z)\mathbf{1}\{f^0(z) > h\}/f^0(z)$, a ratio of a sharply peaked pdf to the true density. Thus $f^h_{YTX}(y,t,x) = K_h(Y-y)K_h(T-t)K_h(X-x)\mathbf{1}\{f^0(z) > h\}$. It follows that

$$\lim_{h\to 0} \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{y - \gamma(t,x)}{f_{T|X}(t|x)} f^h_{YTX}(y,t,x) dy dx = \frac{Y - \gamma(t,X)}{f_{T|X}(t|X)} \lim_{h\to 0} K_h(T-t).$$

[9] Note that a nascent delta function to approximate the Dirac delta function is $K_h(T-t) \equiv k((T-t)/h)/h$ such that $\delta_t(T) = \lim_{h\to 0} K_h(T-t)$.

References

Athey, S. and G. Imbens (2019). Machine learning methods economists should know about. arxiv:1903.10075v1.

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

Athey, S. and S. Wager (2019). Efficient policy learning. arxiv:1702.02896.

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

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies 81*(2), 608–650.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blundell, R. and J. L. Powell (2003). Endogeneity in nonparametric and semiparametric regression models. In L. H. M. Dewatripont and S.J.Turnovsky (Eds.), *Advances in Economics and Econometrics, Theory and Applications, Eighth World Congress*, Volume II. Cambridge University Press, Cambridge, U.K.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association 113*(522), 767–779.

Carone, M., A. R. Luedtke, and M. J. van der Laan (2018). Toward computerized efficient estimation in infinite-dimensional models. *Journal of the American Statistical Association 0*(0), 1–17.

Cattaneo, M. D. and M. Jansson (2019). Average density estimators: Efficiency and bootstrap consistency. arxiv:1904.09372v1.

Chen, X. (2007). *Large sample sieve estimation of semi-nonparametric models*, Volume 6B. Amsterdam: Elsevier.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2018). Locally robust semiparametric estimation. arxiv:1608.00033.

Chernozhukov, V., W. Newey, J. Robins, and R. Singh (2019). Double/de-biased machine learning of global and local parameters using regularized riesz representers. arxiv:1802.08667v3.

Chernozhukov, V. and V. Semenova (2019). Simultaneous inference for best linear predictor of the conditional average treatment effect and other structural functions. Working paper, department of economics, mit.

Criminisi, A., J. Shotton, and E. Konukoglu (2012). Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. In *Foundations and Trends in Computer Graphics and Vision: Vol. 7: Nos. 2-3*, pp. 81–227. NOW Publishers.

Demirer, M., V. Syrgkanis, G. Lewis, and V. Chernozhukov (2019). Semi-parametric efficient policy learning with continuous actions. arxiv:1905.10116v1.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2019). Estimation of conditional average treatment effects with high-dimensional data. arxiv:1908.02399.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 189*(1), 1–23.

Farrell, M. H., T. Liang, and S. Misra (2018). Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. arxiv:1809.09953.

Flores, C. A. (2007). Estimation of dose-response functions and optimal doses with a continuous treatment. Working paper.

Flores, C. A., A. Flores-Lagunes, A. Gonzalez, and T. C. Neumann (2012). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *The Review of Economics and Statistics 94*(1), 153–171.

Galvao, A. F. and L. Wang (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association 110*(512), 1528–1542.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–332.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2017, 06–11 Aug). Deep IV: A flexible approach for counterfactual prediction. In D. Precup and Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning*, Volume 70 of *Proceedings of Machine Learning Research*, International Convention Centre, Sydney, Australia, pp. 1414–1423. PMLR.

Hirano, K. and G. W. Imbens (2004). The propensity score with continuous treatments. In A. Gelman and X.-L. Meng (Eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pp. 73–84. New York: Wiley.

Hsu, Y.-C., M. Huber, Y.-Y. Lee, and L. Pipoz (2018). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. SES Working Paper 495, University of Fribourg.

Ichimura, H. and W. Newey (2017). The influence function of semiparametric estimators. Working paper.

Imbens, G. W. and W. K. Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512.

Kallus, N. and A. Zhou (2018). Policy evaluation and optimization with continuous treatments. *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 84*, 1243–1251.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017). Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B 79*(4), 1229–1245.

Kitagawa, T. and A. Tetenov (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica 86*, 591–616.

Kluve, J., H. Schneider, A. Uhlendorff, and Z. Zhao (2012). Evaluating continuous training programs using the generalized propensity score. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*(2), 587–617.

Lee, Y.-Y. (2015). Partial mean processes with generated regressors: Continuous treatment effects and nonseparable models. Working paper.

Lee, Y.-Y. (2018). Partial mean processes with generated regressors: Continuous treatment effects and nonseparable models. arxiv:1811.00157.

Lee, Y.-Y. and H.-H. Li (2018). Partial effects in binary response models using a special regressor. *Economics Letters 169*, 15–19.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory 10*(2), 233–253.

Newey, W. K. and J. R. Robins (2018). Cross-fitting and fast remainder rates for semiparametric estimation. arxiv:1801.09138.

Oprescu, M., V. Syrgkanis, and Z. S. Wu (2019). Orthogonal random forest for causal inference. arxiv:1806.03467v3.

Pospisil, T. and A. B. Lee (2018). Rfcde: Random forests for conditional density estimation. arxiv:1804.05753.

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Rothe, C. and S. Firpo (2018). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. *Econometric Theory*, forthcoming.

Su, L., T. Ura, and Y. Zhang (2019). Non-separable models with high-dimensional data. arxiv:1702.04625.

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

Zimmert, M. and M. Lechner (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. arxiv:1908.08779.