Colangelo, Kyle; Lee, Ying-Ying

Working Paper

Double debiased machine learning nonparametric inference with continuous treatments

cemmap working paper, No. CWP72/19

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Colangelo, Kyle; Lee, Ying-Ying (2019) : Double debiased machine learning nonparametric inference with continuous treatments, cemmap working paper, No. CWP72/19, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2019.7219

This Version is available at:
https://hdl.handle.net/10419/241875

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 CWP72/19

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

December 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 nonparametric convergence rates. The nuisance estimators for the conditional expectation function and the conditional density 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. We implement various ML methods in Monte Carlo simulations and an empirical application on a job training program evaluation to support the theoretical results and demonstrate the usefulness of our DML estimator in practice.

Keywords: Average structural function, cross-fitting, dose-response function, 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, Midwest Econometrics Group, and California Econometrics Conference.

†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 achieve valid inference and to employ machine learning (ML) methods, we use a double debiased ML approach that combines a doubly robust moment function and cross-fitting.

We consider the fully nonparametric outcome equation $Y = g(T, X, \varepsilon)$. No functional form assumption is imposed on the general disturbances ε, such as restrictions on dimensionality, monotonicity, or separability. The potential outcome is $Y(t) = g(t, X, \varepsilon)$ indexed by the hypothetical treatment value t. The 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). The well-studied average treatment effect of switch from treatment t_1 to t_2 is $\beta_{t_2} - \beta_{t_1}$. We further define the partial or marginal effect of the first component of the continuous treatment T at $t = (t_1, ...t_{d_t})'$ to be $\theta_t = \partial\beta_t/\partial t_1$. In program evaluation, the average dose response function β_t shows how participants' labor market outcomes vary with the length of exposure to a job training program. In demand analysis when T contains price and income, the average structural function β_t can be the Engel curve. The partial effect θ_t reveals the average price elasticity at given values of price and income and hence captures the unrestricted heterogenous effects.

We define the *doubly robust* estimator for continuous treatments by

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

where $\hat{\gamma}(t, x)$ is an estimator of the conditional expectation function $\gamma(t, x) = \mathbb{E}[Y|T = t, X = x]$, $\hat{f}_{T|X}(t|x)$ is an estimator of the conditional density (or generalized propensity score) $f_{T|X}(t|x)$, and a kernel $K_h(T_i - t)$ weights observation i with treatment value around t in a certain distance of h. The number of such observations shrinks as the bandwidth h vanishes with the sample size n. Based on $\hat{\beta}_t^{DR}$, we propose a *double debiased machine learning* (DML) estimator with cross-fitting

[1]This commonly used identifying assumption based on observational data, also known as conditional independence and selection on observables, assumes that conditional on observables X, T is as good as randomly assigned, or conditionally exogenous.

via sample-splitting. Specifically a L-fold cross-fitting splits the sample into L subsamples. The nuisance estimators $\hat{\gamma}(t, X_i)$ and $\hat{f}_{T|X}(t|X_i)$ use observations in the other $L-1$ subsamples that do not contain the observation i. Then we estimate the partial effect θ_t by a numerical differentiation.

We show that the kernel-based DML estimators are asymptotically normal and converge at nonparametric rates. The 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 DML estimators. Thus the estimators of $\mathbb{E}[Y|T, X]$ and $f_{T|X}$ can be conventional nonparametric estimators, such as kernels or series, as well as modern ML methods, such as Lasso or neural networks; see Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) (CCDDHNR, hereafter) and Athey and Imbens (2019) for potential methods, such as ridge, boosted trees, random forest, and various ensembles of these methods. We also propose a ML estimator for the conditional density $f_{T|X}(t|x)$ for the low-dimensional T and high-dimensional X, which may be of independent interest.

We aim for a tractable inference procedure that is flexible to employ nonparametric or ML nuisance estimators and delivers a reliable distributional approximation in practice. Toward that end, the DML method contains two key ingredients: a doubly robust influence function and cross-fitting. The doubly robust influence function reduces sensitivity in estimating β_t with respect to nuisance parameters.[2] Using cross-fitting further removes bias induced by overfitting and achieves stochastic equicontinuity without strong entropy condition.[3] The usefulness of our DML estimator is demonstrated in simulations and an empirical example on the Job Corps program evaluation.

Our work builds on the results for semiparametric models in Ichimura and Newey (2017), Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018), and CCDDHNR and extends the literature to nonparametric continuous treatment/structural effects. 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

[2]Our estimator is doubly robust in the sense that $\hat{\beta}_t^{DR}$ consistently estimates β_t if either one of the nuisance functions $\mathbb{E}[Y|T, X]$ or $f_{T|X}$ is misspecified. The rapidly growing ML literature has utilized 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 (2019), Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018), CCDDHNR, Rothe and Firpo (2018), and references therein.

[3]CCDDHNR point out that the commonly used results in empirical process theory, such as Donsker properties, could break down in high-dimensional settings. For example, Belloni, Chernozhukov, Fernández-Val, and Hansen (2017) show how cross-fitting weakens the entropy condition and hence the sparsity assumption on nuisance Lasso estimator. 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.

the recent ML literature. 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 mean square convergence rates of the nuisance estimators can be weaker than that for the binary treatment due to the bandwidth h in our nonparametric DML estimator. Thus the ML and nonparametric nuisance estimators used in the semiparametric models in CCDDHNR can be applied here.

Our DML estimator is a simple modification of the binary treatment case in practice, yet we make non-trivial novel observations of distinct features of continuous treatments in theory: First we motivate the kernel-based moment function in $\hat{\beta}_t^{DR}$ by analytically calculating the limit of the Gateaux derivative, as in Ichimura and Newey (2017) and Carone, Luedtke, and van der Laan (2018). This calculation approximates the influence of a single observation on an estimator of β_t localized at t and hence is fundamental to construct estimators with desired properties, such as bias reduction, double robustness, and efficiency. The kernel function is a natural choice to approximate the distribution of a point mass and provides a simple moment function to characterize the *partial mean* structure of β_t, which fixes T at t and averages over the marginal distribution of X (Newey, 1994b). Neyman orthogonality holds for the moment function in $\hat{\beta}_t^{DR}$ as $h \to 0$. We can then define a "local Riesz representer" to be $K_h(T-t)/f_{T|X}(t|X)$.

A second motivation of the moment function is adding to the influence function of the regression (or imputation) 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 $\hat{\gamma}$ yields a different adjustment. These distinct features of continuous treatments are in contrast to the binary treatment case, where different nonparametric nuisance estimators of γ result in the same efficient influence function and unique Riesz representer. Therefore we provide a foundational justification for the proposed kernel-based DML estimator.

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 partially 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 partially linear model has many applications and is one of the leading examples in the recent ML literature, where the nuisance function $g(X)$ can be high-dimensional and estimated by a ML method.[4] Another semiparametric parameter of interest is the weighted average of β_t or θ_t over a range of treatment values t, such as the average derivative that reveals certain aggregate effects

[4]See Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018), CCDDHNR, and references therein. Demirer, Syrgkanis, Lewis, and Chernozhukov (2019) and Oprescu, Syrgkanis, and Wu (2019) extend to more general functional forms. Cattaneo, Jansson, and Newey (2018a,b), and Cattaneo, Jansson, and Ma (2019) propose different approaches.

(Powell, Stock, and Stoker, 1989) and the bound of the average welfare effect in Chernozhukov, Hausman, and Newey (2019). In contrast, our average structural function β_t and the partial effect θ_t capture the fully nonparametric heterogenous effect. Our estimator utilizes the kernel function $K_h(T_i - t)$ for the continuous treatments T of fixed dimension and averages out the high-dimensional covariates X, so we can maintain the nonparametric nature and circumvent the complexity of the nuisance parameter space.

To the best of our knowledge, we are among the first to apply the DML approach to inference on the average structural function and the partial effect of continuous treatments. They are nonparametric objects that cannot be estimated at the regular root-n rate. There is a small yet growing literature on employing the DML approach for nonparametric objects. 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 partial effect of continuous treatments. The causal objects of interest are different.[5]

Our paper also adds to the literature on 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 nuisance estimators to select the high-dimensional covariates X via a localized method of L_1-penalization at each t. In contrast, we use cross-fitting and provide high-level conditions that facilitates a variety of nonparametric and ML methods under mild assumptions. Kennedy, Ma, McHugh, and Small (2017) and Kallus and Zhou (2018) propose different versions of the doubly robust estimators.[6] 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. The empirical applications in Flores, Flores-Lagunes, Gonzalez, and Neumann (2012) and Kluve, Schneider, Uhlendorff,

[5]In particular, Chernozhukov, Newey, Robins, and Singh (2019) provide L_1-regularization 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 $\lim_{h\to 0} \int\int \gamma(T, X)K_h(T - t)/\mathbb{E}[K_h(T - t)]dF_{TX}(T, X) = \int \gamma(t, X)dF_{X|T}(X|t) = \mathbb{E}[Y(t)|T = t]$, while we identify the average structural function $\beta_t = \mathbb{E}[Y(t)]$ by $\lim_{h\to 0} \int\int \gamma(T, X)K_h(T - t)/f_{T|X}(t|X)dF_{TX}(T, X) = \int \gamma(t, X)dF_X(X)$. See more details in Section 3.1.1.

[6]Kallus and Zhou (2018) assume a known $f_{T|X}$. Kennedy, Ma, McHugh, and Small (2017) construct a "pseudo-outcome" that is motivated from the doubly robust and efficient influence function of the regular semiparametric parameter $\int \beta_t f_T(t)dt$. Then they regress the pseudo-outcome on T at t to estimate β_t. In contrast, we motivate the moment function of our DML estimator directly from β_t via the Gateaux derivative or the first-step adjustment. Moreover cross-fitting weakens the standard uniform entropy condition on the first-step nuisance estimators for high-dimensional X.

and Zhao (2012) focus on semiparametric results. We extend this literature to high-dimensional settings enabling ML methods for nonparametric inference.

The paper proceeds as follows. We introduce the framework and estimation procedure in Section 2. Section 3 presents the asymptotic theory. Section 4 provides numerical examples of Monte Carlo simulations and an empirical illustration using various ML methods. All the proofs are in the Appendix.

2 Setup and estimation

We give identifying assumptions and introduce the double debiased machine learning estimator.

Assumption 1 *Let $Y = g(T, X, \varepsilon)$. Let $\{Y_i, T_i', X_i'\}_{i=1}^n$ be an i.i.d. sample from $Z = \{Y, T', X'\}' \in \mathcal{Z} = \mathcal{Y} \times \mathcal{T} \times \mathcal{X} \subseteq \mathcal{R}^{1+d_t+d_x}$ with a cumulative distribution function (CDF) $F_{YTX}(Y, T, X)$.*
(i) (Conditional independence) T and ε are independent conditional on X.[7]
(ii) (Common support) For any $t \in \mathcal{T}$ and $x \in \mathcal{X}$, $f_{T|X}(t|x)$ is bounded away from zero.

The product kernel is defined by $K_h(T_i - t) = \Pi_{j=1}^{d_t} k((T_{ji} - t_j)/h)/h$, where T_{ji} is the j^{th} compoment of T_i and the kernel function $k()$ satisfies Assumption 2.

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 for any $t \in \mathcal{T}$,

$$\beta_t = \mathbb{E}[Y(t)] = \int_{\mathcal{X}} \mathbb{E}[Y|T = t, X] dF_X(X) = \mathbb{E}\big[\gamma(t, X)\big] \tag{2}$$

$$= \lim_{h \to 0} \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} \frac{K_h(T - t)Y}{f_{T|X}(t|X)} dF_{YTX}(Y, T, X) = \lim_{h \to 0} \mathbb{E}\left[\frac{K_h(T - t)Y}{f_{T|X}(t|X)}\right]. \tag{3}$$

The expression in equation (2) motivates the class of regression (or imputation) based estimators, while equation (3) motivates the class of inverse probability weighting estimators; see Section 3.2 for further discussion. Now we introduce the double debiased machine learning estimator.

[7] Equivalently T and the potential outcome $Y(t) = g(t, X, \varepsilon)$ are independent conditional on X for any $t \in \mathcal{T}$.

Estimation procedure

Step 1. (Cross-fitting) For some $L \in \{2, ..., n\}$, partition the observation indices into L groups I_ℓ, $\ell = 1, ..., L$. For each $\ell = 1, ..., L$, the estimators $\hat{\gamma}_\ell(t, x)$ for $\gamma(t, x) = \mathbb{E}[Y|T = t, X = x]$ and $\hat{f}_\ell(t|x)$ 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

$$\hat{\beta}_t = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \left\{ \hat{\gamma}_\ell(t, X_i) + \frac{K_h(T_i - t)}{\hat{f}_\ell(t|X_i)} \left(Y_i - \hat{\gamma}_\ell(t, X_i)\right) \right\}. \tag{4}$$

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 $n \to \infty$. We estimate the partial effect of the first component of the continuous treatment $\theta_t = \partial \beta_t / \partial t_1$ by $\hat{\theta}_t = (\hat{\beta}_{t^+} - \hat{\beta}_{t^-})/\eta$.

To simplify notation, we denote the $L_2(F)$-norm of a random vector (T, X) with distribution F_{TX} by $\|\hat{f}_\ell - f_{T|X}\|_{F,2} = \left(\int_{\mathcal{T}} \int_{\mathcal{X}} \left(\hat{f}_\ell(T|X) - f_{T|X}(T|X)\right)^2 dF_{TX}(T, X)\right)^{1/2}$ and $\|\hat{\gamma}_\ell - \gamma\|_{F,2} = \left(\int_{\mathcal{T}} \int_{\mathcal{X}} \left(\hat{\gamma}_\ell(T, X) - \gamma(T, X)\right)^2 dF_{TX}(T, X)\right)^{1/2}$, for each $\ell = 1, ..., L$.

Assumption 3 (Nuisance estimators) *For each $\ell = 1, ..., L$, (i) $\|\hat{\gamma}_\ell - \gamma\|_{F,2} \xrightarrow{p} 0$, $\|\hat{f}_\ell - f_{T|X}\|_{F,2} \xrightarrow{p} 0$; (ii) $\sqrt{nh^{d_t}} \|\hat{\gamma}_\ell - \gamma\|_{F,2} \|\hat{f}_\ell - f_{T|X}\|_{F,2} \xrightarrow{p} 0$.*

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 faster than $1/\sqrt{nh^{d_t}}$, which is slower than $1/\sqrt{n}$ in the semiparametric problem. The convergence rates in Assumption 3 are available for kernel or series estimators, the neural networks in Chen (2007) and Farrell, Liang, and Misra (2019), the Lasso and its variants in Belloni, Chernozhukov, and Hansen (2014), Farrell (2015), and Su, Ura, and Zhang (2019), for example.

When there is no sample splitting $L = 1$, $\hat{\gamma}_1$ and $\hat{f}_1$ use all observations in the full sample. Then the DML estimator $\hat{\beta}_t$ in (4) is the doubly robust estimator $\hat{\beta}_t^{DR}$ in equation (1). When $L = n$, $\hat{\beta}_t$ is known as the leave-one-out estimator. The numerical differentiation estimator $\hat{\theta}_t$ is simple and avoids estimating the derivatives of the nuisance parameters. All our results are readily extended to include discrete treatments D at the cost of notational complication. Specifically the doubly robust estimator in (1) becomes $\hat{\beta}_{td}^{DR} = n^{-1} \sum_{i=1}^{n} \left\{ \hat{\gamma}(t, d, X_i) + \mathbf{1}\{D_i = d\} K_h(T_i - t) \big(Y_i - \hat{\gamma}(t, d, X_i)\big) / \hat{f}_{TD|X}(t, d|X_i) \right\}$, where $\gamma(t, d, X_i) = \mathbb{E}[Y|T = t, D = d, X = X_i]$ and $f_{TD|X}(t, d|X_i) = f_{T|DX}(t|d, X_i) \Pr(D = d|X = X_i)$.

2.1 Conditional density estimation

We propose a simple estimator of the generalized propensity score (GPS) $f_{T|X}$ that allows us to use various nonparametric and ML methods designed for the conditional mean. We provide a uniform convergence rate to verify Assumption 3. The theory of ML methods in estimating the conditional density is less developed comparing with estimating the conditional mean. Alternative estimators can be the kernel density estimator, the artificial neural network in Chen and White (1999), or the Lasso in Su, Ura, and Zhang (2019). In Section 3.1.1, we discuss an alternative estimator using the L_1-regularized methods in Chernozhukov, Newey, Robins, and Singh (2019) to estimate the local Riesz representer without estimating the conditional density.

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, $\sup_{x\in\mathcal{X}}\left|\hat{\mathbb{E}}[W|X=x]-\mathbb{E}[W|X=x]\right| = O_p(R_1)$. Let $G(u)=\int_{-\infty}^{u} g(v)dv$ with a standard second-order kernel function $g()$. Let h_1 and ϵ be positive sequences vanishing as n grows. When $d_T=1$, we define the CDF estimator to be $\hat{F}_{T|X}(t|x)=\hat{\mathbb{E}}[G((t-T)/h_1)|X=x]$. Then we estimate the GPS by the numerical derivative estimator $\hat{f}_{T|X}(t|x)=(2\epsilon)^{-1}\big(\hat{F}_{T|X}(t+\epsilon|x)-\hat{F}_{T|X}(t-\epsilon|x)\big)$.[8] Lemma 1 below shows that the convergence rate $\sup_{x\in\mathcal{X}}\left|\hat{f}_{T|X}(t|x)-f_{T|X}(t|x)\right| = O_p(R_1\epsilon^{-1}+h_1^2\epsilon^{-1}+\epsilon^2)$ that is used to verify Assumption 3.

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 $\widehat{\partial F_{T|X}/\partial t_1}(t_1,t_2|x)=\big(\hat{F}_{T|X}(t_1+\epsilon,t_2|x)-\hat{F}_{T|X}(t_1-\epsilon,t_2|x)\big)/(2\epsilon)$. 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 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 $\sup_{x\in\mathcal{X}}\left|\hat{f}_{T|X}(t|x)-f_{T|X}(t|x)\right| = O_p(R_1\epsilon^{-d_T}+h_1^2\epsilon^{-d_T}+\epsilon^2)$ for any $t\in\mathcal{T}$.*

[8] Similar numerical derivative approaches for Lasso have been used in Belloni, Chernozhukov, Fernández-Val, and Hansen (2017) and Su, Ura, and Zhang (2019). In particular, Su, Ura, and Zhang (2019) estimate $F_{T|X}$ by a logistic distributional Lasso regression. In contrast, we use a kernel g to smooth CDF estimates (such as a Gaussian kernel) and accommodate other ML methods.

3 Asymptotic theory

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

$$\sqrt{nh^{d_t}}\left(\hat{\beta}_t - \beta_t\right) = \sqrt{\frac{h^{d_t}}{n}} \sum_{i=1}^{n} \left\{ \frac{K_h(T_i - t)}{f_{T|X}(t|X_i)} (Y_i - \mathbb{E}[Y|T = t, X = X_i]) + \mathbb{E}[Y|T = t, X = X_i] - \beta_t \right\} + o_p(1) \tag{5}$$

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* $\mathsf{V}_t = \mathbb{E}\big[var[Y|T = t, X]/f_{T|X}(t|X)\big] \int_{-\infty}^{\infty} k(u)^2 du$ *and* $\mathsf{B}_t = \sum_{j=1}^{d_t} \mathbb{E}\left[\frac{1}{2}\frac{\partial^2}{\partial t_j^2}\mathbb{E}\left[Y|T = t, X\right] + \frac{\partial}{\partial t_j}\mathbb{E}\left[Y|T = t, X\right]\frac{\partial}{\partial t_j} f_{T|X}(t|X)/f_{T|X}(t|X)\right] \int_{-\infty}^{\infty} u^2 k(u) du$.

Note interestingly that the second part in the influence function in (5) $n^{-1}\sum_{i=1}^{n} \mathbb{E}[Y|T = t, X = X_i] - \beta_t = O_p(1/\sqrt{n}) = o_p(1/\sqrt{nh^{d_t}})$ 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 asymptotically linear representation. This is in contrast to the binary treatment case where $K_h(T_i - t)$ is replaced by $\mathbf{1}\{T_i - t\}$ in $\hat{\beta}_t$, so $\hat{\beta}_t$ converges at a root-n rate. Then the second part in (5) is of first-order for a binary treatment, resulting in 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.

Theorem 1 is fundamental for inference, such as constructing confidence intervals and the optimal bandwidth h that minimizes the asymptotic mean squared error. The leading bias arises

from the term associated with the kernel $K_h(T-t)$ in the influence function. 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 (5). Specifically $\hat{\mathsf{V}}_t = h^{d_t} n^{-1} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \hat{\psi}_{li}^2$, where the estimated influence function $\hat{\psi}_{li} = K_h(T_i - t)(Y_i - \hat{\gamma}_\ell(t, X_i))/\hat{f}_\ell(t|X_i) + \hat{\gamma}_\ell(t, X_i) - \hat{\beta}_t$. Then we can estimate the optimal bandwidth that minimizes the asymptotic mean squared error (AMSE) or the asymptotic integrated MSE given in the following corollary.

Corollary 1 (AMSE optimal bandwidth) *Let the conditions in Theorem 1 hold.*

(i) For $t \in \mathcal{T}$, if B_t is non-zero, then the bandwidth that minimizes the asymptotic mean squared error is $h_t^* = \left(d_t \mathsf{V}_t / \left(4\mathsf{B}_t^2\right)\right)^{1/(d_t+4)} n^{-1/(d_t+4)}$.

(ii) Consider an integrable weight function $w(t) : \mathcal{T} \mapsto \mathcal{R}$. The bandwidth that minimizes the asymptotic integrated MSE $\int_{\mathcal{T}} \left(\mathsf{V}_t/(nh) + h^4 \mathsf{B}_t^2\right) w(t) dt$ is $h_w^* = \left(d_t \mathsf{V}_w / \left(4\mathsf{B}_w\right)\right)^{1/(d_t+4)} n^{-1/(d_t+4)}$, where $\mathsf{V}_w = \int_{\mathcal{T}} \mathsf{V}_t w(t) dt$ and $\mathsf{B}_w = \int_{\mathcal{T}} \mathsf{B}_t^2 w(t) dt$.

A common approach is to choose an undersmoothing bandwidth h smaller than h_t^* such that the bias is first-order asymptotically negligible, i.e., $h^2\sqrt{nh^{d_t}} \to 0$. Then we can construct the usual $(1-\alpha) \times 100\%$ point-wise confidence interval $\left[\hat{\beta}_t \pm \Phi^{-1}(1-\alpha/2)\sqrt{\hat{\mathsf{V}}_t/(nh^{d_t})}\right]$, where Φ is the CDF of $\mathcal{N}(0,1)$.

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 $\partial_t^\nu = \partial^\nu g(t,\cdot)/\partial t^\nu$ denote the ν^{th} order partial derivative of a generic function g with respect to t and $\partial_t = \partial_t^1$ for brevity.

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) $\eta^{-1}h\|\hat{\gamma}_\ell - \gamma\|_{F,2} \xrightarrow{p} 0$, $\eta^{-1}h\|\hat{f}_\ell - f_{T|X}\|_{F,2} \xrightarrow{p} 0$; (b) $\eta^{-1}h\sqrt{nh^{d_t}}\|\hat{f}_\ell - f_{T|X}\|_{F,2}\|\hat{\gamma}_\ell - \gamma\|_{F,2} \xrightarrow{p} 0$. 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 $\mathsf{B}_t^\theta = \sum_{j=1}^{d_t} \mathbb{E}\Big[\Big(\partial_{t_j}^2 \partial_{t_1}\gamma(t,X) f_{T|X}(t|X)/2 + \partial_{t_j}\partial_{t_1}\gamma(t,X)\partial_{t_j} f_{T|X}(t|X) + \partial_{t_j}\gamma(t,X)\big(\partial_{t_j}\partial_{t_1} f_{T|X}(t|X) - \partial_{t_j} f_{T|X}(t|X)\partial_{t_1} f_{T|X}(t|X) f_{T|X}(t|X)^{-1}\big)\Big) f_{T|X}(t|X)^{-1}\Big] \int u^2 k(u) du$ and $\mathsf{V}_t^\theta = \mathbb{E}\left[var(Y|T=t,X)/f_{T|X}(t|X)\right] \int k'(u)^2 du$.

(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 for any* $t \in \mathcal{T}$, $\sqrt{nh^{d_t}\eta^2}(\hat{\theta}_t - \theta_t - h^2\mathsf{B}_t^\theta) \xrightarrow{d} \mathcal{N}(0, \mathsf{V}_t^\theta)$, *where* $\mathsf{V}_t^\theta = 2\mathbb{E}\big[var[Y|T=t,X]/f_{T|X}(t|X)\big]\big(\int_{-\infty}^{\infty} k(u)^2 du - \bar{k}(\rho)\big)$ *with the convolution kernel* $\bar{k}(\rho) = \int_{-\infty}^{\infty} k(u)k(u-\rho)du$ *and* $\mathsf{B}_t^\theta = \partial \mathsf{B}_t/\partial t_1$ *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 in line 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) and Carone, Luedtke, and van der Laan (2018). The partial mean β_t is a marginal integration over the conditional distribution of Y given (T,X) and the marginal distribution of 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 $Z = (Y, T', X')'$ 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

$$\begin{aligned}\lim_{h\to 0}\frac{d}{d\tau}\beta_t(F^{\tau h})\Big|_{\tau=0} &= \gamma(t,X) - \beta_t + \lim_{h\to 0}\int_{\mathcal{X}}\int_{\mathcal{Y}} \frac{y-\gamma(t,x)}{f_{T|X}(t|x)} f_{YTX}^h(y,t,x)dydx \\ &= \gamma(t,X) - \beta_t + \frac{Y-\gamma(t,X)}{f_{T|X}(t|X)}\lim_{h\to 0} f_T^h(t).\end{aligned} \tag{6}$$

Note that the last term in (6) is a partial mean that is a marginal integration over $\mathcal{Y}\times\mathcal{X}$, fixing the value of T at t. Thus the Gateaux derivative depends on the choice of f_T^h. We then choose

$f_Z^h(z) = K_h(Z-z)\mathbf{1}\{f^0(z) > h\}$, following Ichimura and Newey (2017), so

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

Theorem 1 in Ichimura and Newey (2017) shows that if a semiparametric estimator is asymptotically 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 the influence function.

Importantly the expression in (6) is the building block to construct estimators for β_t. To illustrate this point, next we discuss an alternative estimator.

3.1.1 Local Riesz representer

The above discussion on the Gateaux derivative suggests that the Riesz representer for the non-regular β_t is not unique and depends on the kernel or other methods for localization at t. We define the "local Riesz representer" to be $\alpha_{th}(T,X) = K_h(T-t)/f_{T|X}(T|X)$ indexed by the evaluation value t and the bandwidth of the kernel h. Our local Riesz representer $\alpha_{th}(T,X)$ satisfies $\beta_t = \int_{\mathcal{X}} \gamma(t,X) dF_X(X) = \lim_{h\to 0} \int_{\mathcal{X}} \int_{\mathcal{T}} \alpha_{th}(T,X)\gamma(T,X) dF_{TX}(T,X)$ for all γ with finite second moment, following the insight of the Riesz representation theorem for a regular parameter (Newey, 1994a). Then we can obtain the influence function by adding an adjustment term $\alpha_{th}(T,X)(Y - \gamma(T,X))$, which is the product of the local Riesz representer and the regression residual.

Instead of estimating the closed form of the Riesz representer, e.g., the conditional density in our case, Chernozhukov, Newey, Robins, and Singh (2019), Chernozhukov, Newey, and Singh (2019), and Chernozhukov, Hausman, and Newey (2019) directly approximate the Riesz representer by Lasso or Dantzig regularized learners. Then an alternative DML estimator of β_t is $n^{-1}\sum_{\ell=1}^L \sum_{i\in I_\ell} \{\hat{\gamma}_\ell(t,X_i) + \hat{\alpha}_\ell(T_i,X_i)\,(Y_i - \hat{\gamma}_\ell(t,X_i))\}$. Below we briefly discuss a new estimator that builds on and extends their approach to the average structural function of continuous treatments.[9]

Let $b(T,X)$ be a p-dimensional dictionary of basis functions, such as polynomials or splines. The estimator $\hat{\alpha}_\ell(T_i,X_i) = b(T_i,X_i)'\hat{\rho}_\ell$ uses the L_1-regularized methods developed for non-regular functionals of $\gamma(T,X)$ in Chernozhukov, Newey, Robins, and Singh (2019) but with a modified weight function ℓ_h and a modified $\hat{M}_\ell$, in their notations. Specifically we define $\ell_h(T,X) = \zeta_{th}(T) f_T(T)/f_{T|X}(T|X)$, where $\zeta_{th}(T) = K_h(T-t)/\mathbb{E}[K_h(T-t)]$. We use the novel form of $\hat{M}_\ell$ proposed by Chernozhukov, Hausman, and Newey (2019) with k^{th} component $\hat{M}_{\ell k} = (n -$

[9] We thank Whitney Newey for an insightful discussion that was the seed for this idea.

$n_\ell)^{-2} \sum_{i \notin I_\ell} \sum_{j \notin I_\ell} \hat{\zeta}_{th}(T_i) b_k(T_i, X_j)$, where n_ℓ is the sample size of group I_ℓ, for $\ell = 1, ..., L$ and $k = 1, ..., p$. These new modifications are motivated by re-expressing our average structural function as $\beta_t = \lim_{h \to 0} \int_{\mathcal{T}} \int_{\mathcal{X}} \zeta_{th}(T) \gamma(T, X) dF_X(X) dF_T(T)$. The key is that this expression contains two integrations over the marginal distributions of X and T respectively. Thus the sample analogue $\hat{M}_\ell$ and $\zeta_{th}(T)$ for localization at t account for the partial mean structure of β_t that is defined by a marginal integration over the marginal distribution of X, rather than the *joint* distribution of (T, X). It follows that the minimum distance Lasso or Dantzig method with an estimated $\mathbb{E}[K_h(T - t)]$ in ζ_{th} for the low-dimensional T avoids estimating $f_{T|X}(T|X)$ in $\ell_h(T, X)$.[10] A formal theory for this alternative estimator by extending Chernozhukov, Hausman, and Newey (2019) or Chernozhukov, Newey, Robins, and Singh (2019) is left for future research.

3.2 Adjustment for first-step kernel estimation

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

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

that is based on the identification in (2), and the inverse probability weighting (IPW) estimator

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

that is based on the identification in (3). 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 discussed 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 (1994b) and Lee (2018) have derived the asymptotically 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

[10] Chernozhukov, Hausman, and Newey (2019) estimate bounds on consumer surplus that is a weighted average of the average structural function β_t weighted by a specific $\zeta(T)$ and can be estimated at the regular root-n rate. Our expression of β_t shares the same form of the weighted average yet with a distinct weight function $\zeta_{th}(T)$ for localization and hence is estimated at a nonparametric rate.

adjustment for estimating the nuisance function $\gamma(t,X)$ to the original moment 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 asymptotically 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 moment 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 an 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.[11] Heuristically, let us consider a basis function $b(T,X)$, including raw variables (T,X) as well as interactions and other transformations of these variables. Computing $\int \hat{\gamma}(t,x)f(x)dx$ implies the adjustment term of the form $\mathbb{E}[b(t,X)]\left(n^{-1}\sum_{i=1}^n b(T_i,X_i)b(T_i,X_i)'\right)^{-1} n^{-1}\sum_{i=1}^n b(T_i,X_i)'\big(y_i-\gamma(T_i,X_i)\big) = n^{-1}\sum_{i=1}^n \lambda_{ti}\big(y_i-\gamma(T_i,X_i)\big)$, 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 conditional density $f_{T|X}(t|X)$ is not explicit in the weight λ_{ti}. Such adjustment term may motivate different estimators of β_t; see the approximate residual balancing estimator in Athey, Imbens, and Wager (2018), CEINR, and Demirer, Syrgkanis, Lewis, and Chernozhukov (2019), for example.

[11] 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.

3.3 Asymptotically linear representation

We give an outline of deriving the asymptoticaly linear representation in Theorem 1, following CEINR. The moment function for identification is $m(Z_i, \beta_t, \gamma) = \gamma(t, X_i) - \beta_t$ by equation (2), i.e., $\mathbb{E}[m(Z_i, \beta_t, \gamma(t, X_i))] = 0$ uniquely defines β_t. The adjustment term is $\phi(Z_i, \beta_t, \gamma, \lambda) = K_h(T_i - t)\lambda(t, X_i)\,(Y_i - \gamma(t, X_i))$, where $\lambda(t, x) = 1/f_{T|X}(t|x)$. The doubly robust moment function is $\psi(Z_i, \beta_t, \gamma, \lambda) = m(Z_i, \beta_t, \gamma(t, X_i)) + \phi(Z_i, \beta_t, \gamma(t, X_i), \lambda(t, X_i))$, as in equation (1).

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

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

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 in Assumption 3(i) and cross-fitting. The second-order remainder term (R2) is controlled by Assumption 3(ii).

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 influence function. In our continuous treatment case, the Neyman orthogonality holds as $h \to 0$. Under the conditions in Theorem 1, (R1-DR) is $O_p\big((\|\hat{\gamma} - \gamma\|_{F,2} + \|\hat{\lambda} - \lambda\|_{F,2})\sqrt{nh^{4+d_t}}\big) = o_p(1)$.

4 Numerical examples

This section provides numerical examples of Monte Carlo simulations and an empirical illustration. The estimation procedure of the proposed double debiased machine learning (DML) estimator is described in Section 2. For both the regression $\gamma(t, x) = \mathbb{E}[Y|T = t, X = x]$ and the generalized propensity score (GPS) $f_{T|X}$, we employ three machine learning methods: Lasso, random forest

(RF), and neural network (NN). We implement our DML estimator in Python, using the packages skelearn, pytorch, numpy, and scipy. Software is available from the authors.

4.1 Simulation study

We begin by describing the nuisance estimators for the simulation in more detail. **Lasso:** The penalization parameter is chosen via grid search utilizing tenfold cross validation in both estimators of γ and $f_{T|X}$ separately. The basis functions contain third-order polynomials of X and T, and interactions among X and T. **Random forest:** We use forests with 1,000 trees and 40 minimum observations per leaf, selected based on tenfold cross validation. **Neural network:** To estimate $\gamma(t, X)$, we use a neural network with 4 hidden layers. Each hidden layer has 10 neurons and uses scaled exponential linear unit (SELU) activation functions. For the GPS estimation we use a network with 2 hidden layers, each with 10 neurons and with sigmoid (logistic) activation functions. The weights are fit using stochastic gradient descent with a weight decay of 0.2.[12] For the selection of the neural network models, we perform a train-test split of the data and choose the models based on out-of-sample performance.

We consider the data-generating process: $\nu \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{N}(0,1)$,

$$X = (X_1, ..., X_{100})' \sim \mathcal{N}(0, \Sigma), \quad T = \Phi(3X'\theta) + 0.75\nu, \quad Y = 1.2T + 1.2X'\theta + T^2 + TX_1 + \varepsilon,$$

where $\theta_j = 1/j^2$, $diag(\Sigma) = 1$, the (i,j)-entry $\Sigma_{ij} = 0.5$ for $|i-j| = 1$ and $\Sigma_{ij} = 0$ for $|i-j| > 1$ for $i, j = 1, ..., 100$, and Φ is the CDF of $\mathcal{N}(0,1)$. Thus the potential outcome $Y(t) = 1.2t + 1.2X'\theta + t^2 + tX_1 + \varepsilon$. Let the parameter of interest be the average dose response function at $t = 0$, i.e., $\beta_0 = \mathbb{E}[Y(0)] = 0$.

We compare estimations with fivefold cross-fitting and without cross-fitting, and with a range of bandwidths to demonstrate robustness to bandwidth choice. We consider sample size $n \in \{500, 1000\}$ and the number of subsamples used for cross-fitting $L \in \{1, 5\}$. We use the second-order Epanechnikov kernel with bandwidth h. For the GPS estimator described in Section 2.1, we choose h_1 and ϵ to be h. Let the bandwidth of the form $h = c\sigma_T n^{-0.2}$ for a constant $c \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$ and the standard deviation σ_T of T. We compute the AMSE-optimal bandwidth h_0^* given in Corollary 1(i) that has the corresponding $c^* = 1.43$. Thus using some undersmoothing bandwidth with $c < c^*$, the 95% confidence interval $\left[\hat{\beta}_t \pm 1.96 s.e.\right]$ is asymptotically valid, where the standard error $(s.e.)$ is computed using the sample analogue of the estimated in-

[12] Weight decay is a form of regularization to prevent overfitting. Weight decay is a penalty where after each iteration the weights in the network are multiplied by $(1 - \alpha\lambda)$ before adding the adjustment in the direction of the gradient, where α is the learning rate (step size) and λ is the weight decay.

fluence function, as described in Section 3. We impose common support assumption by trimming the lowest 2.5% of the GPS estimates. All the results are based on 1,000 Monte Carlo simulations.

Table 1 reports the results. The DML estimators using these ML methods perform well in the case of cross-fitting ($L = 5$). Under no cross-fitting ($L = 1$), the confidence intervals based on Lasso and random forest have coverage rates lower than the nominal 95%. Cross-fitting substantially improves biases and coverage rates, as predicted by the theoretical results. Neural network performs relatively well without cross-fitting. Cross-fitting slightly increases the root-MSE (RMSE). Overall cross-fitting improves the coverage rates, and the results are more robust to bandwidth choice under cross-fitting.

Table 1: Simulation Results

			Lasso			RF			NN		
n	L	c	Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage
500	1	0.50	0.004	0.127	0.924	0.154	0.220	0.814	0.093	0.194	0.910
		0.75	0.015	0.119	0.922	0.098	0.177	0.875	0.091	0.175	0.904
		1.00	0.022	0.119	0.889	0.067	0.158	0.894	0.089	0.162	0.894
		1.25	0.037	0.118	0.897	0.055	0.136	0.934	0.087	0.158	0.879
		1.50	0.051	0.120	0.865	0.039	0.134	0.933	0.087	0.152	0.880
	5	0.50	-0.107	2.741	0.951	-0.011	0.205	0.960	-0.072	0.242	0.964
		0.75	-0.001	0.230	0.938	-0.009	0.173	0.955	-0.075	0.208	0.951
		1.00	0.014	0.193	0.941	-0.016	0.165	0.938	-0.081	0.192	0.944
		1.25	0.033	0.173	0.941	-0.011	0.141	0.959	-0.086	0.181	0.941
		1.50	0.045	0.161	0.948	-0.018	0.141	0.952	-0.093	0.179	0.930
1000	1	0.50	0.003	0.125	0.885	0.126	0.164	0.762	-0.018	0.139	0.957
		0.75	0.003	0.114	0.883	0.092	0.137	0.823	-0.014	0.117	0.958
		1.00	0.018	0.105	0.867	0.063	0.114	0.877	-0.013	0.106	0.958
		1.25	0.024	0.105	0.838	0.057	0.109	0.883	-0.013	0.098	0.959
		1.50	0.036	0.109	0.829	0.048	0.098	0.910	-0.008	0.090	0.957
	5	0.50	0.004	0.152	0.933	-0.009	0.140	0.968	-0.062	0.165	0.952
		0.75	0.002	0.132	0.931	-0.001	0.120	0.948	-0.060	0.142	0.942
		1.00	0.020	0.118	0.922	-0.008	0.108	0.947	-0.061	0.129	0.931
		1.25	0.024	0.112	0.914	0.000	0.102	0.948	-0.058	0.120	0.933
		1.50	0.035	0.111	0.894	0.000	0.094	0.944	-0.057	0.114	0.919

Notes: $L = 1$: no cross-fitting. $L = 5$: fivefold cross-fitting. The bandwidth is $h = c\sigma_T n^{-0.2}$, and $c = 1.43$ for the AMSE-optimal bandwidth. The nominal coverage rate of the confidence interval is 0.95.

4.2 Empirical illustration

We illustrate our method by reanalysing the Job Corps program in the United States, which was conducted in the mid-1990s. The largest publicly founded job training program targets disadvantaged youth. The participants are exposed to different numbers of actual hours of academic and vocational training. The participants' labor market outcomes may differ if they accumulate different amounts of human capital acquired through different lengths of exposure. We estimate the average dose response functions to investigate the relationship between employment and the length of exposure to academic and vocational training. As our analysis builds on Flores, Flores-Lagunes, Gonzalez, and Neumann (2012), Hsu, Huber, Lee, and Pipoz (2018), and Lee (2018), we refer the readers to the reference therein for further details of Job Corps.

We use the same dataset in Hsu, Huber, Lee, and Pipoz (2018). We consider the outcome variable (Y) to be the proportion of weeks employed in the second year following the program assignment. The continuous treatment variable (T) is the total hours spent in academic and vocational training in the first year. We follow the literature to assume the conditional independence Assumption 1(i), meaning that selection into different levels of the treatment is random, conditional on a rich set of observed covariates, denoted by X. The identifying Assumption 1 is indirectly assessed in Flores, Flores-Lagunes, Gonzalez, and Neumann (2012). Our sample consists of 4,024 individuals who completed at least 40 hours (one week) of academic and vocational training and 40 covariates measured at the baseline survey. Figure 1 shows the distribution of T by a histogram, and Table 2 provides brief descriptive statistics.

Implementation details We estimate the average dose response function $\beta_t = \mathbb{E}[Y(t)]$ and partial effect $\theta_t = \partial\mathbb{E}[Y(t)]/\partial t$ by the proposed DML estimator with fivefold cross-fitting. We implement the DML estimator with Lasso, random forest, and neural network for the nuisance parameters, respectively. The parameters for Lasso and random forest are selected as described in the simulation Section 4.1. For random forest, in the regression estimation of γ, we use 1000 trees and a minimum of 40 observations per leaf. For the GPS estimation we use 1000 trees with a minimum 220 observations per leaf. For neural network, the regression estimation of γ uses a neural network with two hidden layers and a weight decay of 0.1. The first hidden layer has one-hundred neurons and the second hidden layer has twenty neurons. The hidden layers use scaled exponential linear unit (SELU) activation functions. The output layer uses a linear activation function. The GPS estimation uses a network with 2 hidden layers and a weight decay of 0.2. Each with 10 neurons and with sigmoid (logistic) activation functions. The output layer uses a linear activation function.

We use the second-order Epanechnikov kernel with bandwidth h. For the GPS estimator, we choose h_1 and ϵ to be h. We compute the optimal bandwidth h_w^* that minimizes an asymptotic integrated MSE derived in Corollary 1(ii). A practical implementation is to choose the weight function $w(t) = \mathbf{1}\{t \in [\underline{t}, \bar{t}]\}/(\bar{t}-\underline{t})$ to be the density of $Uniform[\underline{t}, \bar{t}]$ on the interior support $[\underline{t}, \bar{t}] \subset \mathcal{T}$ of the support of the continuous treatment. Set m equally spaced grid points over $[\underline{t}, \bar{t}]$: $\{\underline{t} = t_1, t_2, ..., t_m = \bar{t}\}$. A plug-in estimator $\hat{h}_w^* = \left(\hat{\mathsf{V}}_w / \left(4\hat{\mathsf{B}}_w\right)\right)^{1/5} n^{-1/5}$, where $\hat{\mathsf{V}}_w = m^{-1} \sum_{j=1}^m \hat{\mathsf{V}}_{t_j} \mathbf{1}\{t_j \in [\underline{t}, \bar{t}]]\}/(\bar{t} - \underline{t})$ and $\hat{\mathsf{B}}_w = m^{-1} \sum_{j=1}^m \hat{\mathsf{B}}_{t_j}^2 \mathbf{1}\{t_j \in [\underline{t}, \bar{t}]]\}/(\bar{t} - \underline{t})$. We use $[\underline{t}, \bar{t}] = [160, 1840]$ and $t_2 - t_1 = 40$ in this empirical application. We then obtain bandwidths $0.8h_w^*$ for undersmoothing that are 349.27 for Lasso, 336.93 for RF, and 308.23 for NN.

Results Figure 2 presents the estimated average dose response function β_t along with 95% point-wise confidence intervals. The results for the three ML nuisance estimators have similar patterns. The estimates suggest an inverted-U relationship between the employment and the length of participation.

Figure 3 reports the partial effect estimates $\hat{\theta}_t$ with step size $\eta = 160$ (one month). Across all procedures, we see positive partial effects when hours of training are less than 500 and around 1000. Taking the estimates by neural network for example, $\hat{\beta}_{1000} = 45.49$ with standard error $s.e. = 0.466$ and $\hat{\theta}_{1000} = 0.0043$ with $s.e. = 0.0013$. This estimate implies that increasing the training from six months to seven months increases the average proportion of weeks employed in the second year by 0.7% with $s.e. = 0.215\%$.

We note that the empirical practice has focused on semiparametric estimation; see Flores, Flores-Lagunes, Gonzalez, and Neumann (2012), Hsu, Huber, Lee, and Pipoz (2018), Lee (2018), for example. Although our nonparametric DML estimates for β_t do not give significantly different results than the semiparametric estimates in Lee (2018), the semiparametric methods are subject to the risk of misspecification. Our DML estimator provides a solution to the challenge of implementing a fully nonparametric inference in practice.

5 Conclusion and outlook

This paper provides a nonparametric inference method for continuous treatment effects under unconfoundedness and in the presence of high-dimensional or nonparametric nuisance parameters. The proposed double debiased machine learning (DML) 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. For a future extension, our DML estimator serves as the

preliminary element for policy learning and optimization with a continuous decision, following Manski (2004), Hirano and Porter (2009), Kitagawa and Tetenov (2018), Kallus and Zhou (2018), Demirer, Syrgkanis, Lewis, and Chernozhukov (2019), Athey and Wager (2019), Farrell, Liang, and Misra (2019), among others.

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 average structural function of continuous treatments, as discussed in Section 3. Using such an influence function to construct the corresponding DML estimator is left for future research.

Figure 1: Histogram of Hours of Training

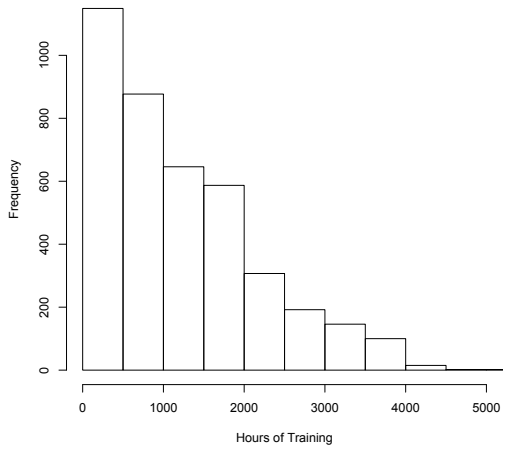


Table 2: Descriptive statistics

Variable	Mean	Median	StdDev	Min	Max
share of weeks employed in 2nd year (Y)	44	40.38	37.89	0	100
total hours spent in 1st-year training (T)	1219.8	992.86	961.74	40	6188.57

Notes: Summary statistics for 4,024 individuals who completed at least 40 hours of academic and vocational training.

Figure 2: Estimated average dose response functions and 95% confidence intervals

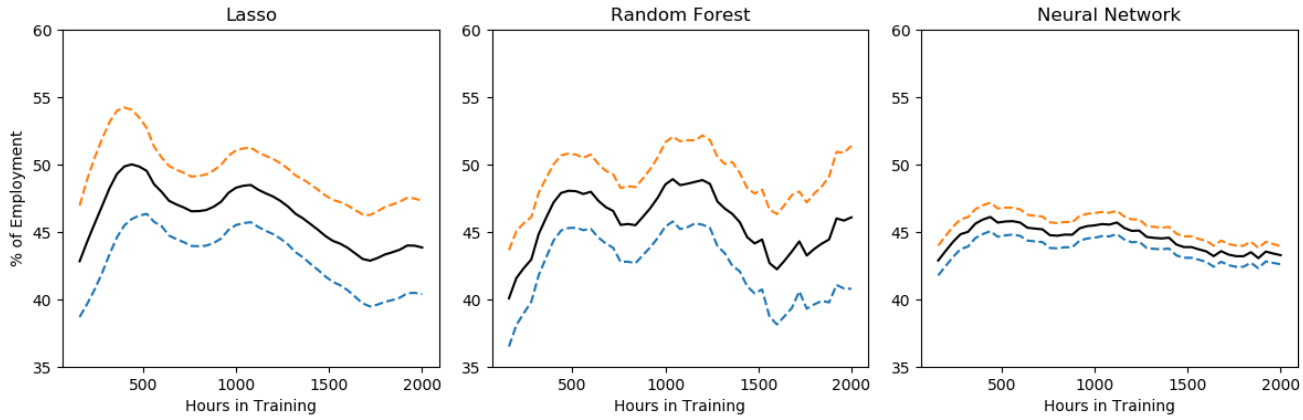
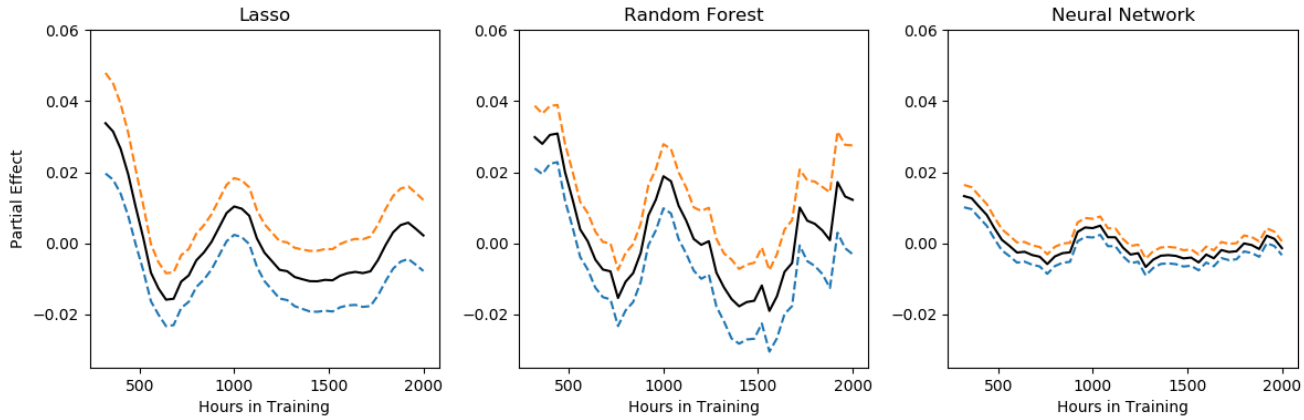


Figure 3: Estimated partial effects and 95% confidence interval



Appendix

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

$$\begin{aligned}
\left\|\hat{f}_{T|X}(t|X) - f_{T|X}(t|X)\right\|_\infty \leq\ & \left\|\hat{\mathbb{E}}\left[G((T-t)/h_1)|X\right] - \mathbb{E}[G((T-t)/h_1)|X]\right\|_\infty \epsilon^{-d_T} && (7)\\
&+ \left\|\mathbb{E}[G((T-t)/h_1)|X] - F_{T|X}(t|X)\right\|_\infty \epsilon^{-d_T} && (8)\\
&+ \left\|(\tilde{f}_{T|X}(t|X) - f_{T|X}(t|X)\right\|_\infty && (9)\\
&= O_p(R_1\epsilon^{-d_T} + h_1^2\epsilon^{-d_T} + \epsilon^2).
\end{aligned}$$

For (7), we give a crude bound by exploiting the convergence rate of the ML or nonparametric estimators. For (8), we follow the standard algebra for kernel, using integration by parts and change of variables. We analyze (9) below.

We first prove the results for $d_T = 1$. By a Taylor expansion, $F_{T|X}(t \pm \epsilon|x) = F_{T|X}(t|x) \pm \epsilon f_{T|X}(t|x) + \epsilon^2/2\ \partial f_{T|X}(t|x)/\partial t \pm \epsilon^3/3!\ \partial^2 f_{T|X}(t_\pm|x)/\partial t^2$ for some mean values $t_+ \in (t, t+\epsilon)$ 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 $F = F_{T|X}(t_1, t_2|x)$. For any positive sequences $\epsilon = (\epsilon_1, \epsilon_2)' \to 0$, let $F_{++} = F_{T|X}(t_1+\epsilon_1, t_2+\epsilon_2|x)$, $F_{+-} = F_{T|X}(t_1+\epsilon_1, t_2-\epsilon_2|x)$, $F_{-+} = F_{T|X}(t_1-\epsilon_1, t_2+\epsilon_2|x)$, $F_{--} = F_{T|X}(t_1-\epsilon_1, t_2-\epsilon_2|x)$, and $\partial_j^\nu F = \partial^\nu F_{T|X}(t|x)/\partial t_j^\nu$ 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 some 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 $\ell = 1, ..., L$ and for any $t \in \mathcal{T}$,*

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

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

Under Assumption 1(ii), Assumption 4 is implied by Assumption 3.[13] 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 $i \neq I_\ell$. Let $\hat{\gamma}_{il} = \hat{r}_\ell(t, X_i)$ using Z_ℓ^c for $i \in I_\ell$. 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 $i \in I_\ell$. By construction and independence of Z_ℓ^c and $z_i, i \in I_\ell$, $\mathbb{E}[\hat{\Delta}_{il}|Z_\ell^c] = 0$ and $\mathbb{E}[\hat{\Delta}_{il}\hat{\Delta}_{jl}|Z_\ell^c] = 0$ for $i, j \in I_\ell$. For $i \in I_\ell$ and for all t, $h\mathbb{E}[\hat{\Delta}_{il}^2|Z_\ell^c] = h\int(\hat{\gamma}_{il}-\gamma_i)^2 f_X(X_i)dX_i \xrightarrow{p} 0$ by Assumption 4(i). Then $\mathbb{E}\left[\left(\sqrt{h^{d_t}/n}\sum_{i\in I_\ell}\hat{\Delta}_{il}\right)^2 \middle| Z_\ell^c\right] = (h/n)\sum_{i\in I_\ell}\mathbb{E}\left[\hat{\Delta}_{il}^2\middle|Z_\ell^c\right] \leq h\int(\hat{\gamma}_{il}-\gamma_i)^2 f_X(X_i)dX_i \xrightarrow{p} 0$. The conditional Markov inequality implies that $\sqrt{h^{d_t}/n}\sum_{i\in I_\ell}\hat{\Delta}_{il} \xrightarrow{p} 0$.

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).

[13] We claim that Assumption 3(i) is implied by Assumption 4(i). Other conditions can be shown by analogous arguments. Denote $\hat{A}(t) = \int\left(\hat{\gamma}_\ell(t,x) - \gamma(t,x)\right)^2 f_{TX}(t,x)dx \geq 0$. 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 $n \geq N$.

In particular, for (R1-2), $h\mathbb{E}[\hat{\Delta}_{il}^2|Z_\ell^c] = O_p\left(\int k(u)^2du \int_{\mathcal{X}} \left(\hat{\lambda}_{il} - \lambda_i\right)^2 f_X(X_i)dX_i\right) \xrightarrow{p} 0$ by the smoothness condition and Assumption 4(i). So (R1-1)$\xrightarrow{p} 0$ and (R1-2)$\xrightarrow{p} 0$.

For (R2),

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

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)$= O_p((\|\hat{\gamma} - \gamma\|_{F,2} + \|\hat{\lambda} - \lambda\|_{F,2})\sqrt{nh^{d_t}}h^2) = o_p(1)$.

By the triangle inequality, we obtain the asymptoticaly 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, $\mathbb{E}\left[\frac{K_h(T-t)}{f_{T|X}(t|X)}\left(Y - \gamma(t, X)\right)\right] = \mathbb{E}\left[\frac{1}{f_{T|X}(t|X)}\mathbb{E}\left[K_h(T-t)\left(\gamma(T, X) - \gamma(t, X)\right)|X\right]\right]$. 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

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

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 (i) By Theorem 1, the asymptotic MSE is $h^4\mathsf{B}_t^2 + \mathsf{V}_t/(nh^{d_t})$. (ii) The asymptotic integrated MSE is $\int_{\mathcal{T}}\left(h^4\mathsf{B}_t^2 + \mathsf{V}_t/(nh^{d_t})\right)w(t)dt$. The results follow by solving the first-order conditions. □

Proof of Theorem 2 We decompose $\hat{\theta}_t - \theta_t = (\hat{\theta}_t - \theta_{t\eta}) + (\theta_{t\eta} - \theta_t)$, where $\theta_{t\eta} = (\beta_{t^+} - \beta_{t^-})/\eta$. 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 $\hat{\theta}_t - \theta_{t\eta} = \eta^{-1}n^{-1}\sum_{i=1}^n \left(\psi_{t^+i} - \psi_{t^-i} + R_{t^+i} - R_{t^-i}\right)$. Denote $f_{t|X_i} = f_{T|X}(t|X_i)$.

(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 $\int \left(\partial_{t_1}K_h(T-t)\right)^2\mathbb{E}\big[(Y - \gamma(t,X))^2|T,X\big]f_{T|X}/f_{t|X}^2 dT = h^{-(d_t+2)}\mathbb{E}\big[var(Y|T=t,X)/f_{T|X}(t|X)\big]\int k'(u)^2du + o(h^{-(d_t+2)}) = O\big(h^{-(d_t+2)}\big)$.

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,

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

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.[14] For any $F \in \mathcal{F}$,

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

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

The influence function can be calculated as

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

In particular, we specify F^h_Z following equation (3.1) in Ichimura and Newey (2017). Let $K_h(Z) = \Pi_{j=1}^{d_z} k(Z_j/h)/h$, 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

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

[14] Note that a nascent delta function to approximate the Dirac delta function is $K_h(T-t) = k((T-t)/h)/h$ 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.

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.

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.

Cattaneo, M. D., M. Jansson, and X. Ma (2019). Two-step estimation and inference with possibly many included covariates.

Cattaneo, M. D., M. Jansson, and W. Newey (2018a). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory 34*(2), 277–301.

Cattaneo, M. D., M. Jansson, and W. Newey (2018b). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association 113*(523), 1350–1361.

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

Chen, X. and H. White (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory 45*.

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., J. A. Hausman, and W. K. Newey (2019). Demand analysis with many prices. cemmap Working Paper, CWP59/19.

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., W. K. Newey, and R. Singh (2019). Learning L_2-continuous regression functionals via regularized riesz representers. arxiv:1809.05224v2.

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.

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 (2019). 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.

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.

Hirano, K. and J. Porter (2009). Asymptotics for statistical treatment rules. *Econometrica 77*, 1683–1701.

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.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica 72*(4), 1221–1246.

Newey, W. (1994a). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–1382.

Newey, W. K. (1994b). 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.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*(6), 1403–30.

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.