

Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments*

Kyle Colangelo Ying-Ying Lee[†]

University of California Irvine

December 2021

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 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 or ML methods. Utilizing a kernel-based doubly robust moment function and cross-fitting, we give high-level conditions under which the nuisance estimators do not affect the first-order large sample distribution of the DML estimators. We further provide sufficient low-level conditions for kernel and series estimators, as well as modern ML methods - generalized random forests and deep neural networks. We justify the use of kernel to localize the continuous treatment at a given value by the Gateaux derivative. We implement various ML methods in Monte Carlo simulations and an empirical application on a job training program evaluation.

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 Lee (February 2019), “Double machine learning nonparametric inference on continuous treatment effects.” We are grateful to Max Farrell, Whitney Newey, and Takuya Ura for valuable discussion. We thank seminar participants at Harvard-MIT, UC Irvine, Bonn-Mannheim, and USC, 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, California Econometrics Conference, and 2020 North American Winter Meeting of the Econometric Society, 2021 IAAE.

[†]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, under the unconfoundedness assumption¹ and in the presence of high-dimensional or nonparametric nuisance parameters. We focus on the heterogeneous 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 a fully nonparametric outcome equation $Y = g(T, X, \varepsilon)$. No functional form assumption is imposed on the unobserved 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 expected value of the potential outcome across observations with the observed and unobserved heterogeneity (X, ε) , i.e., $\beta_t \equiv \mathbb{E}[Y(t)] = \int \int g(t, X, \varepsilon) dF_{X\varepsilon}$. It is also known as the *average structural function* in nonseparable models in Blundell and Powell (2003). The well-studied average treatment effect of switching from treatment t to s is $\beta_s - \beta_t$. We further define the *partial (or marginal) effect* of the first component of the continuous treatment T at $t = (t_1, \dots, t_{d_t})'$ to be the partial derivative $\theta_t \equiv \partial \beta_t / \partial t_1$. 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 heterogeneous effects.

We are among the first to apply the double debiased machine learning approach to inference on the average structural function β_t and the partial effect θ_t of continuous treatments, to our knowledge. They are *non-regular nonparametric* objects that cannot be estimated at a root- n convergence rate. We propose a kernel-based *double debiased machine learning* (DML) estimator that utilizes a doubly robust moment function and cross-fitting via sample-splitting. The DML estimator uses the moment function

$$\gamma(t, X_i) + \frac{K_h(T_i - t)}{f_{T|X}(t|X_i)} (Y_i - \gamma(t, X_i)), \quad (1)$$

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

an estimator $\hat{\gamma}(t, x)$ of the conditional expectation function $\gamma(t, x) \equiv \mathbb{E}[Y|T = t, X = x]$, an estimator $\hat{f}_{T|X}(t|x)$ of the conditional density (or generalized propensity score) $f_{T|X}(t|x)$, and a kernel $K_h(T_i - t)$ that weights observation i with treatment value around t in a distance of h . The number of such observations shrinks as the bandwidth h vanishes with the sample size n . 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 . The DML estimator averages over the subsamples. Then we estimate the partial effect θ_t by a numerical differentiation.

The doubly robust moment function in equation (1) has appeared in Kallus and Zhou (2018) without asymptotic theory and has been extensively studied in Su, Ura, and Zhang (2019) for Lasso-type nuisance estimators. We utilize cross-fitting and provide high-level and low-level conditions that facilitate a variety of nonparametric and ML methods.

We show that the kernel-based DML estimators are asymptotically normal and converge at nonparametric rates. We provide high-level conditions under which the nuisance estimators $\hat{\gamma}(t, X_i)$ and $\hat{f}_{T|X}(t|X_i)$ do not affect the first-order asymptotic distribution of the DML estimators. Specifically our high-level conditions on the convergence rates use a *partial L_2 norm* that fixes the treatment value at t , in contrast to the standard L_2 norm that integrates over the joint distribution of (T, X) , i.e., the root-mean-squared rates.

We further give low-level conditions for conventional nonparametric kernel and series estimators, as well as modern ML methods: the generalized random forests in Athey, Tibshirani, and Wager (2019) and the deep neural networks in Farrell, Liang, and Misra (2021b). These results on the convergence rates of the nuisance estimators are new to the literature, to our best knowledge. In addition, we propose a generic 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. See Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) (CCDDHNR, hereafter) and Athey and Imbens (2019) for potential ML methods, such as ridge, boosted trees, and various ensembles of these methods.

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 moment function and cross-fitting. The doubly robust moment function reduces sensitivity in estimating β_t with respect to nuisance parameters.² Cross-fitting further removes bias induced by overfitting and achieves

²Our estimator is doubly robust in the sense that it 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 et al. (2014), Farrell (2015), Belloni et al. (2017), Farrell et al. (2021b), Chernozhukov et al.

stochastic equicontinuity without strong entropy conditions.³

Our work builds on the results for semiparametric models in Ichimura and Newey (2017), Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2018) (CEINR, hereafter), 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 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.

Our DML estimator utilizes the kernel function $K_h(T_i - t)$ for the continuous treatments T of fixed low 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. Our kernel-based estimator appears to be a simple modification of the binary treatment case in practice, yet we make non-trivial new observations on distinct features of continuous treatments in theory:

First, from the literature on estimating regular parameters, the Gateaux derivative is fundamental to construct estimators with desired properties, such as bias reduction and double robustness (Ichimura and Newey (2017) and Carone, Luedtke, and van der Laan (2018)). We show that one important feature of *non-regular nonparametric* parameters is that the Gateaux derivative and the Riesz representer are not unique, depending on what we choose to approximate the continuous treatment distribution of a point mass. And the kernel function is a natural choice. Neyman orthogonality holds as $h \rightarrow 0$ (Neyman, 1959). This is in strong contrast to the binary treatment case and regular semiparametric parameters. Moreover, to construct the DML estimator of a linear functional of β_t that preserves the good properties, the corresponding moment function is simply the linear functional of the moment function of β_t . Therefore we provide a foundational justification for the proposed kernel-based DML estimator, relative to alternative approaches, such as Kennedy, Ma, McHugh, and Small (2017) and Semenova and Chernozhukov (2020).⁴ To the best of our knowledge, this is the first explicit calculation of Gateaux derivative for such a non-regular

(2018), CCDDHNR, Rothe and Firpo (2019), and references therein.

³CCDDHNR 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 et al. (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.

⁴Kennedy et al. (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 locally regress the pseudo-outcome on T at t using a kernel to estimate β_t . Semenova and Chernozhukov (2020), an updated version of Chernozhukov and Semenova (2019), illustrate in an example to estimate β_t by the best linear projection of an “orthogonal signal of the outcome” which is the same “pseudo-outcome” proposed by Kennedy et al. (2017). In contrast, we motivate the moment function of our DML estimator directly from β_t via the Gateaux derivative or the first-step adjustment.

nonparametric parameter.

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 $\hat{\gamma}$ under the low-dimensional case when the dimension of X_i is fixed. A series estimator $\hat{\gamma}$ yields a different adjustment. These distinct features of continuous treatments are again in contrast to the regular binary treatment case, where different nonparametric nuisance estimators $\hat{\gamma}$ result in the same efficient influence function.

There is a small yet growing literature on employing the DML approach for non-regular 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 of the high-dimensional X is studied in Chernozhukov, Newey, Robins, and Singh (2019), Chernozhukov and Semenova (2019), Fan, Hsu, Lieli, and Zhang (2021), and Zimmert and Lechner (2019). Their causal objects of interest are different from our average structural function and partial effect of continuous treatments. As our DML estimator, most of the papers mentioned above use a kernel to localize the low-dimensional X_1 , while Chernozhukov and Semenova (2019) use a series-based localization.

Our paper also adds to the literature on continuous treatment effects estimation. In low-dimensional settings, see Imbens (2000), 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 Lettry (2020) study a class of inverse probability weighting estimators. The empirical applications in Flores, Flores-Lagunes, Gonzalez, and Neumann (2012) and Kluge, Schneider, Uhlenborff, and Zhao (2012) focus on semiparametric results. We extend this literature to high-dimensional settings and enable ML methods for nonparametric inference in practice.

A 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 nonseparable 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.⁵ 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 summarizes certain aggregate effects (Powell, Stock, and Stoker, 1989) and the bound of the average welfare effect in

⁵See CEINR, CCDDHNR, and references therein. Demirer et al. (2019) and Oprescu et al. (2019) extend to more general functional forms. Cattaneo et al. (2018a), Cattaneo et al. (2018b), Cattaneo et al. (2019), and Farrell et al. (2021a) propose different approaches.

Chernozhukov, Hausman, and Newey (2019). In contrast, our average structural function β_t and the partial effect θ_t capture the fully nonparametric heterogeneous effects of T .

The paper proceeds as follows. We introduce the framework and estimation procedure in Section 2. Section 3 presents the asymptotic theory point-wise in t and low-level conditions for various ML methods. We also discuss uniform inference over t by a multiplier bootstrap method. Section 4 justifies our kernel-based DML estimator. Section 5 demonstrates the usefulness of our DML estimator with various ML methods in Monte Carlo simulations and an empirical example on the Job Corps program evaluation. All the proofs are in the Appendix.

2 Setup and estimation

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) $F_{YTX}(Y, T, X)$. We give assumptions and introduce the double debiased machine learning estimator.

Assumption 1 (i) (Conditional independence) T and ε are independent conditional on X .⁶
(ii) (Common support) $f_{T|X}(t|x)$ is bounded away from zero almost everywhere (a.e.).
(iii) For $(y, t', x')' \in \mathcal{Z}$, $f_{YTX}(y, t, x)$ is three-times differentiable with respect to t .

Define the product kernel as $K_h(T_i - t) \equiv \prod_{j=1}^{d_t} k((T_{ji} - t_j)/h)/h^{d_t}$, where T_{ji} is the j^{th} component of T_i and the kernel function $k(\cdot)$ satisfies Assumption 2.

Assumption 2 (Kernel) The second-order symmetric kernel function $k(\cdot)$ (i.e., $\int k(u)du = 1$, $\int uk(u)du = 0$, and $0 < \int u^2k(u) < \infty$.) is bounded differentiable. For some finite positive constants C, \bar{U} , and for some $\nu > 1$, $|dk(u)/du| \leq C|u|^{-\nu}$ for $|u| > \bar{U}$.

Assumption 2 is standard in nonparametric kernel estimation and holds for commonly used kernel functions, such as Epanechnikov and Gaussian. By Assumptions 1-2 and the same reasoning for the binary treatment, it is straightforward to show the identification for any interior $t \in \mathcal{T}$,

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

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

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 4.2

⁶Equivalently T and the potential outcome $Y(t) = g(t, X, \varepsilon)$ are independent conditional on X for any t .

for further discussion. Now we introduce the double debiased machine learning estimator.

Estimation procedure

Step 1. (Cross-fitting) For some fixed $L \in \{2, \dots, n\}$, partition the observation indices into L groups I_ℓ , $\ell = 1, \dots, L$. For each $\ell = 1, \dots, L$, the estimators $\hat{\gamma}_\ell(t, x)$ for $\gamma(t, x) \equiv \mathbb{E}[Y|T = t, X = x]$ 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. (Double robustness) The double debiased ML (DML) estimator is defined as

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

Step 3. (Partial effect) Let $t^+ \equiv (t_1 + \eta/2, t_2, \dots, t_{d_t})'$ and $t^- \equiv (t_1 - \eta/2, t_2, \dots, t_{d_t})'$, where η is a positive sequence converging to zero as $n \rightarrow \infty$. We estimate the partial effect of the first component of the continuous treatment $\theta_t \equiv \partial \beta_t / \partial t_1$ by $\hat{\theta}_t \equiv (\hat{\beta}_{t^+} - \hat{\beta}_{t^-}) / \eta$.

Assumption 3 For each $\ell = 1, \dots, L$ and for any $t \in \mathcal{T}$,

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

$$(ii) \sqrt{nh^{d_t}} \left(\int_{\mathcal{X}} (\hat{f}_\ell(t|x) - f_{T|X}(t|x))^2 f_{TX}(t, x) dx \right)^{1/2} \left(\int_{\mathcal{X}} (\hat{\gamma}_\ell(t, x) - \gamma(t, x))^2 f_{TX}(t, x) dx \right)^{1/2} \xrightarrow{p} 0.$$

In Section 3.1, we provide sufficient low-level conditions for Assumption 3 when the nuisance estimators are kernel estimators, series, the generalized random forests in Athey, Tibshirani, and Wager (2019), and the deep neural networks in Farrell, Liang, and Misra (2021b).

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 considered in Kallus and Zhou (2018) and Su, Ura, and Zhang (2019). The numerical differentiation estimator $\hat{\theta}_t$ is simple and avoids estimating the derivatives of the nuisance parameters.

Our results are readily extended to include binary/multivalued treatments D at the cost of notational complication, e.g., Cattaneo (2010) for the low-dimensional setting. Specifically, the frequency method replaces the kernel with an indicator function: $\hat{\beta}_{td} = n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} \{ \hat{\gamma}_\ell(t, d, X_i) + \mathbf{1}\{D_i = d\} K_h(T_i - t) (Y_i - \hat{\gamma}_\ell(t, d, X_i)) / \hat{f}_{TD|X_i}(t, d|X_i) \}$, where $\gamma(t, d, X_i) = \mathbb{E}[Y|T = t, D = d, X = X_i]$ and $f_{TD|X}(t, d|X_i) = f_{T|DX}(t|d, X_i) \Pr(D = d|X = X_i)$. There is a literature on the kernel smoothing of discrete (categorical) variables (Aitchison and Aitken (1976), Ouyang, Li, and Racine (2009) and reference therein); such extension to smoothing discrete treatments is out of the scope of this paper.

Remark 1 (Common support) Assumption 1(ii) implies that we need to observe sufficient individuals in the population who can find a match sharing the same value of the control variable X and receiving the counterfactual value t . An analogous assumption in the binary treatment case is that the propensity score is bounded away from zero, e.g., Hirano, Imbens, and Ridder (2003). The common support assumption is standard, although it may be strong in some applications. For the binary treatment case, Khan and Tamer (2010) study extensively irregular identification and inverse weight estimation, when the propensity score can be close to zero as a small denominator. For the continuous treatment case, the convergence rate of $\hat{\beta}_t$ might similarly be affected if the generalized propensity score can be close to zero. We believe that this interesting extension is beyond the scope of the paper and is worthy of a separate research project. See also Su, Ura, and Zhang (2019) for the related discussion.

Another possible approach to relaxing Assumption 1(ii) is to define the object of interest by a common support via fixed trimming, e.g., Lee (2018). Without imposing this common support assumption, we instead focus on the causal object defined by a common support for the subpopulation whose control variable takes values in a common support $\mathcal{X}^* \equiv \{x : \inf_{t \in \mathcal{T}^*} f_{T|X}(t|x) \geq c\} \subseteq \cap_{t \in \mathcal{T}^*} \text{Supp}(X|T = t)$, which is a subset of the intersection of the supports of X conditional on $T = t$ for $t \in \mathcal{T}^* \subset \mathcal{T}$ for a positive constant c . Intuitively it is reasonable to focus on the subpopulation who has nontrivial chance to receive the counterfactual value t . More specifically, define the trimming function $\pi(x) \equiv \mathbf{1}\{\inf_{t \in \mathcal{T}^*} f_{T|X}(t|x) \geq c\}$ to select the common support \mathcal{X}^* . Then define the causal object of interest by $\int_{\mathcal{X}} \mathbb{E}[Y(t)|X = x] \pi(x) f_X(x) dx = \int_{\mathcal{X}} \mathbb{E}[Y|T = t, X = x] \pi(x) f_X(x) dx$. It is straightforward to include $\pi(x)$ in all our results at the cost of notational complication.

2.1 Conditional density estimation

We propose an 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 convergence rate to verify Assumption 3. The theory of ML methods in estimating the conditional density is less developed compared with estimating the conditional mean. Alternative estimators can be the kernel density estimator, the artificial neural networks in Chen and White (1999), or the Lasso method in Su, Ura, and Zhang (2019).

Let $\hat{\mathbb{E}}[W|X]$ be an estimator of the conditional mean $\mathbb{E}[W|X]$ for a bounded random variable W . Suppose a root-mean-squared convergence rate is available, $(\int_{\mathcal{X}} (\hat{\mathbb{E}}[W|X = x] - \mathbb{E}[W|X = x])^2 f_X(x) dx)^{1/2} = O_p(R_1)$ for a sequence of constants $R_1 = R_{1n}$. We estimate $f_{T|X}(t|x)$ by $\hat{f}_{T|X}(t|x) = \hat{\mathbb{E}}[g_{h_1}(T - t)|X = x]$, where the bandwidth h_1 is a positive sequence vanishing as n grows, the product kernel $g_{h_1}(T_i - t) \equiv \prod_{j=1}^{d_t} g((T_{ji} - t_j)/h_1)/h_1^{d_t}$, and $g(\cdot)$ satisfies Assumption 2

with $g()$ replacing $k()$ and with an unbounded support. We can choose $g()$ to be the Gaussian kernel.⁷ In Section 3.1, we demonstrate this generic GPS estimator using the kernel, series, the generalized random forests in Athey, Tibshirani, and Wager (2019), and the deep neural networks in Farrell, Liang, and Misra (2021b) and how Assumption 3 can be verified.

We can view our approach as estimating the ratio $K_h(T_i - t)/f_{T|X}(t|X_i)$ (or the Riesz representer in Section 4.1) by $K_h(T_i - t)/\hat{\mathbb{E}}[g_{h_1}(T - t)|X = X_i]$ with flexible nonparametric and ML methods. A possible drawback of using the proposed GPS estimator is that the estimate could be negative or small in finite samples. In practice, the small denominator problem often occurs due to observations near the tails of the distribution. An advantage of our kernel-based approach is that intuitively, when we choose the kernel k with a bounded support and $h < h_1$, $K_h(T_i - t)$ in the numerator serves as a trimming function to mitigate the possibly small denominator. When T_i is far from the target value t and $\hat{f}_{T|X}(t|X_i)$ is small, a bounded-support kernel $K_h(T_i - t) = 0$ excluding this observation i . The Monte Carlo simulations in Section 5.1 support that our DML estimator with various ML methods performs well without additional trimming.⁸

Lemma 1 (GPS) *Let $f_{T|X}(t|x)$ be $(d_t + 1)$ -times differentiable with respect to t for any $x \in \mathcal{X}$ and $\|f_{T|X}(t|\cdot)\|_\infty \equiv \sup_{(t,x) \in \mathcal{T} \times \mathcal{X}} \leq C$ that is a positive constant. Then $(\int_{\mathcal{X}} (\hat{f}_{T|X}(t|x) - f_{T|X}(t|x))^2 f_{TX}(t, x) dx)^{1/2} = O_p(R_1 h_1^{-d_t} + h_1^2)$ for any $t \in \mathcal{T}$.*

Assumption 3 specifies the root-mean-squared convergence rate of our conditional density estimator. Note that our conditional density estimator is a regression of a kernel function of the continuous treatment on the covariates X , i.e., the treatment value t is in the dependent variable and the conditioning variables are the covariates. Thus as long as the root-mean-squared convergence rate of a ML method is available, Assumption 3 is satisfied with a suitable bandwidth h_1 . Then we are able to use such a ML method to estimate the conditional density, as discussed in Section 3.1.

We note that the convergence rate provided in Lemma 1 is not sharp. We may attain a tighter bound if we could compute the mean-squared error directly; for example, see a kernel-based estimator in Section 3.1.1. Nevertheless the proposed estimator and Lemma 1 offer a convenient

⁷We thank one referee who suggested this estimator that we had considered at the early stage of this project, where we used a kernel function g with a *bounded* support in Monte Carlo simulations. The resulting dependent variable $g_{h_1}(T - t)$ in the regression estimator $\hat{\mathbb{E}}[g_{h_1}(T - t)|X]$ has a mass point at zero, so the estimator and the corresponding DML estimator $\hat{\beta}_t$ performed poorly in the simulation study. In this version, we require the kernel function g to have an *unbounded* support, so the distribution of $g_{h_1}(T - t)$ is continuous.

⁸We may adopt the same trimming approach to addressing this concern in the literature. For example, following Hsu et al. (2020), we can use the estimate $\max\{\hat{f}_{T|X}(t|X_i), \epsilon_n\}$ for some positive sequence $\epsilon_n \rightarrow 0$. We may further normalize the ratio weight by dividing $K_h(T_i - t)/\hat{f}_{T|X}(t|X_i)$ with $\sum_{i=1}^n K_h(T_i - t)/\hat{f}_{T|X}(t|X_i)$, following Imbens (2004) for the binary treatment, Flores et al. (2012), and Kallus and Zhou (2018).

bound for estimating the GPS using various ML methods.

3 Asymptotic theory

We first derive the asymptotically linear representation and asymptotic normality. We provide low-level conditions for various nuisance estimators in Section 3.1. We discuss uniform inference over t in Section 3.2.

Theorem 1 (Asymptotic normality) *Let Assumptions 1-3 hold. Let $h \rightarrow 0$, $nh^{d_t} \rightarrow \infty$, and $nh^{d_t+4} \rightarrow C \in [0, \infty)$. Assume that $\text{var}(Y|T = t, X = x)f_{T|X}(t|x)$ is bounded above uniformly over $x \in \mathcal{X}$. Then for any t in the interior of \mathcal{T} ,*

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

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

Note that the second part in the influence function⁹ 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 \mathbf{V}_t . We keep these smaller-order terms to show that the nuisance estimators have no first-order influence on the asymptotic distribution of $\hat{\beta}_t$. 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).

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, so we may estimate the leading bias $h^2 \mathbf{B}_t$ by the sample analogue $h^2 n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} K_h(T_i - t)(Y_i - \hat{\gamma}_\ell(t, X_i))/\hat{f}_\ell(t|X_i)$. We can estimate the asymptotic variance \mathbf{V}_t by the sample variance of the estimated influence function

⁹For our non-regular parameters, we borrow the terminology “influence function” in estimating a regular parameter that is \sqrt{n} -estimable. An influence function gives the first-order asymptotic effect of a single observation on the estimator. The estimator is asymptotically equivalent to a sample average of the influence function. See Hampel (1974) and Ichimura and Newey (2017), for example.

$\hat{V}_t = h^{d_t} n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} \hat{\psi}_{li}^2$, where $\hat{\psi}_{li} = K_h(T_i - t)(Y_i - \hat{\gamma}_\ell(t, X_i)) / \hat{f}_\ell(t|X_i) + \hat{\gamma}_\ell(t, X_i) - \hat{\beta}_t$. 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 \mathbf{B}_t is non-zero, then the bandwidth that minimizes the asymptotic mean squared error is $h_t^* = (d_t \mathbf{V}_t / (4\mathbf{B}_t^2))^{1/(d_t+4)} n^{-1/(d_t+4)}$.*
- (ii) *Consider an integrable weight function $w(t) : \mathcal{T} \mapsto \mathcal{R}$. The bandwidth that minimizes the asymptotic integrated MSE $\int_{\mathcal{T}} (\mathbf{V}_t / (nh) + h^4 \mathbf{B}_t^2) w(t) dt$ is $h_w^* = (d_t \mathbf{V}_w / (4\mathbf{B}_w))^{1/(d_t+4)} n^{-1/(d_t+4)}$, where $\mathbf{V}_w \equiv \int_{\mathcal{T}} \mathbf{V}_t w(t) dt$ and $\mathbf{B}_w \equiv \int_{\mathcal{T}} \mathbf{B}_t^2 w(t) dt$.*

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., $h^2 \sqrt{nh^{d_t}} \rightarrow 0$. Then we can construct the usual $(1 - \alpha) \times 100\%$ point-wise confidence interval $[\hat{\beta}_t \pm \Phi^{-1}(1 - \alpha/2) \sqrt{\hat{V}_t / (nh^{d_t})}]$, where Φ is the CDF of $\mathcal{N}(0, 1)$. 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 left for future research.

Next we present the asymptotic theory for $\hat{\theta}_t$. We consider two conditions for the tuning parameter η via $\eta/h \rightarrow \rho$ 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 and $\partial_t \equiv \partial_t^1$.

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

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

$$\sqrt{nh^{d_t+2}}(\hat{\theta}_t - \theta_t) = \sqrt{\frac{h^{d_t+2}}{n}} \sum_{i=1}^n \frac{\partial}{\partial t_1} K_h(T_i - t) \frac{Y_i - \gamma(t, X_i)}{f_{T|X}(t|X_i)} + o_p(1) \quad (6)$$

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

(ii) Let $\eta/h \rightarrow \rho \in (0, \infty]$, $nh^{d_t}\eta^2 \rightarrow \infty$, and $nh^{d_t}\eta^4 \rightarrow 0$. Then for any $t \in \mathcal{T}$, $\sqrt{nh^{d_t}\eta^2}(\hat{\theta}_t - \theta_t - h^2\mathbf{B}_t^\theta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_t^\theta)$, where $\mathbf{V}_t^\theta = 2\mathbb{E}[\text{var}(Y|T=t, X)/f_{T|X}(t|X)](\int_{-\infty}^{\infty} k(u)^2 du - \bar{k}(\rho))$ with the convolution kernel $\bar{k}(\rho) \equiv \int_{-\infty}^{\infty} k(u)k(u-\rho)du$ and $\mathbf{B}_t^\theta \equiv \partial\mathbf{B}_t/\partial t_1$ 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 $\eta/h \rightarrow \infty$, $\bar{k}(\eta/h) = 0$ and hence $\mathbf{V}_t^\theta = 2\mathbf{V}_t$. This is in line with the special case of a fixed η implied by the result in Theorem 1.

3.1 Nuisance estimators

In this section, we illustrate that the high-level conditions on the convergence rates in Assumption 3 are attainable by the nonparametric and ML methods: kernel, series, the generalized random forests in Athey, Tibshirani, and Wager (2019), and the deep neural networks in Farrell, Liang, and Misra (2021b). Lasso methods have been extensively studied in Su, Ura, and Zhang (2019), Sasaki and Ura (2021), and Sasaki, Ura, and Zhang (2021). These ML methods require different low-level conditions and have their own (dis)advantages depending on the settings and data generating processes.

We seek theoretical results to provide insights on choosing the tuning parameters in practice, which is challenging and under-developed in the ML literature. We can use the optimal choices for the nuisance estimators, as they do not affect the first-order asymptotics. A common method for selection of the tuning parameters is cross-validation.

To simplify notations, let the root-mean-squared norm, or the $L_2(TX)$ norm, of a random vector (T, X) with distribution F_{TX} be $\|\hat{\gamma} - \gamma\|_{F_{TX}} \equiv (\int_{\mathcal{T} \times \mathcal{X}} (\hat{\gamma}(t, x) - \gamma(t, x))^2 f_{TX}(t, x) dx dt)^{1/2}$. We define the *partial $L_2(tX)$ norm* for any $t \in \mathcal{T}$ as $\|\hat{\gamma} - \gamma\|_{F_{tX}} \equiv \|\hat{\gamma}(T, X) - \gamma(T, X)\|_{F_{tX}} \equiv (\int_{\mathcal{X}} (\hat{\gamma}(t, x) - \gamma(t, x))^2 f_{TX}(t, x) dx)^{1/2}$, where the joint distribution $F_{TX}(t, X)$ is evaluated at a fixed value of T equal to t .

Assumption 3(ii) requires $\sqrt{nh^{d_t}}\|\hat{f} - f_{T|X}\|_{F_{tX}}\|\hat{\gamma} - \gamma\|_{F_{tX}} \xrightarrow{p} 0$. Consider the conditional density estimator \hat{f} given in Section 2.1 that uses a conditional mean estimator with a $L_2(X)$ convergence rate $\|\hat{\mathbb{E}}[W|X] - \mathbb{E}[W|X]\|_{F_X} = (\int_{\mathcal{X}} (\hat{\mathbb{E}}[W|X=x] - \mathbb{E}[W|X=x])^2 f_X(x) dx)^{1/2} = O_p(R_1)$ for a bounded random variable W . Then by Lemma 1, Assumption 3(ii) requires

$$\sqrt{nh^{d_t}}\|\hat{\gamma} - \gamma\|_{F_{tX}} \left(h_1^{-d_t} \left\| \hat{\mathbb{E}}[W|X] - \mathbb{E}[W|X] \right\|_{F_X} + h_1^2 \right) \xrightarrow{p} 0. \quad (7)$$

Therefore, we need to obtain the partial $L_2(tX)$ convergence rate $\|\hat{\gamma} - \gamma\|_{F_{tX}}$ and the standard $L_2(X)$ convergence rate $\|\hat{\mathbb{E}}[W|X] - \mathbb{E}[W|X]\|_{F_X}$.

The condition in (7) provides insights on selection of the tuning parameters. We could choose the optimal bandwidths for $\hat{\gamma}$ and $\hat{\mathbb{E}}[W|X]$ that respectively minimize $\|\hat{\gamma} - \gamma\|_{F_{TX}}$ and $\|\hat{\mathbb{E}}[W|X] - \mathbb{E}[W|X]\|_{F_X}$, which might be available in the literature. Similarly we can derive the optimal $h_1^* \propto R_1^{1/(2+d_t)}$.

Notice that cross-validation is an approximately unbiased estimator of the integrated MSE, which is the expected integrated squared error, or $\mathbb{E}[\|\hat{\gamma} - \gamma\|_{F_{TX}}]$; see, e.g., Theorem 19.7 in Hansen (2021). As we show in the following sections that some estimators have the same partial $L_2(tX)$ convergence rate as the standard $L_2(TX)$ convergence rate, it is reasonable to use cross-validation for the nuisance estimators as a rule-of-thumb method to choose the tuning parameters in practice.

We derive the $L_2(tX)$ convergence rate for series in Section 3.1.2. In Section 3.1.4, we propose a deep MLP-ReLU network kernel estimator for $\gamma(t, x)$ and derive its $L_2(tX)$ convergence rate. These results are new and non-trivial extensions of existing results in the literature.

The partial $L_2(tX)$ convergence rate may seem non-standard. We provide Result 1 below to obtain the $L_2(tX)$ convergence rate when more common convergence rates, such as point-wise convergence rate, of a ML method are available. These results for the L_2 convergence rates are also useful for the semiparametric cases in CCDDHNR. We will use Result 1(i) for a kernel-based estimator in Section 3.1.1 and Result 1(ii) for the generalized random forests in Athey, Tibshirani, and Wager (2019) in Section 3.1.3.

Result 1 *For any $(t, x) \in \mathcal{T} \times \mathcal{X}$, suppose either*

(i) the mean-squared error $MSE(\hat{\gamma}(t, x)) = \mathbb{E}[(\hat{\gamma}(t, x) - \gamma(t, x))^2] = O(a_n^2)$ and is bounded a.e.,¹⁰ or

(ii) $|\hat{\gamma}(t, x) - \gamma(t, x)| = O_p(a_n)$ for $\hat{\gamma}(t, x)$ and $\gamma(t, x)$ that are uniformly bounded over $\mathcal{T} \times \mathcal{X}$.¹¹

Then $\|\hat{\gamma} - \gamma\|_{F_{TX}} = O_p(a_n)$ and $\|\hat{\gamma} - \gamma\|_{F_{tX}} = O_p(a_n)$ for any $t \in \mathcal{T}$.

In the following sections, we focus on conditional mean estimators for γ . For the conditional density estimator for $f_{T|X}$ described in Section 2.1, we use the results of the $L_2(X)$ convergence rate. The corresponding estimators using the following methods are constructed by replacing the dependent variable Y with $g_{h_1}(T - t)$ and replacing the regressors (T, X) with X .

¹⁰Note that $\mathbb{E}[\|\hat{\gamma} - \gamma\|_{F_{TX}}^2] = \int_{\mathcal{T} \times \mathcal{X}} \mathbb{E}[(\hat{\gamma}(t, x) - \gamma(t, x))^2] dF_{TX}(t, x)$ is an integrated MSE. When $MSE(\hat{\gamma}(t, x))$ is bounded a.e., $\mathbb{E}[\|\hat{\gamma} - \gamma\|_{F_{TX}}^2] = O(a_n^2)$. Then Chebyshev's inequality implies $\|\hat{\gamma} - \gamma\|_{F_{TX}} = O_p(a_n)$. The same argument applies to show $\mathbb{E}[\|\hat{\gamma} - \gamma\|_{F_{tX}}^2] = \int_{\mathcal{X}} \mathbb{E}[(\hat{\gamma}(t, x) - \gamma(t, x))^2] f_{TX}(t, x) dx = O(a_n^2)$.

¹¹Since $\hat{\gamma} - \gamma$ is bounded, there exists some constant M such that $|\hat{\gamma}(t, x) - \gamma(t, x)|/a_n < M$, for all n and $(t, x) \in \mathcal{T} \times \mathcal{X}$, with probability approach one (w.p.a.1). It follows that $\int_{\mathcal{T} \times \mathcal{X}} (\hat{\gamma}(t, x) - \gamma(t, x))^2 f_{TX}(t, x) dt dx \leq \int_{\mathcal{T} \times \mathcal{X}} M^2 a_n^2 f_{TX}(t, x) dt dx = M^2 a_n^2$, w.p.a.1. Similarly for the $L_2(tX)$ convergence rate, for any $t \in \mathcal{T}$, $\int_{\mathcal{X}} (\hat{\gamma}(t, x) - \gamma(t, x))^2 f_{TX}(t, x) dx \leq \int_{\mathcal{X}} M^2 a_n^2 f_{TX}(t, x) dx = M^2 a_n^2$, w.p.a.1.

3.1.1 Kernel

The mean-squared error of a kernel-based estimator is well-studied (see, e.g., Chapter 19 in Hansen (2021)), so we can use Result 1(i) to verify Assumption 3. Consider the Nadaraya-Watson regression estimator $\hat{\gamma}(t, x) = \sum_{i=1}^n Y_i K_{h_\gamma}(T_i - t) K_{h_\gamma}(X_i - x) / \sum_{i=1}^n K_{h_\gamma}(T_i - t) K_{h_\gamma}(X_i - x)$ with a bandwidth h_γ and a kernel of order r_γ . The mean-squared error is $O(a_{n\gamma}^2)$ for any $(t, x) \in \mathcal{T} \times \mathcal{X}$, where $a_{n\gamma} = (nh_\gamma^d)^{-1/2} + h_\gamma^{r_\gamma}$ and $d = d_t + d_x$.

Estimate the GPS $f_{T|X}$ by the standard kernel estimator $\hat{f}_{T|X}(t|x) = \sum_{i=1}^n K_{h_f}(T_i - t) K_{h_f}(X_i - x) / \sum_{i=1}^n K_{h_f}(X_i - x)$ with a bandwidth h_f and a kernel of order r_f . The mean-squared error of $\hat{f}_{T|X}(t|x)$ is $O(a_{nf}^2)$ for any $(t, x) \in \mathcal{T} \times \mathcal{X}$, where $a_{nf} = (nh_f^d)^{-1/2} + h_f^{r_f}$.¹² Assumption 4 ensures that the MSEs of $\hat{\gamma}$ and $\hat{f}_{T|X}$ are bounded *a.e.*

Assumption 4 (First-step kernel) $f_{TX}(t, x)$ is bounded away from zero. $\text{var}(Y|T = t, X = x)$ is bounded and continuous. the second derivatives of $\gamma(t, x)$ and $f_{TX}(t, x)$ are bounded and continuous *a.e.*

By Assumption 4 and Result 1(i), $\|\hat{\gamma} - \gamma\|_{F_{tX}} = \|\hat{\gamma} - \gamma\|_{F_{TX}} = O_p(a_{n\gamma})$ and $\|\hat{f}_{T|X} - f_{T|X}\|_{F_{tX}} = O_p(a_{nf})$. Then Assumption 3(ii) requires $\sqrt{nh^{d_t}}((nh_f^d)^{-1/2} + h_f^{r_f})((nh_\gamma^d)^{-1/2} + h_\gamma^{r_\gamma}) \rightarrow 0$. It further implies that we can choose the AMSE optimal bandwidths of $\hat{\gamma}$ and $\hat{f}_{T|X}$, respectively, i.e., $h_\gamma \propto n^{-1/(2r_\gamma+d)}$ and $h_f \propto n^{-1/(2r_f+d)}$.

3.1.2 Series

We illustrate how a series estimator satisfies Assumption 3 using the results in Newey (1997) summarized in Chapter 20 in Hansen (2021). For a detailed review of series methods, see Chen (2007).

Let $Z_K \equiv Z_K(T, X)$ be a $K \times 1$ vector of regressors obtained by making transformations of (T, X) , such as polynomial. The series approximation to $\gamma(t, x)$ is $\gamma_K(t, x) \equiv Z_K(t, x)' \beta_K$, where $\beta_K \equiv \mathbb{E}[Z_K Z_K']^{-1} \mathbb{E}[Z_K Y]$. Consider a least squares estimator $\hat{\beta}_K \equiv (\sum_{i=1}^n Z_{Ki} Z_{Ki}')^{-1} \sum_{i=1}^n Z_{Ki} Y_i$ and $\hat{\gamma}_K(t, x) \equiv Z_K(t, x)' \hat{\beta}_K$.

To analyze the asymptotic properties, define $\mathbf{Q}_K \equiv \int_{\mathcal{T} \times \mathcal{X}} Z_K(t, x) Z_K(t, x)' dF_{TX}(t, x)$, $\zeta_K \equiv \sup_{(t, x)} (Z_K(t, x)' \mathbf{Q}_K^{-1} Z_K(t, x))^{1/2}$, and the projection approximation error $r_K(t, x) \equiv \gamma(t, x) - Z_K(t, x)' \beta_K$.

¹²Consider the GPS estimator proposed in Section 2.1, i.e., $\hat{f} = \hat{f}_{T|X}(t|x) = \sum_{i=1}^n g_{h_1}(T_i - t) K_{h_\gamma}(X_i - x) / \sum_{i=1}^n K_{h_\gamma}(X_i - x)$. Lemma 1 implies $\|\hat{f} - f_{T|X}\|_{F_{tX}} = O_p(h_1^2 + h_1^{-d_t}((nh_f^d)^{-1/2} + h_f^{r_f}))$, which is a larger bound than a_n with $h_1 = h_f$ and the same kernel $g = K$. This shows that the bound provided in Lemma 1 is not sharp.

Assumption 5 (Series) (i) The smallest eigenvalue of \mathbf{Q}_K is bounded away from zero.

(ii) $\zeta_K^2 \log(K)/n \rightarrow 0$ as $n, K \rightarrow \infty$.

(iii) There are α and β_K such that $\sup_{(t,x) \in \mathcal{T} \times \mathcal{X}} |r_K(t, x)| = O(K^{-\alpha})$ as $K \rightarrow \infty$.

(iv) $\text{var}(Y|T = t, X = x)$ and $f_{T|X}(t, x)$ is bounded above uniformly over $\mathcal{T} \times \mathcal{X}$.

Theorem 3 (Series) Let Assumption 5 hold. Then $\|\hat{\gamma}_K - \gamma\|_{F_{tX}} = O_p(\sqrt{K/n} + K^{-\alpha})$ for $t \in \mathcal{T}$.

Theorem 20.7 in Hansen (2021) provides the convergence rate of an integrated squared error which is defined as $\|\hat{\gamma}_K - \gamma\|_{F_{TX}}^2$, i.e., under Assumption 5, the $L_2(TX)$ convergence rate $\|\hat{\gamma}_K - \gamma\|_{F_{TX}} = O_p(\sqrt{K/n} + K^{-\alpha})$. Theorem 3 contributes the convergence rate for the partial $L_2(tX)$ norm for a series estimator that is the same as the standard $L_2(TX)$ convergence rate.

To verify Assumption 5(ii), $\zeta_K \leq O(K)$ for power series and $\zeta_K \leq O(K^{1/2})$ for splines under the assumption that $f_{TX}(t, x)$ is strictly positive on $\mathcal{T} \times \mathcal{X}$ (Theorem 20.3 in Hansen (2021)). Assumption 5(iii) is from Assumption 3 in Newey (1997) and is satisfied for splines and power series by $\alpha = s/d$, where s is the number of continuous derivatives of $\gamma(t, x)$. It implies that $\|r_K\|_{F_{TX}} = O(K^{-\alpha})$ and $\|r_K\|_{F_{tX}} = O(K^{-\alpha})$. Note that Assumption 5(iii) is sufficient to obtain $\|r_K\|_{F_{tX}}$ for the partial $L_2(tX)$ norm, but it may be stronger than necessary; see Theorem 20.2 in Hansen (2021) for the $L_2(TX)$ norm $\|r_K\|_{F_{TX}}$.

3.1.3 Generalized random forests

Theoretical properties of random forests have not been fully understood; see, e.g., Biau and Scornet (2016), Athey, Tibshirani, and Wager (2019), and references therein. Choosing tuning parameters justified by theory is challenging. In practice, it is recommended either cross-validating or using the defaults in the software. We show that in theory, Assumption 3 is attainable by the generalized random forests in Athey, Tibshirani, and Wager (2019).

We use Result 1(ii) that the pointwise convergence rate of a uniformly bounded estimator $\hat{\gamma}(t, x)$ of a uniformly bounded function $\gamma(t, x)$ implies the $L_2(TX)$ convergence rate and the $L_2(tX)$ convergence rate. Theorem 5 in Athey, Tibshirani, and Wager (2019) implies that $|\hat{\gamma}(t, x) - \gamma(t, x)| = O_p(n^{-(1-\beta)/2})$ for any $(t, x) \in \mathcal{T} \times \mathcal{X}$. The convergence rate depends on β that corresponds to the subsample size $s = n^\beta$ and satisfies $\beta_{\min} < \beta < 1$, where $\beta_{\min} \equiv 1 - (1 + \pi^{-1}(\log(\omega^{-1})))/(\log((1 - \omega)^{-1}))^{-1}$, π is the lower bound of the probability that the tree splits on each variable, and ω is the minimum fraction of the observations in the parent node into

each child in every split.¹³ Assumption 6 imposes further regularization conditions from Athey, Tibshirani, and Wager (2019).

Assumption 6 (Generalized random forests) (i) $\mathcal{T} \times \mathcal{X} = [0, 1]^{d_t + d_x}$ and \mathcal{Y} is compact. (ii) For $(t, x) \in \mathcal{T} \times \mathcal{X}$, $\text{var}(Y|T = t, X = x) > 0$ and $\gamma(t, x)$ is Lipschitz in (t, x) . (iii) For $(t, x) \in \mathcal{T} \times \mathcal{X}$, $\text{var}(T|X = x) > 0$ and $f_{T|X}(t|x)$ is Lipschitz in x .

When \mathcal{Y} is compact, the generalized random forest estimator $\hat{\gamma}(t, x)$ and $\gamma(t, x)$ are bounded uniformly over $\mathcal{T} \times \mathcal{X}$. It then implies that under our setup and Assumption 6(i) and (ii), the $L_2(TX)$ and $L_2(tX)$ convergence rates are $n^{(1-\beta)/2}$.

The estimator for the conditional density $f_{T|X}$ proposed in Section 2.1 is constructed by replacing the dependent variable Y with $g_{h_1}(T - t)$ and replacing the regressors (T, X) with X . Further let Assumption 6(iii) hold.

To see Assumption 3 is attainable, let $h = Cn^{-a}$ and $h_1 = C_1n^{-b}$ for some positive constants C, C_1 . Assumption 3(ii) requires $\sqrt{nh^{d_t}}n^{-(1-\beta)/2}(n^{-(1-\beta)/2}h_1^{-d_t} + h_1^2) \rightarrow 0$ that implies $(1 - ad_t)/2 - (1 - \beta) + bd_t < 0$ and $(1 - ad_t)/2 - (1 - \beta)/2 - 2b < 0$. For one continuous treatment variable $d_t = 1$, the conditions are $(-a + \beta)/4 < b < (1 + a)/2 - \beta$ that imply $a > (5\beta - 2)/3$.

To see these conditions can be satisfied, consider an example of $\pi = 0.5$ and $\omega = 0.05$. Then $\beta_{\min} = 0.9915$. It follows that $a > 0.987$, which is larger than 0.2 for the AMSE optimal bandwidth h_t^* given in Corollary 1, resulting in undersmoothing.

We remark that for the semiparametric cases in CCDDHNR, the parallel condition of Assumption 3(ii) is $\sqrt{nn}^{-1+\beta} \rightarrow 0$ that implies $\beta < 0.5$. For our continuous treatment case, $\beta < (2 + 3a)/5$. Since $a > 1/5$, our nonparametric case of continuous treatments requires a less restrictive β .

3.1.4 Deep neural networks

We consider the deep neural networks in Farrell, Liang, and Misra (2021b) (FLM, hereafter) that use the fully connected feedforward neural networks (multilayer perceptron, or MLP) and the nonsmooth rectified linear units (ReLU) activation function. We propose a deep MLP-ReLU network kernel estimator for $\gamma(t, x)$. The proposed estimator serves the purpose to conveniently apply the $L_2(TX)$ convergence rate given in FLM to obtain the $L_2(tX)$ convergence rate. So we can deliver valid asymptotic inference for β_t and θ_t following deep learning. In this section, we closely follow the notations in FLM for easy reference, by slightly abusing our notations.

¹³Theorem 5 in Athey et al. (2019) shows that $(\hat{\gamma}(t, x) - \gamma(t, x))/\sigma_n(t, x) \xrightarrow{d} \mathcal{N}(0, 1)$, where the sequence $\sigma_n^2(t, x) = \text{polylog}(n/s)^{-1}s/n$. Since $\text{polylog}(n/s)$ is bounded away from zero and increases at most polynomially with $\log(n/s)$, $\text{polylog}(n/s)^{-1}$ is bounded above. So we obtain a loose bound $\hat{\gamma}(t, x) - \gamma(t, x) = O_p(\sigma_n(t, x)) = O_p(\sqrt{s/n}) = O(n^{-(1-\beta)/2})$.

We consider a kernel-weighted loss function for any $t \in \mathcal{T}$,

$$\ell_{tb}(f, Z) \equiv \frac{1}{2} (Y - f(X))^2 \mathbf{K}_b(T - t),$$

where a product kernel $\mathbf{K}_b(T - t) \equiv \prod_{j=1}^{d_t} \mathbf{k}((T_j - t_j)/b)/b^{d_t}$ with the kernel function $\mathbf{k}()$ satisfying Assumption 2 and a positive sequence of bandwidth b . We define the *deep MLP-ReLU network kernel estimator* for any $t \in \mathcal{T}$ as

$$\hat{f}_{tb}(X_i) \equiv \arg \min_{f_\theta \in \mathcal{F}_{MLP}, \|f_\theta\|_\infty \leq 2M} \sum_{i=1}^n \ell_{tb}(f_\theta, Z_i), \quad (8)$$

where \mathcal{F}_{MLP} is the MLP class, M is an absolute constant, and θ collects the weights and constants over all nodes. We refer the details of the MLP-ReLU network estimators to FLM. Then we obtain $\hat{\gamma}(t, x) = \hat{f}_{tb}(x)$.

Assumption 7(i) is due to the kernel function \mathbf{k} in the loss function. Assumption 7(ii)-(iv) collect assumptions from FLM. To simplify exposition, X is assumed to be continuous, but discrete variables in X are allowed. Theorem 1 in Farrell, Liang, and Misra (2021a) find that the rate only depends on the dimension of the continuously distributed components.

Assumption 7 (DNN) For any $t \in \mathcal{T}$,

- (i) The second derivatives of $f_{T|X}(t, x)$ and $\gamma(t, x)$ with respect to t are bounded and continuous.
- (ii) X are continuously distributed with support $\mathcal{X} = [-1, 1]^{d_x}$. For an absolute constant $M > 0$, $\|\gamma(t, \cdot)\|_\infty \leq M$ and $\mathcal{Y} \subset [-M, M]$.
- (iii) $\gamma(t, \cdot)$ lies in the Hölder ball $\mathcal{W}^{\beta, \infty}([-1, 1]^{d_x})$, with smoothness $\beta \in \mathcal{N}_+$ and $\mathcal{W}^{\beta, \infty}([-1, 1]^{d_x}) \equiv \{f : \max_{\alpha, |\alpha| \leq \beta} \text{ess sup}_{x \in [-1, 1]^{d_x}} |D^\alpha f(x)| \leq 1\}$, where $\alpha = (\alpha_1, \dots, \alpha_{d_x})$, $|\alpha| = \alpha_1 + \dots + \alpha_{d_x}$ and $D^\alpha f$ is the weak derivative.
- (iv) $\|f_{T|X}(t|\cdot)\|_\infty \leq M$.

Theorem 4 (DNN) Let Assumption 7 hold. Let \hat{f}_{tb} be the deep MLP-ReLU network kernel estimator defined by (8).

- (i) Let width $H \asymp (nb^{2d_t})^{\frac{d_x}{2(\beta+d_x)}} \log^2(nb^{2d_t})$ and depth $L \asymp \log(nb^{2d_t})$. Then $\|\hat{f}_{tb} - \gamma\|_{F_{tX}}^2 \leq C \cdot \left\{ (nb^{2d_t})^{-\frac{\beta}{\beta+d_x}} \log^8 n + \log \log n / (nb^{d_t}) + b^2 \right\}$ with probability approaching one as $n \rightarrow \infty$, for a constant $C > 0$ independent of n , which may depend on d , M , and other fixed constants.

(ii) For \hat{f}_{tb} with a uniform kernel, width $H \asymp (nb^{d_t})^{\frac{d_x}{2(\beta+d_x)}} \log^2(nb^{d_t})$ and depth $L \asymp \log(nb^{d_t})$, $\|\hat{f}_{tb} - \gamma\|_{F_{tX}}^2 = O_p\left((nb^{d_t})^{-\frac{\beta}{\beta+d_x}} \log^8 n + \log \log n / (nb^{d_t}) + b^4\right)$, as $nb^{d_t} \rightarrow \infty$.¹⁴

To estimate the conditional density $f_{T|X}$, we use the estimator proposed in Section 2.1, where the conditional mean estimator $\hat{\mathbb{E}}[g((T-t)/h_1)|X=x] = \hat{f}_{MLP}(x)$ is the MLP estimator in FLM who use the unweighted loss function $\ell(f, Z) = (g((T-t)/h_1) - f)^2$. Assuming $f_{T|X}(t|\cdot) \in \mathcal{W}^{\beta,\infty}([-1,1]^{d_x})$, Theorem 1 in FLM provides the $L_2(X)$ convergence rate $\|\hat{f}_{MLP}(X) - \mathbb{E}[g((T-t)/h_1)|X]\|_{F_X}^2 = O_p\left(n^{-\frac{\beta}{\beta+d_x}} \log^8 n + \log \log n/n\right)$ for any t .

We are ready to show that Assumption 3 is attainable by the MLP-ReLU network estimators. Assumption 3(ii) for $d_t = 1$ is $nh\left((nb^2)^{-\frac{\beta}{\beta+d_x}} \log^8 n + \log \log n/(nb) + b^2\right)\left(h_1^{-2}\left(n^{-\frac{\beta}{\beta+d_x}} \log^8 n + \log \log n/n\right) + h_1^4\right) \rightarrow 0$. When $h = h_1 = b$, letting smoothness $\beta > 1.6d_x$ is sufficient for the above inequality to hold. FLM discuss the condition $\beta > d_x$ for the average treatment effect of a binary treatment variable. Similarly we note that this condition is not minimal but to justify the practical use of the MLP-ReLU network estimators for valid inference on the average structural function and the partial effect of continuous treatments by our approach.

Remark 2 To estimate $\gamma(T, X)$, we could use the unweighted loss function for least squares $\ell(f, Z) = (Y - f)^2/2$ to obtain the MLP estimator $\hat{f}_{MLP}(T, X)$ in FLM. FLM provide the corresponding $L_2(TX)$ convergence rate $\|\hat{f}_{MLP} - \gamma\|_{F_{TX}}^2 = O_p\left(n^{-\frac{\beta}{\beta+d_x+d_t}} \log^8 n + \log \log n/n\right)$ and illustrate its usefulness for semiparametric inference on the average treatment effect of a binary treatment. But we find it challenging to obtain the $L_2(tX)$ convergence rate $\|\hat{f}_{MLP} - \gamma\|_{F_{tX}}$ for our case of continuous treatments by extending FLM's results. So we use a different loss function $\ell_{tb}(f, Z)$; details are in the proof of Theorem 4 in the Appendix.

The partial L_2 convergence rate also appears in Noack, Olma, and Rothe (2021) who study covariate adjustments in regression discontinuity designs. The following corollary provides the partial L_2 convergence rate for t at the boundary that can be applied to the policy threshold in regression discontinuity designs.

Corollary 2 *Let t be at the boundary of \mathcal{T} . Let the conditions in Theorem 4 hold.*

(i) *Then $\|\hat{f}_{tb} - \gamma\|_{F_{tX}}^2 \leq C \cdot \left\{ (nb^{2d_t})^{-\frac{\beta}{\beta+d_x}} \log^8 n + \log \log n / (nb^{d_t}) + b \right\}$ with probability approaching one as $n \rightarrow \infty$, for a constant $C > 0$ independent of n , which may depend on d , M , and other fixed constants.*

¹⁴We thank Christoph Rothe for the insight on the effective sample size nb^{d_t} .

- (ii) For \hat{f}_{tb} with a uniform kernel, width $H \asymp (nb^{d_t})^{\frac{d_x}{2(\beta+d_x)}} \log^2(nb^{d_t})$ and depth $L \asymp \log(nb^{d_t})$, $\|\hat{f}_{tb} - \gamma\|_{F_{tX}}^2 = O_p\left((nb^{d_t})^{-\frac{\beta}{\beta+d_x}} \log^8 n + \log \log n / (nb^{d_t}) + b^2\right)$, as $nb^{d_t} \rightarrow \infty$.

3.2 Uniform inference

We extend the asymptotic theory to uniformity over $t \in \mathcal{T}_0$ which is a compact subset of \mathcal{T} . The uniform asymptotic representation in Theorem 5 is the basis for a uniform inference procedure for β_t and θ_t . Assumption 8 strengthens Assumption 3 for the nuisance estimators.

Assumption 8 For each $\ell = 1, \dots, L$,

- (i) $\sup_{t \in \mathcal{T}_0} \sqrt{nh^{d_t}} \left(\int_{\mathcal{X}} (\hat{f}_\ell(t|x) - f_{T|X}(t|x))^2 f_{TX}(t, x) dx \right)^{1/2} \left(\int_{\mathcal{X}} (\hat{\gamma}_\ell(t|x) - \gamma(t, x))^2 f_{TX}(t, x) dx \right)^{1/2} \xrightarrow{p} 0$.
- (ii) There exist positive sequences $A_{1n} \rightarrow 0$ and $A_{2n} \rightarrow 0$ such that $\sup_{(t,x) \in \mathcal{T}_0 \times \mathcal{X}} |\hat{\gamma}_\ell(t, x) - \gamma(t, x)| = O_p(A_{1n})$ and $\sup_{(t,x) \in \mathcal{T}_0 \times \mathcal{X}} |\hat{f}_\ell(t|x) - f_{T|X}(t|x)| = O_p(A_{2n})$.
- (iii) $\hat{\gamma}_\ell(t, x)$ and $\hat{f}_\ell(t|x)$ are Lipschitz continuous in $t \in \mathcal{T}_0$, for any $x \in \mathcal{X}$.

Assumption 8(i) is the uniform analogs of Assumption 3(ii). In addition, Assumption 8(ii) requires the nuisance estimators to converge uniformly at some rates. Fan, Hsu, Lieli, and Zhang (2021) make similar assumptions for the conditional average binary treatment effect. We may verify Assumption 8 for the kernel and series estimators by extending our results in Sections 3.1.1 and 3.1.2, respectively, and using the existing results in the literature (e.g., Newey (1994b) and Cattaneo, Farrell, and Feng (2020)). Verifying Assumption 8 for the modern ML methods, the generalized random forests in Athey, Tibshirani, and Wager (2019) and the deep neural networks in Farrell, Liang, and Misra (2021b), is more involved and beyond the scope of this paper.

Theorem 5 Let the conditions in Theorem 1 and Assumption 8 hold. Then (i) the asymptotically linear representation of $\hat{\beta}_t$ in (5) holds uniformly in $t \in \mathcal{T}_0$. (ii) Furthermore let the conditions in Theorem 2 hold. Then the asymptotically linear representation of $\hat{\theta}_t$ in (6) holds uniformly in $t \in \mathcal{T}_0$.

We consider a multiplier bootstrap method for uniform inference on β_t and θ_t over $t \in \mathcal{T}_0$. The method and proof closely follow Su, Ura, and Zhang (2019) and Theorem 4.1 in Fan, Hsu, Lieli, and Zhang (2021), where the moment functions share a similar structure with that of $\hat{\beta}_t$. Let $\{U_i\}_{i=1}^n$ be a sequence of i.i.d. random variables satisfying Assumption 9.

Assumption 9 (Multiplier bootstrap) *The random variable U_i is independent of Z_i , $\mathbb{E}[U_i] = \text{var}(U_i) = 1$, and its distribution has sub-exponential tails.*¹⁵

Assumption 9 is standard for multiplier bootstrap inference and can be satisfied by a normal random variable, for example. Then we compute

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

We use $\sqrt{nh^{d_t}}(\hat{\beta}_t^* - \hat{\beta}_t)$ to simulate the limiting process of $\sqrt{nh^{d_t}}(\hat{\beta}_t - \beta_t)$ indexed by $t \in \mathcal{T}_0$. That is, repeat the above procedure for B times and obtain a bootstrap sample of $\{\hat{\beta}_{t,b}^*\}_{b=1}^B$.

For the partial effect, compute the numerical differentiation estimator $\hat{\theta}_t^*$ following Step 3 in the estimation procedure with $\hat{\beta}_t^*$. Then we simulate the limiting process of $\sqrt{nh^{d_t+2}}(\hat{\theta}_t - \theta_t)$ by $\sqrt{nh^{d_t+2}}(\hat{\theta}_t^* - \hat{\theta}_t)$. Theorem 6 below shows the validity of the multiplier bootstrap.

Theorem 6 (Multiplier bootstrap) *Let the conditions in Theorem 5 and Assumption 9 hold. Then $\sqrt{nh^{d_t}}(\hat{\beta}_t^* - \hat{\beta}_t) = \sqrt{h^{d_t}/n} \sum_{i=1}^n U_i \{K_h(T_i - t)(Y_i - \gamma(t, X_i)/f_{T|X}(t|X_i) + \gamma(t, X_i) - \beta_t)\} + o_p(1)$ and $\sqrt{nh^{d_t+2}}(\hat{\theta}_t^* - \hat{\theta}_t) = \sqrt{h^{d_t+2}/n} \sum_{i=1}^n U_i \partial K_h(T_i - t)/\partial t_1(Y_i - \gamma(t, X_i))/f_{T|X}(t|X_i) + o_p(1)$ uniformly in $t \in \mathcal{T}_0$.*

Next we discuss inference using the multiplier bootstrap in Su, Ura, and Zhang (2019) and Fan, Hsu, Lieli, and Zhang (2021). Theorem 6 implies that $\sqrt{nh^{d_t}}(\hat{\beta}_t^* - \hat{\beta}_t)$ and $\sqrt{nh^{d_t+2}}(\hat{\theta}_t^* - \hat{\theta}_t)$ converge in distribution to the limiting distribution of $\sqrt{nh^{d_t}}(\hat{\beta}_t - \beta_t)$ and $\sqrt{nh^{d_t+2}}(\hat{\theta}_t - \theta_t)$, respectively, conditional on the sample path with probability one. Following Su, Ura, and Zhang (2019), obtain $\hat{q}_t(\alpha)$ as the α^{th} quantile of the sequence $\{\hat{\beta}_{t,b}^* - \hat{\beta}_t\}_{b=1}^B$. The standard $100(1 - \alpha)\%$ percentile bootstrap confidence interval for β_t is $(\hat{q}_t(\alpha/2) + \hat{\beta}_t, \hat{q}_t(1 - \alpha/2) + \hat{\beta}_t)$ or $(-\hat{q}_t(\alpha/2) + \hat{\beta}_t, \hat{q}_t(\alpha/2) + \hat{\beta}_t)$.

We can follow the approach in Fan, Hsu, Lieli, and Zhang (2021) to construct uniform confidence bands. Specifically obtain $\hat{Q}(\alpha)$ as the α^{th} quantile of the sequence $\{\sup_{t \in \mathcal{T}_0} \sqrt{nh^{d_t}} |\hat{\beta}_{t,b}^* - \hat{\beta}_t| / \hat{\sigma}_t\}_{b=1}^B$, where $\hat{\sigma}_t^2$ is a uniformly consistent estimator of \mathbf{V}_t . The supremum is approximated by the maximum over a chosen fine grid over \mathcal{T} . Then construct the $100(1 - \alpha)\%$ uniform confidence band as $(\hat{\beta}_t - \hat{Q}(1 - \alpha)\hat{\sigma}_t/\sqrt{nh}, \hat{\beta}_t + \hat{Q}(1 - \alpha)\hat{\sigma}_t/\sqrt{nh})$. For example, we could use $\hat{\sigma}_t^2 = \hat{\mathbf{V}}_t$ the sample variance estimator described in Section 3. Following the proof of Theorem 3.2 in Fan et al. (2021), one could show $\sup_{t \in \mathcal{T}_0} |\hat{\mathbf{V}}_t - \mathbf{V}_t| = o_p(1)$. Based on Theorem 6, the asymptotic validity of the confidence band could follow the proof of Theorem 4.2 in Fan et al. (2021). We do not include the formal theoretical details in this paper to conserve space and focus on the new results.

¹⁵A random variable U has sub-exponential tails if $P(|U| > u) \leq c_1 \exp(-c_2 u)$ for every u and some constants c_1 and c_2 ,

4 Kernel localization

We discuss the construction of the doubly robust moment function by Gateaux derivative and a local Riesz representer in Section 4.1. In Section 4.2, we discuss the adjustment for the first-step kernel estimators in the moment 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.

4.1 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 from the true distribution, 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 $h \rightarrow 0$. 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 $K_h(T - t)$. To the best of our knowledge, this is the first explicit calculation of Gateaux derivative for a non-regular nonparametric parameter. Importantly the expression in (9) below is the building block to construct estimators for β_t and the linear functionals of β_t .

More specifically, for any $t \in \mathcal{T}$, let $\beta_t(\cdot) : \mathcal{F} \rightarrow \mathcal{R}$, 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 $h \rightarrow 0$. 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

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{d}{d\tau} \beta_t(F^{\tau h}) \Big|_{\tau=0} &= \gamma(t, X) - \beta_t + \lim_{h \rightarrow 0} \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{y - \gamma(t, x)}{f_{T|X}(t|x)} f_{YTX}^h(y, t, x) dy dx \\ &= \gamma(t, X) - \beta_t + \frac{Y - \gamma(t, X)}{f_{T|X}(t|X)} \lim_{h \rightarrow 0} f_T^h(t). \end{aligned} \quad (9)$$

Note that the last term in (9) 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

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

Theorem 1 in Ichimura and Newey (2017) shows that if a semiparametric estimator is asymptotically linear and locally regular, then the influence function is $\lim_{h \rightarrow 0} d\beta_t(F^{\tau h})/d\tau|_{\tau=0}$. Here, we use the Gateaux derivative limit calculation to motivate our moment function that depends on F_T^h . Then we show that our estimator is asymptotically equivalent to a sample average of the moment function.

Remark 3 (Linear functional of β_t) Consider a non-regular function-valued linear functional of β_t , denoted by $\alpha_t = \mathbb{A}[\beta_t]$ for a linear operator \mathbb{A} . So α_t is also a nonparametric function of t . To construct the DML estimator of α_t , the Gateaux derivative of α_t is simply the linear functional of the Gateaux derivative of β_t in (9), i.e.,

$$\lim_{h \rightarrow 0} \frac{d}{d\tau} \alpha_t(F^{\tau h}) \Big|_{\tau=0} = \lim_{h \rightarrow 0} \frac{d}{d\tau} \mathbb{A}[\beta_t(F^{\tau h})] \Big|_{\tau=0} = \mathbb{A}[\gamma(t, X)] - \mathbb{A}[\beta_t] + \lim_{h \rightarrow 0} \mathbb{A} \left[\frac{Y - \gamma(t, X)}{f_{T|X}(t|X)} f_T^h(t) \right].$$

We may work out the close-form expression of this Gateaux derivative of α_t and use its estimated sample analogue to construct the DML estimator of α_t . An alternative DML estimator is simply $\hat{\alpha}_t = \mathbb{A}[\hat{\beta}_t]$.

Note that the partial effect can be expressed as a linear functional of β_t : $\theta_t = \mathbb{A}[\beta_t] = \partial\beta_t/\partial t_1$. An example of a weighted conditional average partial derivative given $T_1 = t_1$ can be defined as $\alpha_t = \int \theta_t w(t) dt_2 \dots dt_{d_t}$ that is a function of t_1 with a weight function $w(t) = w(t_1, t_2, \dots, t_{d_t})$. In contrast, the weighted average derivative $\alpha_t = \int_{\mathcal{T}} \theta_t w(t) dt = \alpha$ does not depend on t . The DML estimation for such regular real-valued parameter α has been well-studied in the semiparametric literature. We contribute to the DML literature by focusing on the function-valued non-regular nonparametric objects based on β_t .

Remark 4 (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 $\alpha_{th}(T, X) = f_T^h(t)/f_{T|X}(T|X) = K_h(T - t)/f_{T|X}(T|X)$ indexed by the evaluation value t and the bandwidth of the kernel h . Our local Riesz representer $\alpha_{th}(T, X)$ satisfies $\beta_t = \int_{\mathcal{X}} \gamma(t, X) dF_X(X) = \lim_{h \rightarrow 0} \int_{\mathcal{X}} \int_{\mathcal{T}} \alpha_{th}(T, X) \gamma(T, X) dF_{TX}(T, X)$ for all γ with finite second moment, following the in-

sight of the local 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. For a series localization, Chen, Liao, and Sun (2014) define the *sieve Riesz representer* for plug-in sieve M estimators of irregular functionals; see also Chen and Pouzo (2015) for general semi/nonparametric conditional moment models.

4.2 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 CEINR and Bravo, Escanciano, and van Keilegom (2020) for the semiparametric empirical likelihood inference in a low dimensional nonparametric setting.

For $\hat{\beta}_t^{REG}$, consider $\hat{\gamma}(t, x)$ to be a local constant or local polynomial estimator with bandwidth h for low-dimensional X . 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 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 Lettry (2020) 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 $\int_{\mathcal{Z}} K_h(T - t)Y/\hat{f}_{T|X}(t|X)dF_{YTX} = n^{-1} \sum_{i=1}^n \gamma(t, X_i) (1 - K_h(T_i - t)/f_{T|X}(t|X_i)) + o_p((nh^{d_t})^{-1/2})$ 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 $f_{T|X}$ to the original moment function $K_h(T - t)Y/f_{T|X}(t|X)$.

Remark 5 (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 $h^2\mathbf{B}_t$, $\hat{\beta}_t$ 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 $h_1^r \sqrt{nh_1^{d_t}} \rightarrow 0$. 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 $h_1^{2r} \sqrt{nh^{d_t}} \rightarrow 0$. It follows that the DML estimator need not undersmooth the nuisance estimators, while the regression estimator $\hat{\beta}_t^{REG}$ requires an undersmoothing $\hat{\gamma}$. Moreover we observe that the condition is weaker than $h_1^r \sqrt{n} \rightarrow 0$ for the binary treatment that has a regular root- n convergence rate.

Remark 6 (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.¹⁶ 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 $\int_{\mathcal{X}} \hat{\gamma}(t, x)f(x)dx$ implies the adjustment term of the form $\mathbb{E}[b(t, X)](n^{-1} \sum_{i=1}^n b(T_i, X_i)b(T_i, X_i)')^{-1} n^{-1} \sum_{i=1}^n b(T_i, X_i)'(y_i - \gamma(T_i, X_i)) = n^{-1} \sum_{i=1}^n \lambda_{ti}(y_i - \gamma(T_i, X_i))$, 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 ; for example, the approximate residual balancing estimator in Athey, Imbens, and Wager (2018), CEINR, and Demirer, Syrgkanis, Lewis, and Chernozhukov (2019).

¹⁶For 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.

5 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. To estimate the conditional mean function $\gamma(t, x) = \mathbb{E}[Y|T = t, X = x]$ and the conditional density (or the generalized propensity score GPS) $f_{T|X}$, we employ three machine learning methods: Lasso, the generalized random forests in Athey, Tibshirani, and Wager (2019), and the deep neural networks based on Farrell, Liang, and Misra (2021b) as described in Section 3.1.4. We implement our DML estimator with these three algorithms respectively in Python, using the packages scikit-learn, pytorch, numpy, pandas, rpy2 and scipy. We use the R package “grf” for the generalized random forest implementation, implementing it in Python via the rpy2 package. Software is available from the authors.

5.1 Simulation study

We describe the nuisance estimators in more detail.

Lasso: The penalization parameter is chosen via grid search utilizing tenfold cross validation for $\hat{\gamma}$ and $\hat{f}_{T|X}$ separately. The basis functions contain third-order polynomials of X and T , and interactions among X and T .

Generalized Random forest (GRF): We use the generalized random forests in Athey, Tibshirani, and Wager (2019), with 2,000 trees and all other parameters chosen via cross validation in every Monte Carlo replication. The parameters tuned via cross validation are: The fraction of data used for each tree, the number of variables tried for each split, the minimum number of observations per leaf, whether or not to use “honesty splitting,” whether or not to prune trees such that no leaves are empty, the maximum imbalance of a split, and the amount of penalty for an imbalanced split. Unlike Lasso, we do not add any additional basis functions as inputs into GRF.

Deep neural network (DNN): We use the deep neural networks described in Section 3.1.4 with four hidden layers. Each hidden layer has 10 neurons and uses rectified linear units (ReLU) activation functions. The weights are fit using stochastic gradient descent with a weight decay of 0.2 and a learning rate of 0.01.¹⁷ For the selection of the neural network models, we perform a train-test split of the data and chose the models based on out-of-sample performance.

¹⁷Weight 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.

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

$$X = (X_1, \dots, X_{100})' \sim \mathcal{N}(0, \Sigma), \quad T = \Phi(3X'\theta) + 0.75\nu - 0.5, \quad Y = 1.2T + 1.2X'\theta + T^2 + TX_1 + \varepsilon,$$

where $\theta_j = 1/j^2$, $\text{diag}(\Sigma) = 1$, the (i, j) -entry $\Sigma_{ij} = 0.5$ for $|i - j| = 1$ and $\Sigma_{ij} = 0$ for $|i - j| > 1$ for $i, j = 1, \dots, 100$, 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 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 $L \in \{1, 2, 5\}$. We use the second-order Epanechnikov kernel with bandwidth h . For the GPS estimator described in Section 2.1, we choose bandwidth $h_1 = h$. 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 computed 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 $[\hat{\beta}_t \pm 1.96s.e.]$ is asymptotically valid, where the standard error (*s.e.*) is computed using the sample analogue of the estimated influence function, as described in Section 3.

Table 1 reports the results based on 1,000 Monte Carlo replications. The estimators using these machine learning methods perform well in the case of cross-fitting ($L = 2, 5$), with coverage rates near the nominal 95% for small bandwidths. Under no cross-fitting ($L = 1$), the confidence intervals generally have lower coverage rates than under cross-fitting. The coverage rate and bias are improved the most for GRF with cross-fitting. Cross-fitting should improve our estimation in the case that the machine learning algorithm is over-fitting. Given that cross-fitting only results in small improvements for Lasso and DNN, it might suggest that those algorithms do not have a severe over-fitting problem for this data-generating process, but GRF does. It may be possible to alleviate this over-fitting via more regularization.

We do not find significant difference between $L = 2$ and $L = 5$. However one issue with choosing a smaller L is that the machine learning algorithm is fit on a smaller subset of the data. For example, we see some large RMSE for Lasso for $L = 2$ in $n = 500$. CCDDHNR also discuss that the fivefold cross-fitting estimates use more observations to learn the nuisance functions than twofold cross-fitting and thus are likely learn them more precisely.

All three methods seem somewhat robust to bandwidth choice under cross-fitting. Overall these results demonstrate consistency with the theoretical results of this paper, confirming the importance of cross-fitting for these ML methods.

Table 1: Simulation Results

n	L	c	Lasso			Random Forest			Neural Network		
			Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage
500	1	0.50	0.010	0.195	0.935	0.137	0.211	0.802	-0.083	0.286	0.964
		0.75	0.010	0.134	0.939	0.115	0.180	0.826	-0.086	0.240	0.963
		1.00	0.017	0.125	0.946	0.102	0.165	0.845	-0.088	0.217	0.962
		1.25	0.027	0.120	0.950	0.096	0.154	0.865	-0.090	0.202	0.959
		1.50	0.040	0.118	0.938	0.093	0.150	0.865	-0.091	0.192	0.947
	2	0.50	1.026	86.508	0.973	-0.011	0.208	0.947	-0.092	0.297	0.966
		0.75	-0.359	9.620	0.965	-0.005	0.170	0.953	-0.093	0.247	0.962
		1.00	-0.020	0.598	0.950	-0.002	0.151	0.953	-0.094	0.221	0.958
		1.25	-0.055	2.720	0.956	0.000	0.138	0.957	-0.095	0.206	0.958
		1.50	0.037	0.124	0.941	0.004	0.130	0.963	-0.096	0.196	0.945
	5	0.50	-0.161	5.785	0.950	-0.005	0.196	0.957	-0.089	0.292	0.964
		0.75	-0.003	0.200	0.948	0.000	0.163	0.950	-0.090	0.244	0.963
		1.00	0.013	0.130	0.944	0.004	0.145	0.957	-0.091	0.220	0.959
		1.25	0.024	0.121	0.951	0.007	0.134	0.963	-0.093	0.205	0.955
		1.50	0.038	0.119	0.943	0.012	0.127	0.962	-0.094	0.195	0.942
1000	1	0.50	0.010	0.121	0.954	0.121	0.169	0.718	-0.084	0.229	0.969
		0.75	0.014	0.106	0.949	0.105	0.149	0.752	-0.081	0.195	0.960
		1.00	0.018	0.097	0.951	0.096	0.137	0.769	-0.082	0.177	0.949
		1.25	0.025	0.093	0.935	0.092	0.130	0.779	-0.083	0.166	0.931
		1.50	0.034	0.092	0.924	0.090	0.126	0.780	-0.083	0.157	0.929
	2	0.50	-0.061	1.566	0.960	0.003	0.149	0.954	-0.091	0.235	0.969
		0.75	0.007	0.158	0.952	0.009	0.125	0.953	-0.086	0.199	0.958
		1.00	0.018	0.098	0.954	0.013	0.113	0.946	-0.086	0.180	0.945
		1.25	0.025	0.094	0.936	0.016	0.105	0.947	-0.087	0.169	0.928
		1.50	0.035	0.092	0.919	0.021	0.099	0.940	-0.086	0.160	0.921
	5	0.50	0.008	0.124	0.956	0.007	0.144	0.954	-0.088	0.232	0.965
		0.75	0.013	0.106	0.952	0.013	0.122	0.949	-0.084	0.197	0.959
		1.00	0.018	0.097	0.952	0.017	0.110	0.948	-0.084	0.179	0.946
		1.25	0.025	0.093	0.933	0.020	0.103	0.939	-0.086	0.168	0.929
		1.50	0.034	0.092	0.923	0.026	0.098	0.932	-0.085	0.158	0.928

Notes: $L = 1$: no cross-fitting. $L = 2$: twofold 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.

5.2 Empirical illustration

We illustrate our method by re-analyzing the Job Corps program in the United States, which was conducted in the mid-1990s. The Job Corps program is the largest publicly funded job training program, which 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 Lettry (2020), 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 Lettry (2020). 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. There are 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 three DML estimators Lasso, the generalized random forest, and the deep neural network. The parameters for these three methods are selected as described in Section 5.1. For the deep neural networks described in Section 3.1.4, 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 100 neurons and the second hidden layer has 20 neurons. The GPS estimation uses a network with four hidden layers and a weight decay of 0.1. Each layer has 10 neurons.

We use the second-order Epanechnikov kernel with bandwidth h . For the GPS estimator, we use the Gaussian kernel with bandwidth $h_1 = h$. We compute the optimal bandwidth h_w^* that minimizes an asymptotic integrated MSE derived in Corollary 1(ii) after an initial choice of bandwidth $3\hat{\sigma}_T n^{-0.2} = 563.34$. 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 $[\underline{t}, \bar{t}] \subset \mathcal{T}$. Set m equally spaced

grid points over $[\underline{t}, \bar{t}]$: $\{\underline{t} = t_1, t_2, \dots, t_m = \bar{t}\}$. A plug-in estimator $\hat{h}_w^* = (\hat{\mathbf{V}}_w / (4\hat{\mathbf{B}}_w))^{1/5} n^{-1/5}$, where $\hat{\mathbf{V}}_w = m^{-1} \sum_{j=1}^m \hat{\mathbf{V}}_{t_j} \mathbf{1}\{t_j \in [\underline{t}, \bar{t}]\} / (\bar{t} - \underline{t})$ and $\hat{\mathbf{B}}_w = m^{-1} \sum_{j=1}^m \hat{\mathbf{B}}_{t_j}^2 \mathbf{1}\{t_j \in [\underline{t}, \bar{t}]\} / (\bar{t} - \underline{t})$. We use $[\underline{t}, \bar{t}] = [160, 1840]$ and $t_j - t_{j-1} = 40$ in this empirical application. We then obtain bandwidths $0.8h_w^*$ for undersmoothing that are 418.87 for Lasso, 363.40 for the generalized random forest, and 318.26 for the deep neural network.

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. DNN estimates appear to be the most erratic, possibly due to the smaller bandwidth compared with other estimators.

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 around 500 (three months) and negative partial effect around 1,500 hours (9 months). Taking the estimates by Lasso for example, $\hat{\beta}_{400} = 47.03$ with standard error $s.e. = 1.31$ and $\hat{\theta}_{400} = 0.0211$ with $s.e. = 0.0062$. This estimate implies that increasing the training from two months to three months increases the average proportion of weeks employed in the second year by 3.38% (about two working weeks) with $s.e. = 1\%$.

Lee (2009) finds that the program had a negative impact on employment propensities in the short term (104 weeks since random assignments) and a positive effect in the long term (104-208 weeks). Note that Lee (2009) considers a binary treatment variable of being in the program or not, with the outcome variable $\mathbf{1}\{Y \geq 0\}$ in our notations. We focus on the employment proportion in the second year following the program assignment (52-104 weeks) and estimate the heterogeneous effects of the total hours spent in academic and vocational training in the first year.

We note that the empirical practice has focused on semiparametric estimation; see Flores, Flores-Lagunes, Gonzalez, and Neumann (2012), Hsu, Huber, Lee, and Lettry (2020), Lee (2018), for example. The semiparametric methods are subject to the risk of misspecification. Our DML estimator provides a feasible approach to implementing a fully nonparametric inference in practice.

6 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 kernel-based double debiased machine learning (DML) estimator uses a doubly robust moment function and cross-fitting. We provide tractable high-level 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. A main contribution is to provide low-level conditions for kernel and series estimators, as well as modern ML methods: the generalized random forests in Athey, Tibshirani, and Wager (2019) and the deep neural networks in Farrell, Liang, and Misra (2021b). We justify the use of the kernel function by calculating the Gateaux derivative.

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 (2021b), among others.

Another extension is robustness against multiway clustering, where the conventional cross-fitting does not ensure the independence between observations in I_ℓ from I_ℓ^c . We may adopt the K^2 -fold multiway cross-fitting proposed by Chiang, Kato, Ma, and Sasaki (2021) that focus on regular DML estimators as in CCDDHNR. Since the form of estimators and the proofs of asymptotic theories for our continuous treatment case are similar to those studied in Chiang, Kato, Ma, and Sasaki (2021), we expect that their proposed algorithm works for $\hat{\beta}_t$; a formal extension is out of the scope of this paper.

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. In particular, Lee (2009) studies the issue of sample selection for the wage effects of the Job Corps program. To extend our empirical application to the wage effect of the length of exposure to the program, we may follow Lee (2009) to estimate bounds on the wage effect of the continuous treatment using the excess number of individuals who are induced to be selected.¹⁸ A closer approach to our estimator is Das, Newey, and Vella (2003), who study a nonparametric sample selection model with endogeneity and show that the propensity score and reduced form residuals lead to a control function method to account for both selection and endogeneity. 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

¹⁸One key identification assumption in Lee (2009) is monotonicity that assumes the potential sample selection indicator to be weakly monotonic in the treatment value, i.e., $\mathbf{1}\{Y(t) > 0\} \geq \mathbf{1}\{Y(t') > 0\}$ for $t > t'$. In words, suppose an individual is employed in the second year. The monotonicity assumption requires that this individual must be employed if he/she received more hours of training. Such monotonicity assumption is not testable but may not be supported by our estimation results. Das et al. (2003) use an exclusion assumption for instrumental variables.

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

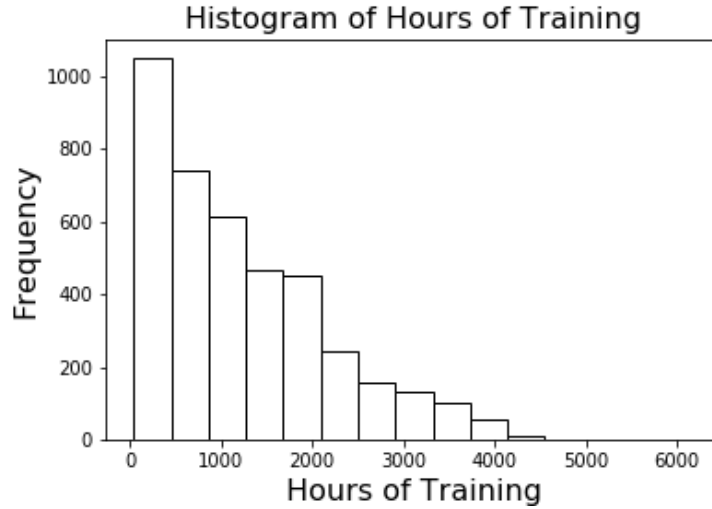


Table 2: Descriptive statistics

Variable	Mean	Median	StdDev	Min	Max
share of weeks employed in 2nd year (Y)	44.00	40.38	37.88	0	100
total hours spent in 1st-year training (T)	1219.80	992.86	961.62	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

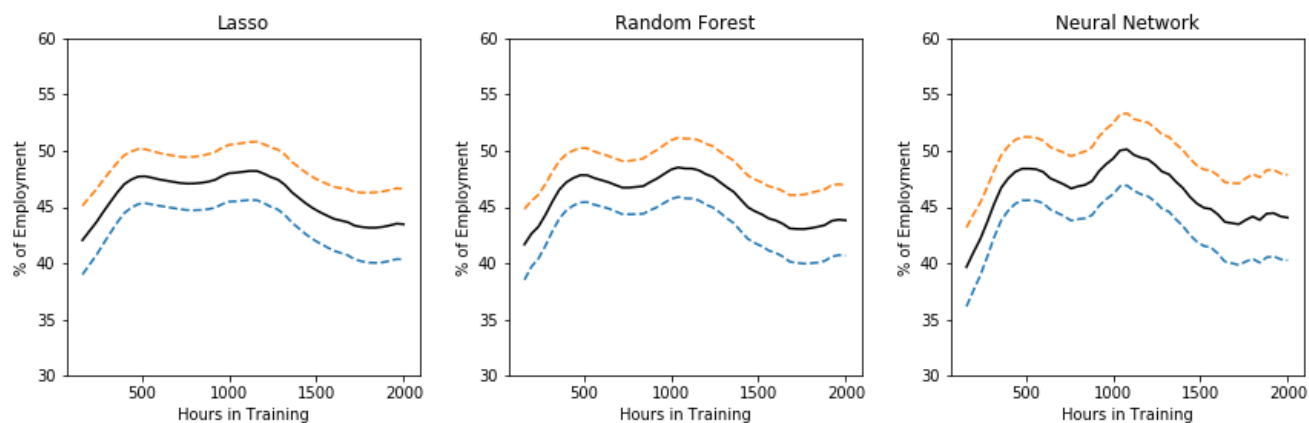
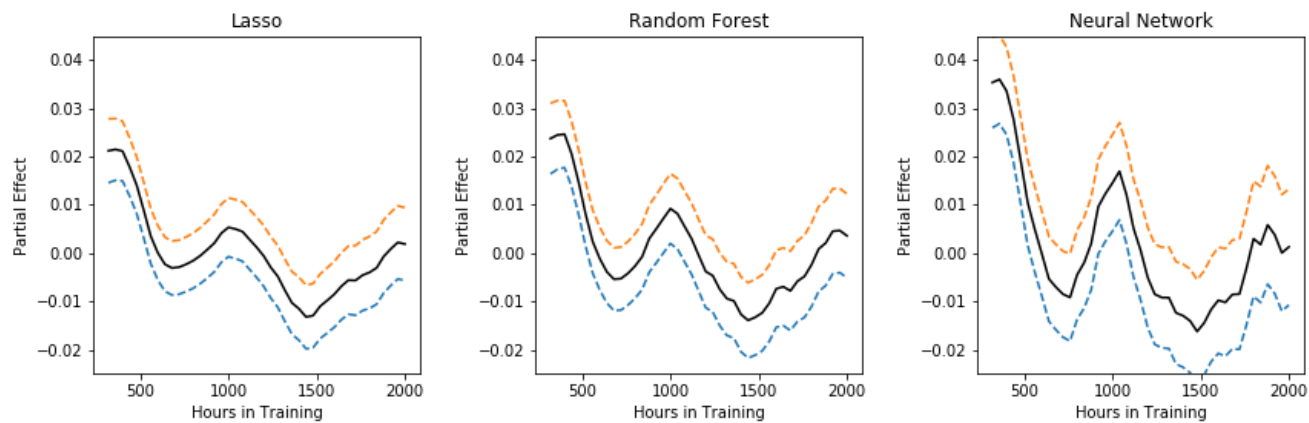


Figure 3: Estimated partial effects and 95% confidence interval



Appendix

Proof of Lemma 1 By the triangle inequality,

$$\begin{aligned}
& \left\{ \int_{\mathcal{X}} (\hat{f}_{T|X}(t|x) - f_{T|X}(t|x))^2 f_{TX}(t, x) dx \right\}^{1/2} \\
& \leq \left\{ \int_{\mathcal{X}} (\hat{f}_{T|X}(t|x) - f_{T|X}(t|x))^2 f_X(x) dx C \right\}^{1/2} \\
& \leq \left\{ \frac{C}{h_1^{2d_t}} \int_{\mathcal{X}} \left(\hat{\mathbb{E}} \left[\Pi_{j=1}^{d_t} g \left(\frac{T_j - t_j}{h_1} \right) \middle| X = x \right] - \mathbb{E} \left[\Pi_{j=1}^{d_t} g \left(\frac{T_j - t_j}{h_1} \right) \middle| X = x \right] \right)^2 f_X(x) dx \right\}^{1/2} \\
& \quad + \left\{ C \int_{\mathcal{X}} \left(\mathbb{E} [g_{h_1}(T - t) | X = x] - f_{T|X}(t|x) \right)^2 f_X(x) dx \right\}^{1/2} \\
& = O_p(R_1 h_1^{-d_t} + h_1^2). \tag{10}
\end{aligned}$$

For the second term (10) to be $O_p(h_1^2)$, we follow the standard algebra for kernel, using integration by parts and change of variables. \square

Asymptotically linear representation We give an outline of deriving the asymptotically 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 $\phi(Z_i, \beta_t, \gamma, \lambda) \equiv K_h(T_i - t)\lambda(t, X_i)(Y_i - \gamma(t, X_i))$, where $\lambda(t, x) \equiv 1/f_{T|X}(t|x)$. The doubly robust moment function $\psi(Z_i, \beta_t, \gamma, \lambda) \equiv m(Z_i, \beta_t, \gamma(t, X_i)) + \phi(Z_i, \beta_t, \gamma(t, X_i), \lambda(t, X_i))$, as in equation (1).

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

$$\begin{aligned}
& \sqrt{nh^{d_t}} \frac{1}{n} \sum_{i=1}^n \left\{ \hat{\psi}(Z_i, \beta_t, \hat{\gamma}, \hat{\lambda}) - \psi(Z_i, \beta_t, \gamma, \lambda) \right\} \\
& = \sqrt{\frac{h^{d_t}}{n}} \sum_{i=1}^n \left\{ \hat{\gamma}_i - \gamma_i - \mathbb{E}[\hat{\gamma}_i - \gamma_i] + K_h(T_i - t)\lambda_i(\gamma_i - \hat{\gamma}_i) - \mathbb{E}[K_h(T_i - t)\lambda_i(\gamma_i - \hat{\gamma}_i)] \right\} \tag{R1-1}
\end{aligned}$$

$$+ \sqrt{\frac{h^{d_t}}{n}} \sum_{i=1}^n \left\{ K_h(T_i - t)(\hat{\lambda}_i - \lambda_i)(Y_i - \gamma_i) - \mathbb{E}[K_h(T_i - t)(\hat{\lambda}_i - \lambda_i)(Y_i - \gamma_i)] \right\} \tag{R1-2}$$

$$+ \sqrt{\frac{h^{d_t}}{n}} \sum_{i=1}^n \left\{ \mathbb{E}[(\hat{\gamma}_i - \gamma_i)(1 - K_h(T_i - t)\lambda_i)] + \mathbb{E}[(\hat{\lambda}_i - \lambda_i)K_h(T_i - t)(Y_i - \gamma_i)] \right\} \tag{R1-DR}$$

$$- \sqrt{\frac{h^{d_t}}{n}} \sum_{i=1}^n K_h(T_i - t)(\hat{\lambda}_i - \lambda_i)(\hat{\gamma}_i - \gamma_i). \tag{R2}$$

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 $h \rightarrow 0$. Under the conditions in Theorem 1, (R1-DR) is $O_p(\mathbb{E}[(\hat{\gamma}(t, X) - \gamma(t, X))] + \mathbb{E}[(\hat{\lambda}(t, X) - \lambda(t, X))]\sqrt{nh^{4+d_t}}) = o_p(1)$.

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 $\hat{\gamma}_{i\ell} = \hat{r}_\ell(t, X_i)$ using Z_ℓ^c for $i \in I_\ell$. Following the proof of Lemma A1 in CEINR, define $\hat{\Delta}_{i\ell} = \hat{\gamma}_{i\ell} - \gamma_i - \mathbb{E}[\hat{\gamma}_{i\ell} - \gamma_i]$ for $i \in I_\ell$. By construction and independence of Z_ℓ^c and $z_i, i \in I_\ell$, $\mathbb{E}[\hat{\Delta}_{i\ell}|Z_\ell^c] = 0$ and $\mathbb{E}[\hat{\Delta}_{i\ell}\hat{\Delta}_{j\ell}|Z_\ell^c] = 0$ for $i, j \in I_\ell$. For $i \in I_\ell$ and for all t , $h^{d_t}\mathbb{E}[\hat{\Delta}_{i\ell}^2|Z_\ell^c] = O_p(h^{d_t} \int (\hat{\gamma}_{i\ell} - \gamma_i)^2 f_X(X_i) dX_i) \xrightarrow{p} 0$ by Assumptions 1(ii) and 3(i). Then $\mathbb{E}\left[\left(\sqrt{h^{d_t}/n} \sum_{i \in I_\ell} \hat{\Delta}_{i\ell}\right)^2 \middle| Z_\ell^c\right] = (h^{d_t}/n) \sum_{i \in I_\ell} \mathbb{E}[\hat{\Delta}_{i\ell}^2|Z_\ell^c] \leq h^{d_t} \int (\hat{\gamma}_{i\ell} - \gamma_i)^2 f_X(X_i) dX_i \xrightarrow{p} 0$. The conditional Markov inequality implies that $\sqrt{h^{d_t}/n} \sum_{i \in I_\ell} \hat{\Delta}_{i\ell} \xrightarrow{p} 0$.

The analogous results also hold for $\hat{\Delta}_{i\ell} = K_h(T_i - t)\lambda_i(\gamma_i - \hat{\gamma}_{i\ell}) - \mathbb{E}[K_h(T_i - t)\lambda_i(\gamma_i - \hat{\gamma}_{i\ell})]$ in (R1-1) and $\hat{\Delta}_{i\ell} = K_h(T_i - t)(\hat{\lambda}_{i\ell} - \lambda_i)(Y_i - \gamma_i) - \mathbb{E}[K_h(T_i - t)(\hat{\lambda}_{i\ell} - \lambda_i)(Y_i - \gamma_i)]$ in (R1-2). In particular, for (R1-2), $h^{d_t}\mathbb{E}[\hat{\Delta}_{i\ell}^2|Z_\ell^c] = O_p\left(\int k(u)^2 du \int_{\mathcal{X}} (\hat{\lambda}_{i\ell} - \lambda_i)^2 f_{TX}(t, X_i) dX_i\right) = o_p(1)$ by the smoothness condition and Assumption 3(i). So (R1-1) $\xrightarrow{p} 0$ and (R1-2) $\xrightarrow{p} 0$.

For (R2),

$$\begin{aligned} & \mathbb{E}\left[\left|\sqrt{h^{d_t}/n} \sum_{i \in I_\ell} K_h(T_i - t)(\hat{\lambda}_{i\ell} - \lambda_i)(\gamma_i - \hat{\gamma}_{i\ell})\right| \middle| Z_\ell^c\right] \\ & \leq \sqrt{nh^{d_t}} \int_{\mathcal{X}} \int_{\mathcal{T}} \left|K_h(T_i - t)(\hat{\lambda}_{i\ell} - \lambda_i)(\gamma_i - \hat{\gamma}_{i\ell})\right| f_{TX}(T_i, X_i) dT_i dX_i \\ & \leq \sqrt{nh^{d_t}} \int_{\mathcal{X}} f_{T|X}(t|X_i) \left|(\hat{\lambda}_{i\ell} - \lambda_i)(\gamma_i - \hat{\gamma}_{i\ell})\right| f_X(X_i) dX_i + o_p(\sqrt{nh^{d_t}}h^2) \\ & \leq \sqrt{nh^{d_t}} \left(\int_{\mathcal{X}} f_{T|X}(t|X_i)(\hat{\lambda}_{i\ell} - \lambda_i)^2 f_X(X_i) dX_i\right)^{1/2} \left(\int_{\mathcal{X}} f_{T|X}(t|X_i)(\hat{\gamma}_{i\ell} - \gamma_i)^2 f_X(X_i) dX_i\right)^{1/2} + o_p(1) \\ & \xrightarrow{p} 0 \end{aligned} \tag{11}$$

by Cauchy-Schwartz inequality, Assumption 3(ii), and $nh^{d_t+4} \rightarrow C$. So (R2) $\xrightarrow{p} 0$ follows by the conditional Markov and triangle inequalities.

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

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

For \mathbf{B}_t , $\mathbb{E} \left[\frac{K_h(T-t)}{f_{T|X}(t|X)} (Y - \gamma(t, X)) \right] = \mathbb{E} \left[\frac{1}{f_{T|X}(t|X)} \mathbb{E} [K_h(T-t) (\gamma(T, X) - \gamma(t, X)) | X] \right]$. A standard algebra for kernel yields

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

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

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

The asymptotic variance is determined by $h^{d_t} \mathbb{E} \left[((Y - \gamma(t, X)) K_h(T-t) / f_{T|X}(t|X))^2 \right]$. A standard algebra for kernel as above yields \mathbf{V}_t . Asymptotic normality follows directly from the Lindeberg-Lévy central limit theorem. Specifically, by the above calculation the condition $h^{d_t} \mathbb{E}[\psi(Z_i, \beta_t, \gamma, \lambda)^2] < \infty$ holds under the conditions given in the theorem. \square

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

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 $\hat{\beta}_t = n^{-1} \sum_{i=1}^n \hat{\psi}_{ti} = n^{-1} \sum_{i=1}^n (\psi_{ti} + R_{ti})$, 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 above in (R1-1), (R1-2), (R1-DR), and (R2). Thus $\hat{\theta}_t - \theta_{t\eta} = \eta^{-1} n^{-1} \sum_{i=1}^n (\psi_{t+i} - \psi_{t-i} + R_{t+i} - R_{t-i})$. Denote $f_{t|X_i} = f_{T|X}(t|X_i)$.

(i) By $\eta/h \rightarrow 0$ and a Taylor expansion, the variance of $\eta^{-1}n^{-1} \sum_{i=1}^n (\psi_{t+i} - \psi_{t-i})$ 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 $\eta^{-1}n^{-1} \sum_{i=1}^n (\psi_{t+i} - \psi_{t-i})$ is $\int (\partial_{t_1} K_h(T - t))^2 \mathbb{E}[(Y - \gamma(t, X))^2 | T, X] f_{T|X} / f_{t|X}^2 dT = h^{-(d_t+2)} \mathbb{E}[\text{var}(Y | T = t, X) / f_{T|X}(t | X)] \int k'(u)^2 du + o(h^{-(d_t+2)}) = O(h^{-(d_t+2)})$.

To control $\sqrt{nh^{d_t+2}} \eta^{-1} n^{-1} \sum_{i=1}^n (R_{t+i} - R_{t-i}) = o_p(1)$, the conditions (a) and (b) give a crude 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 ,

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

where the first equality is by integration by parts.

(ii) $\sqrt{nh^{d_t} \eta^2} (\hat{\theta}_t - \theta_{t\eta}) = \sqrt{nh^{d_t}} (\hat{\beta}_{t+} - \hat{\beta}_{t-} - (\beta_{t+} - \beta_{t-})) = \sqrt{nh^{d_t}} n^{-1} \sum_{i=1}^n (\psi_{t+i} - \psi_{t-i} + R_{t+i} - R_{t-i}) = \sqrt{nh^{d_t}} n^{-1} \sum_{i=1}^n (\psi_{t+i} - \psi_{t-i}) + o_p(1)$ 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. $\mathbb{E}[\psi_{t+i}\psi_{t-i}]$ is bounded by the order of

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

□

Proof of Theorem 3 The proofs of Theorem 20.6 and Theorem 20.7 in Hansen (2021) analyze $\|\hat{\gamma}_K - \gamma\|_{F_{TX}}^2$. We follow the same argument to analyze the $L_2(tX)$ norm.

We can write $Y = Z'_K \beta_K + e_K$, where e_K is the projection error. Define $\tilde{Z}_{Ki} \equiv \mathbf{Q}_K^{-1/2} Z_{Ki}$, $\tilde{\mathbf{Q}}_K \equiv n^{-1} \sum_{i=1}^n \tilde{Z}_{Ki} \tilde{Z}'_{Ki}$, and $\mathbf{Q}_{Kt} \equiv \int_{\mathcal{X}} Z_K(t, x) Z_K(t, x)' f_{TX}(t, x) dx$.

$$\begin{aligned}
\|\hat{\gamma}_K - \gamma\|_{F_{tX}}^2 &= \int_{\mathcal{X}} (Z_K(t, x)'(\hat{\beta}_K - \beta_K) - r_K(t, x))^2 f_{TX}(t, x) dx \\
&= (\hat{\beta}_K - \beta_K)' \left(\int_{\mathcal{X}} Z_K(t, x) Z_K(t, x)' f_{TX}(t, x) dx \right) (\hat{\beta}_K - \beta_K) \quad (12)
\end{aligned}$$

$$- 2(\hat{\beta}_K - \beta_K)' \left(\int_{\mathcal{X}} Z_K(t, x) r_K(t, x) f_{TX}(t, x) dx \right) \quad (13)$$

$$\begin{aligned}
&+ \int_{\mathcal{X}} r_K(t, x)^2 f_{TX}(t, x) dx \\
&= O_p \left((\hat{\beta}_K - \beta_K)' \mathbf{Q}_{Kt} (\hat{\beta}_K - \beta_K) + \|r_K\|_{F_{tX}}^2 \right).
\end{aligned}$$

Assumption 5(iii) implies $\|r_K\|_{F_{tX}} = O(K^{-\alpha})$. Consider the term in (13). For the $L_2(TX)$ norm, $\int_{\mathcal{T} \times \mathcal{X}} Z_K(t, x) r_K(t, x) dF_{TX}(t, x) = 0$ due to the regression and projection errors. But this is not the case for the $L_2(tX)$ norm.

$$\begin{aligned}
\left| (\hat{\beta}_K - \beta_K)' \left(\int_{\mathcal{X}} Z_K(t, x) r_K(t, x) f_{TX}(t, x) dx \right) \right| &\leq \int_{\mathcal{X}} |(\hat{\gamma}_K(t, x) - \gamma_K(t, x)) r_K(t, x)| f_{TX}(t, x) dx \\
&\leq \|(\hat{\gamma}_K - \gamma_K) r_K\|_{F_{tX}} \\
&\leq \|\hat{\gamma}_K - \gamma_K\|_{F_{tX}} \|r_K\|_{F_{tX}},
\end{aligned}$$

where $\|\hat{\gamma}_K - \gamma_K\|_{F_{tX}}^2 = \int_{\mathcal{X}} (Z_K(t, x)'(\hat{\beta}_K - \beta_K))^2 f_{TX}(t, x) dx = (\hat{\beta}_K - \beta_K)' \mathbf{Q}_{Kt} (\hat{\beta}_K - \beta_K)$.

Write $\hat{\beta}_K = (\mathbf{Z}'_K \mathbf{Z}_K)^{-1} \mathbf{Z}'_K \mathbf{Y}$, where $\mathbf{Z}_K \equiv (Z_1, \dots, Z_n)'$ and $\mathbf{Y} \equiv (Y_1, \dots, Y_n)'$. We show the

term in (12)

$$\begin{aligned}
(\hat{\beta}_K - \beta_K)' \mathbf{Q}_{Kt} (\hat{\beta}_K - \beta_K) &= (\mathbf{e}'_K \mathbf{Z}_K) (\mathbf{Z}'_K \mathbf{Z}_K)^{-1} \mathbf{Q}_{Kt} (\mathbf{Z}'_K \mathbf{Z}_K)^{-1} (\mathbf{Z}'_K \mathbf{e}_K) \\
&= n^{-2} (\mathbf{e}'_K \tilde{\mathbf{Z}}_K) \tilde{\mathbf{Q}}_K^{-1} \mathbf{Q}_K^{-1/2} \mathbf{Q}_{Kt} \mathbf{Q}_K^{-1/2} \tilde{\mathbf{Q}}_K^{-1} (\tilde{\mathbf{Z}}'_K \mathbf{e}_K) \\
&\leq \left(\lambda_{\max} \left(\tilde{\mathbf{Q}}_K^{-1} \mathbf{Q}_K^{-1/2} \mathbf{Q}_{Kt} \mathbf{Q}_K^{-1/2} \tilde{\mathbf{Q}}_K^{-1} \right) \right)^2 \left(n^{-2} \mathbf{e}'_K \tilde{\mathbf{Z}}_K \tilde{\mathbf{Z}}'_K \mathbf{e}_K \right) \\
&= O_p(1) \left(n^{-2} \mathbf{e}'_K \mathbf{Z}_K \mathbf{Q}_K^{-1} \mathbf{Z}'_K \mathbf{e}_K \right) \\
&= O_p(K/n).
\end{aligned}$$

The last equality follows the proof of Theorem 20.7 in Hansen (2021). The above inequality is by the Quadratic Inequality, where $\lambda_{\max}(\mathbf{Q})$ denotes the largest eigenvalue of a matrix \mathbf{Q} . By the Schwarz Matrix Inequality, $\lambda_{\max}(\tilde{\mathbf{Q}}_K^{-1} \mathbf{Q}_K^{-1/2} \mathbf{Q}_{Kt} \mathbf{Q}_K^{-1/2} \tilde{\mathbf{Q}}_K^{-1}) \leq \lambda_{\max}(\tilde{\mathbf{Q}}_K^{-1}) \lambda_{\max}(\mathbf{Q}_K^{-1/2}) \lambda_{\max}(\mathbf{Q}_{Kt}) \times \lambda_{\max}(\mathbf{Q}_K^{-1/2}) \lambda_{\max}(\tilde{\mathbf{Q}}_K^{-1})$, which is $O_p(1)$ by Assumption 5(i) and Theorem 20.5 in Hansen (2021). We notice that \mathbf{Q}_{Kt} does not affect the bound.

Putting together, we obtain $\|\hat{\gamma} - \gamma\|_{F_{tX}}^2 = O_p(K/n + K^{-2\alpha})$. \square

Proof of Theorem 4 (i) Let $f_{*b} \equiv \arg \min_{f \in \mathcal{W}^{\beta, \infty}} \mathbb{E}[\ell_{tb}(f, Z)]$, $f_n \equiv \arg \min_{f \in \mathcal{F}_{DNN}, \|f\|_{\infty} \leq 2M} \|f - f_{*b}\|_{\infty}$, and $\epsilon_n \equiv \epsilon_{DNN} \equiv \|f_n - f_{*b}\|_{\infty}$. Let the bounded kernel function $\mathbf{k}() < \bar{k}$ for some constant $\bar{k} > 0$.

We modify equation (2.1) in FLM to the following:

$$|\ell_{tb}(f, Z) - \ell_{tb}(g, Z)| \leq \mathbf{K}_b(T - t) M |f(X) - g(X)| \leq \frac{\bar{k}^{dt}}{b^{dt}} M |f(X) - g(X)|, \quad (2.1-1)$$

$$2(\mathbb{E}[\ell_{tb}(f, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)]) = \|f - \gamma\|_{F_{tX}}^2 + O(b^2). \quad (2.1-2)$$

Lemma 8 in FLM and the bounded kernel $\mathbf{k}()$ imply the Lipschitz condition (2.1-1). The key of our modification is the condition (2.1-2) that replaces $c_1 \mathbb{E}[(f - f_*)^2] \leq \mathbb{E}[\ell(f, Z)] - \mathbb{E}[\ell(f_*, Z)] \leq c_2 \|\hat{f} - f_*\|_{L_2(X)}^2$ in FLM's (2.1). We prove (2.1-2) at the end of this proof. In the proof of Theorem 1 in FLM, the main decomposition in equation (A.1) starts with the inequality in their equation (2.1): $c_1 \|\hat{f} - f_*\|_{L_2(X)}^2 \leq \mathbb{E}[\ell(\hat{f}, Z)] - \mathbb{E}[\ell(f_*, Z)]$. This is the only place where this inequality is used. We modify it to (2.1-2) that implies $\|\hat{f}_b - \gamma\|_{F_{tX}}^2 = O_p(\mathbb{E}[\ell_{tb}(\hat{f}_b, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)]) + O_p(b^2)$. Thus we can bound $\|\hat{f}_b - \gamma\|_{F_{tX}}^2$ using the bound of $\mathbb{E}[\ell_{tb}(\hat{f}_b, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)]$. We modify (A.1) in FLM and bound

$$\begin{aligned}
&\mathbb{E}[\ell_{tb}(\hat{f}_b, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)] \\
&\leq (\mathbb{E} - \mathbb{E}_n)[\ell_{tb}(\hat{f}_b, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)] + \mathbb{E}_n[\ell_{tb}(\hat{f}_b, Z) - \ell_{tb}(f_{*b}, Z)].
\end{aligned} \quad (14)$$

To bound the second bias term, FLM only use the inequality in (2.1) $\mathbb{E}[\ell(f, Z)] - \mathbb{E}[\ell(f_*, Z)] \leq c_2 \|\hat{f} - f_*\|_{L_2(X)}^2$ in the second inequality in their (A.2). We use (2.1-2) that implies $\mathbb{E}[\ell_{tb}(f_n, Z) - \ell_{tb}(f_{*b}, Z)] = O_p(\|f_n - \gamma\|_{F_{tX}}^2) + O(b^2) = O_p(\epsilon_n^2 + b^2)$.

Next we modify the proof of Theorem 1 in FLM by replacing all (ℓ, \hat{f}, f_*) with $(\ell_{tb}, \hat{f}_b, f_{*b})$ for

any $t \in \mathcal{T}$ and b . We only point out the key modifications of their proof in the following.

By (2.1-1), $\text{var}[\ell_{tb}(f_n, z) - \ell_{tb}(f_{*b}, z)] = O_p(M^2 \|f_n - f_{*b}\|_\infty^2 / b^{d_t})$. Thus (A.2) in FLM is modified to $c_2 \epsilon_n^2 + \epsilon_n \sqrt{\frac{2C_\ell^2 \tilde{\gamma}}{nb^{d_t}}} + \frac{7C_\ell M \tilde{\gamma}}{nb^{d_t}}$.

Similarly the last equation on page 201 in FLM is modified to $\mathbb{V}[g] = \mathbb{E}[|\ell_{tb}(f, z) - \ell_{tb}(f_{*b}, z)|^2] \leq C_\ell^2 \mathbb{E}[(f - f_{*b})^2 \mathbf{K}_b(T - t)^2] = O(C_\ell^2 M r_0^2 / b^{d_t})$. Thus the statement for (A.7) in FLM is modified to the following: we find that $(\mathbb{E} - \mathbb{E}_n)[\ell_{tb}(\hat{f}, z) - \ell_{tb}(f_{*b}, z)] = O_p\left(6\mathbb{E}_\eta R_n \mathcal{G} + \sqrt{\frac{2C_\ell^2 r_0^2 \tilde{\gamma}}{nb^{d_t}}} + \frac{23 \cdot 3MC_\ell}{3} \frac{\tilde{\gamma}}{nb^{d_t}}\right)$.

On page 202 of FLM, the bound of $\mathbb{E}_n R_n \mathcal{G}$ is multiplied by b^{-d_t} . This is because in Lemma 2 and Lemma 3 in FLM, the Lipschitz condition is modified by $|\phi(f_1) - \phi(f_2)| \leq L|f_1 - f_2|(\bar{k}/b)^{d_t}$.¹⁹ It follows that (A.9) in FLM is modified to $r_0 \cdot \left(\frac{K\sqrt{C}}{b^{d_t}} \sqrt{\frac{WL \log W}{n} \log n} + \sqrt{\frac{2C_\ell^2 \tilde{\gamma}}{nb^{d_t}}}\right) + c_2 \epsilon_n^2 + \epsilon_n \sqrt{\frac{2C_\ell^2 \tilde{\gamma}}{nb^{d_t}}} + 30MC_\ell \frac{\tilde{\gamma}}{nb^{d_t}}$. And the bound in (A.10) in FLM is multiplied by b^{-d_t} . Thus (A.14) in FLM is modified to

$$\bar{r} = \frac{8}{c_1} \left(\frac{K\sqrt{C}}{b^{d_t}} \sqrt{\frac{WL \log W}{n} \log n} + \sqrt{\frac{2C_\ell^2 \tilde{\gamma}}{nb^{d_t}}} \right) + \left(\sqrt{\frac{2(c_2 \vee 1)}{c_1}} \epsilon_n + \sqrt{\frac{120MC_\ell}{c_1} \frac{\tilde{\gamma}}{nb^{d_t}}} \right) + \frac{r_*}{b^{d_t}}.$$

Therefore (A.17) in FLM is modified to

$$C' \left(\sqrt{\frac{WL \log W}{nb^{2d_t}} \log n} + \sqrt{\frac{\log \log n + \gamma}{nb^{d_t}}} + \epsilon_n \right)$$

with some constant $C' > 0$ that does not depend on n . Thus we can optimize the upper bound on page 206 of FLM,

$$\bar{r} \leq C' \left(\sqrt{\frac{\epsilon_n^{-\frac{2d_x}{\beta}} (\log(1/\epsilon_n) + 1)^7}{nb^{2d_t}} \log n} + \sqrt{\frac{\log \log n + \gamma}{nb^{d_t}}} + \epsilon_n \right)$$

by choosing $\epsilon_n = (nb^{2d_t})^{-\frac{\beta}{2(\beta+d_x)}}$, $H \asymp \cdot (nb^{2d_t})^{\frac{d_x}{2(\beta+d_x)}} \log^2(nb^{2d_t})$, $L \asymp \cdot \log(nb^{2d_t})$. Hence, w.p.a.1, we can bound (14)

$$\mathbb{E}[\ell_{tb}(\hat{f}_b, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)] \leq \bar{r}^2 \leq C \left((nb^{2d_t})^{-\frac{\beta}{\beta+d_x}} \log^8 n + \frac{\log \log n + \gamma}{nb^{d_t}} \right).$$

The remaining proof is to show (2.1-2). We add and subtract $\gamma(T, X)$ to the loss function, and by the law of iterated expectations, we obtain $2\mathbb{E}[\ell_{tb}(f, Z)] = \mathbb{E}[\text{var}(Y|T, X) \mathbf{K}_b(T - t)] +$

¹⁹We might obtain a tighter bound by incorporating the kernel function in Lemma 2 and Lemma 3 in FLM. For example, in Lemma 2, we might bound the Rademacher complexity to $2L\mathbb{E}_\eta [\sup_{f \in \mathcal{F}} n^{-1} \sum_{i=1}^n \eta_i f(X_i) \mathbf{K}_b(T_i - t)]$. Such extension is out of the scope of this paper.

$\mathbb{E}[(\gamma(T, X) - f(X))^2 \mathbf{K}_b(T - t)]$. Since the first term does not depend on f , we focus on the second term $Q_b(f) \equiv \mathbb{E}[(\gamma(T, X) - f(X))^2 \mathbf{K}_b(T - t)]$.

Let $Q(f) \equiv \|f - \gamma\|_{F_{tX}}^2$. For any f , a standard algebra yields $Q_b(f) = Q(f) + b^2 \mathbf{B}(f) + o(b^2)$, where $\mathbf{B}(f) \equiv \mathbb{E}[(\partial_t \gamma)^2 f_{t|X} + (\gamma - f)^2 \partial_t^2 f_{t|X} / 2 + 2\partial_t \gamma (\gamma - f) \partial_t f_{t|X} + (\gamma - f) \partial_t^2 \gamma f_{t|X}] \int u^2 \mathbf{k}(u) du$, under Assumption 7(i) that the second derivatives $\partial_t^2 f_{t|X}$ and $\partial_t^2 \gamma$ are bounded and continuous. Note that $\gamma(t, x) = \arg \min_f \lim_{b \rightarrow 0} \mathbb{E}[\ell_{tb}(f, Z_i)] = \arg \min_f Q(f)$. Therefore,

$$\begin{aligned} 2(\mathbb{E}[\ell_{tb}(f, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)]) &= Q_b(f) - Q_b(f_{*b}) \\ &= \|f - \gamma\|_{F_{tX}}^2 - \|f_{*b} - \gamma\|_{F_{tX}}^2 + b^2(\mathbf{B}(f) - \mathbf{B}(f_{*b})) + o(b^2). \end{aligned} \quad (15)$$

Next we show that $Q(f_{*b}) = \|f_{*b} - \gamma\|_{F_{tX}}^2 = O(b^2)$. We find a bound for $Q_b(f_{*b})$ by the definition of the minimizers: $Q(f_{*b}) \geq Q(\gamma) = 0$ and $Q_b(\gamma) \geq Q_b(f_{*b})$. So $Q_b(f_{*b}) - Q(f_{*b}) = b^2 \mathbf{B}(f_{*b}) + o(b^2) \leq Q_b(f_{*b}) \leq Q_b(\gamma) = b^2 \mathbf{B}(\gamma) + o(b^2)$. Therefore, $o(b^2) \leq Q(f_{*b}) = \|f_{*b} - \gamma\|_{F_{tX}}^2 \leq b^2(\mathbf{B}(\gamma) - \mathbf{B}(f_{*b})) + o(b^2)$.

By (16), $\|f - \gamma\|_{F_{tX}}^2 = 2(\mathbb{E}[\ell_{tb}(f, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)]) + \|f_{*b} - \gamma\|_{F_{tX}}^2 + b^2(\mathbf{B}(f_{*b}) - \mathbf{B}(f)) + o(b^2) = O(\mathbb{E}[\ell_{tb}(f, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)]) + O(b^2)$. We obtain (2.1-2).

(ii) The *deep MLP-ReLU network uniform-kernel estimator* for any $t \in \mathcal{T}$ is

$$\hat{f}_{tb}(X_i) \equiv \arg \min_{f_\theta \in \mathcal{F}_{MLP}, \|f_\theta\|_\infty \leq 2M} \sum_{i=1}^n \frac{1}{2} (Y_i - f(X_i))^2 \times \Pi_{j=1}^{d_t} \mathbf{1}\{|T_{ij} - t_j| \leq b\} / (2b).$$

Let $\mathcal{T}_{tb} \equiv \Pi_{j=1}^{d_t} [t_j - b, t_j + b]$. As $\Pi_{j=1}^{d_t} \mathbf{1}\{|T_j - t_j| \leq b\} = \mathbf{1}\{T \in \mathcal{T}_{tb}\}$, we can treat $\mathbf{1}\{T \in \mathcal{T}_{tb}\}$ as a discrete variable for any fixed b and re-write $\hat{f}_{tb}(X_i) = \arg \min_{f_\theta \in \mathcal{F}_{MLP}, \|f_\theta\|_\infty \leq 2M} \sum_{\{i: T \in \mathcal{T}_{tb}\}} \frac{1}{2} (Y - f_\theta)^2$.

Then we can analyze $\hat{f}_{tb}(x)$ as an estimator of $f_{*b}(x) \equiv \mathbb{E}[Y|X = x, T \in \mathcal{T}_{tb}] = \mathbb{E}[Y|X = x, \mathbf{1}\{T \in \mathcal{T}_{tb}\} = 1]$ for any fixed b . Let $n_b \equiv \sum_{i=1}^n \mathbf{1}\{T_i \in \mathcal{T}_{tb}\}$ be the number of observations used in \hat{f}_{tb} . The results in FLM can be applied to every category of the discrete data; see also Theorem 1 in Farrell et al. (2021a). Therefore as $n_b \rightarrow \infty$, $\|\hat{f}_{tb} - f_{*b}\|_{F_X}^2 \leq C \left(n_b^{-\frac{\beta}{\beta+d_x}} \log^8 n_b + \log \log n_b / n_b \right)$, w.p.a.1.

Observe that $n_b / (n(2b)^{d_t}) = n^{-1} \sum_{i=1}^n \Pi_{j=1}^{d_t} \mathbf{1}\{|T_{ij} - t_j| \leq b\} / (2b) \equiv \hat{f}_T(t)$ is the uniform-kernel density estimator of $f_T(t)$. By the consistency of $\hat{f}_T(t)$ and the continuous mapping theorem, $\hat{f}_T(t)^{-1} - f_T(t)^{-1} = o_p(1)$ and hence $\hat{f}_T(t)^{-1} = n(2b)^{d_t} / n_b = O_p(1)$. Therefore $n_b \asymp n(2b)^{d_t}$. It follows that as $n(2b)^{d_t} \rightarrow \infty$, $\|\hat{f}_{tb} - f_{*b}\|_{F_X}^2 \leq C \left((n(2b)^{d_t})^{-\frac{\beta}{\beta+d_x}} \log^8(n(2b)^{d_t}) + \log \log(n(2b)^{d_t}) / (n(2b)^{d_t}) \right)$, w.p.a.1.

Next we show $f_{*b}(x) = \mathbb{E}[Y|X = x, T \in \mathcal{T}_{tb}] = \mathbb{E}[Y|X = x, T = t] + O(b^2)$. Consider $d_t = 1$ for

simple exposition, and the result for $d_t \geq 2$ follows immediately.

$$\begin{aligned}
P(Y \leq y, T \in \mathcal{T}_{tb}) &= F_{YT}(y, t+b) - F_{YT}(y, t-b) \\
&= F_{YT}(y, t) + b\partial_t F_{YT}(y, t) + \frac{b^2}{2}\partial_t^2 F_{YT}(y, t) + \frac{b^3}{3!}\partial_t^3 F_{YT}(y, t) \\
&\quad - \left(F_{YT}(y, t) - b\partial_t F_{YT}(y, t) + \frac{b^2}{2}\partial_t^2 F_{YT}(y, t) - \frac{b^3}{3!}\partial_t^3 F_{YT}(y, t) \right) + O(b^4) \\
&= 2b\partial_t F_{YT}(y, t) + \frac{2b^3}{3!}\partial_t^3 F_{YT}(y, t) + O(b^4).
\end{aligned}$$

Similarly, $P(T \in \mathcal{T}_{tb}) = 2bf_T(t) + \frac{2b^3}{3!}\partial_t^2 f_T(t) + O(b^4)$. Then

$$\begin{aligned}
f_{Y|T \in \mathcal{T}_{tb}}(x) &= \frac{\frac{\partial}{\partial y} P(Y \leq y, T \in \mathcal{T}_{tb})}{P(T \in \mathcal{T}_{tb})} = \frac{2bf_{YT}(y, t) + \frac{2b^3}{3!}\partial_t^2 f_{YT}(y, t) + O(b^4)}{2bf_T(t) + \frac{2b^3}{3!}\partial_t^2 f_T(t) + O(b^4)} \\
&= f_{Y|T}(y|t) + \frac{b^2}{3!} \frac{\partial_t^2 f_{YT}(y, t)}{f_T(t)} + O(b^3).
\end{aligned}$$

It follows that $\mathbb{E}[Y|X = x, T \in \mathcal{T}_{tb}] = \int y f_{Y|X, T \in \mathcal{T}_{tb}}(y|x) dy = \mathbb{E}[Y|X = x, T = t] + O(b^2)$, which implies $\|f_{*b} - \gamma(t, \cdot)\|_{F_X} = O(b^2)$.

Putting everything together, $\|\hat{f}_{tb} - \gamma\|_{F_{tX}}^2 \leq \int (\hat{f}_{tb}(x) - \gamma(t, x))^2 f_X(x) dx \|f_{T|X}(t|\cdot)\|_\infty = O_p(\|\hat{f}_{tb} - \gamma(t, \cdot)\|_{F_X}^2) = O_p\left((nb^{d_t})^{-\frac{\beta}{\beta+d_x}} \log^8(nb^{d_t}) + \log \log(nb^{d_t})/(nb^{d_t}) + b^4\right)$. \square

Proof of Corollary 2 The arguments in the proof of Theorem 4 hold with some modifications noted below.

(i) We modify (2.1-2) to the following:

$$4(\mathbb{E}[\ell_{tb}(f, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)]) = \|f - \gamma\|_{F_{tX}}^2 + O(b). \quad (2.1-2b)$$

The key modification is due to $\int_0^\infty k(u) du = 1/2$. For any f , a standard algebra yields $Q_b(f) = Q(f)/2 + b\mathbf{B}(f) + o(b)$, where $\mathbf{B}(f) \equiv \mathbb{E}[2\partial_t \gamma(\gamma - f)f_{t|X} + (\gamma - f)^2 \partial_t f_{t|X}] \int_0^\infty uk(u) du$. Therefore,

$$\begin{aligned}
2(\mathbb{E}[\ell_{tb}(f, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)]) &= Q_b(f) - Q_b(f_{*b}) \\
&= \|f - \gamma\|_{F_{tX}}^2/2 - \|f_{*b} - \gamma\|_{F_{tX}}^2/2 + b(\mathbf{B}(f) - \mathbf{B}(f_{*b})) + o(b). \quad (16)
\end{aligned}$$

Next we show that $Q(f_{*b}) = \|f_{*b} - \gamma\|_{F_{tX}}^2 = O(b)$. We find a bound for $Q_b(f_{*b})$ by the definition of the minimizers: $Q(f_{*b}) \geq Q(\gamma) = 0$ and $Q_b(\gamma) \geq Q_b(f_{*b})$. So $Q_b(f_{*b}) - Q(f_{*b})/2 = b\mathbf{B}(f_{*b}) + o(b) \leq Q_b(f_{*b}) \leq Q_b(\gamma) = b\mathbf{B}(\gamma) + o(b)$. Therefore, $o(b) \leq Q(f_{*b})/2 = \|f_{*b} - \gamma\|_{F_{tX}}^2/2 \leq b(\mathbf{B}(\gamma) - \mathbf{B}(f_{*b})) + o(b)$.

By (16), $\|f - \gamma\|_{F_{tX}}^2/2 = 2(\mathbb{E}[\ell_{tb}(f, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)]) + \|f_{*b} - \gamma\|_{F_{tX}}^2/2 + b(\mathbf{B}(f_{*b}) - \mathbf{B}(f)) + o(b) = O(\mathbb{E}[\ell_{tb}(f, Z)] - \mathbb{E}[\ell_{tb}(f_{*b}, Z)]) + O(b)$. We obtain (2.1-2b).

(ii) Consider $d_t = 1$ for simple exposition. Let $\mathcal{T}_{tb} \equiv \mathbf{1}\{T \in [t, t+b]\}$. Then $\hat{f}_{tb}(X_i) \equiv \arg \min_{f_\theta \in \mathcal{F}_{MLP}, \|f_\theta\|_\infty \leq 2M} \sum_{i=1}^n \frac{1}{2} (Y_i - f(X_i))^2 \times \mathbf{1}\{T_i \in [t, t+b]\}$. The arguments for the inte-

rior t hold, so as $nb \rightarrow \infty$, $\|\hat{f}_{tb} - f_{*b}\|_{F_X}^2 \leq C \left((nb)^{-\frac{\beta}{\beta+d_x}} \log^8(nb) + \log \log(nb)/(nb) \right)$, w.p.a.1.

Next we show $f_{*b}(x) = \mathbb{E}[Y|X = x, T \in \mathcal{T}_{tb}] = \mathbb{E}[Y|X = x, T = t] + O(b)$.

$$\begin{aligned} P(Y \leq y, T \in \mathcal{T}_{tb}) &= F_{YT}(y, t+b) - F_{YT}(y, t) \\ &= F_{YT}(y, t) + b\partial_t F_{YT}(y, t) + \frac{b^2}{2}\partial_t^2 F_{YT}(y, t) - F_{YT}(y, t) + O(b^3) \\ &= b\partial_t F_{YT}(y, t) + \frac{b^2}{2}\partial_t^2 F_{YT}(y, t) + O(b^3). \end{aligned}$$

Similarly, $P(T \in \mathcal{T}_{tb}) = bf_T(t) + \frac{b^2}{2}\partial_t^2 f_T(t) + O(b^3)$. Then $f_{Y|T \in \mathcal{T}_{tb}}(x) = f_{Y|T}(y|t) + \frac{b}{2} \frac{\partial_t f_{YT}(y, t)}{f_T(t)} + O(b^2)$. It follows that $\mathbb{E}[Y|X = x, T \in \mathcal{T}_{tb}] = \int y f_{Y|X, T \in \mathcal{T}_{tb}}(y|x) dy = \mathbb{E}[Y|X = x, T = t] + O(b)$, which implies $\|f_{*b} - \gamma(t, \cdot)\|_{F_X} = O(b)$. We complete the proof. \square

Proof of Theorem 5 (i) We show the remainder terms (R1-1), (R1-2), (R1-DR), and (R2) are $o_p(1)$ uniformly over $t \in \mathcal{T}_0$. Denote the nuisance estimators $\hat{\gamma}_{i\ell t} = \hat{r}_\ell(t, X_i)$ and $\hat{\lambda}_{i\ell t} = 1/\hat{f}_\ell(t|X_i)$ that use Z_ℓ^c for $i \in I_\ell$. Denote $\hat{g}(t) = n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} K_h(T_i - t) \Delta_{i\ell}(t)$ and $W_{i\ell}(t) = K_h(T_i - t) \Delta_{i\ell}(t) - \mathbb{E}[K_h(T_i - t) \Delta_{i\ell}(t)]$, where $\Delta_{i\ell}(t) = (\hat{\lambda}_{i\ell t} - \lambda_{it})(\hat{\gamma}_{i\ell t} - \gamma_{it})$ for (R2).

We follow similar decompositions in the proof of Theorem 3.1 in Fan, Hsu, Lieli, and Zhang (2021). First, to show that $\sup_{t \in \mathcal{T}_0} \mathbb{E}[\hat{g}(t)] = o_p(\sqrt{\ln(n)/(nh^{d_t})})$, the same argument in (11) holds uniformly over $t \in \mathcal{T}_0$ by Assumption 8(i).

Since \mathcal{T}_0 is compact, it can be covered by a finite number M_n of cubes $\mathcal{C}_{k,n}$ with centered $t_{k,n}$ and length m_n , for $k = 1, \dots, M_n$. So $M_n \propto 1/m_n^{d_t}$. Decompose

$$\begin{aligned} \sup_{t \in \mathcal{T}_0} |\hat{g}(t) - \mathbb{E}[\hat{g}(t)]| &= \max_{1 \leq k \leq M_n} \sup_{t \in \mathcal{T}_0 \cap \mathcal{C}_{k,n}} |\hat{g}(t) - \mathbb{E}[\hat{g}(t)]| \\ &\leq \max_{1 \leq k \leq M_n} \sup_{t \in \mathcal{T}_0 \cap \mathcal{C}_{k,n}} |\hat{g}(t) - \hat{g}(t_{k,n})| \end{aligned} \quad (17)$$

$$+ \max_{1 \leq k \leq M_n} |\hat{g}(t_{k,n}) - \mathbb{E}[\hat{g}(t_{k,n})]| \quad (18)$$

$$+ \max_{1 \leq k \leq M_n} \sup_{t \in \mathcal{T}_0 \cap \mathcal{C}_{k,n}} |\mathbb{E}[\hat{g}(t_{k,n})] - \mathbb{E}[\hat{g}(t)]|. \quad (19)$$

For a positive constant C and positive sequences A_{1n} and A_{2n} given in Assumption 8(ii), let $\mathcal{F}_n(C) \equiv \{(\gamma^\dagger, \lambda^\dagger) : \sup_{(t,x) \in \mathcal{T}_0 \times \mathcal{X}} |\gamma^\dagger(t, x) - \gamma(t, x)| \leq CA_{1n}, \sup_{(t,x) \in \mathcal{T}_0 \times \mathcal{X}} |\lambda^\dagger(t, x) - \lambda(t, x)| \leq CA_{2n}\}$. Let $\mathcal{A}_{\ell n}(C) \equiv \{(\hat{\gamma}_{i\ell t}, \hat{\lambda}_{i\ell t}) \in \mathcal{F}_n(C)\}$ and $\mathcal{A}_n(C) \equiv \cap_{\ell=1}^L \mathcal{A}_{\ell n}(C)$. On $\mathcal{A}_n(C)$, i.e., $(\hat{\gamma}_{i\ell t}, \hat{\lambda}_{i\ell t}) \in \mathcal{F}_n(C)$ for $\ell = 1, \dots, L$, $\sup_{t \in \mathcal{T}_0, i \in I_\ell} |\Delta_{i\ell}(t)| \leq C^2 A_{1n} A_{2n} \equiv C^2 A_n$. Assumption 8(ii) implies that for any $\varepsilon > 0$, there exists a positive constant C , such that $P(\mathcal{A}_n(C)) \geq 1 - \varepsilon$ for n large enough.

Observe that

$$\begin{aligned} P(n^{-1}W_{i\ell}(t) > \eta_n, \mathcal{A}_{\ell n}(C)) &= \mathbb{E}[P(n^{-1}W_{i\ell}(t) > \eta_n | Z_\ell^c) \mathbf{1}\{\mathcal{A}_{\ell n}(C)\}] \\ &= \mathbb{E}\left[\int \mathbf{1}\{n^{-1}W_{i\ell}(t) > \eta_n\} f_Z(Z_i) dZ_i \mathbf{1}\{\mathcal{A}_{\ell n}(C)\}\right]. \end{aligned} \quad (20)$$

The second equality is due to cross-fitting with $(\hat{\gamma}, \hat{\lambda})$ using Z_ℓ^c .

We will use the following inequalities. By $\exp(w) \leq 1 + w + w^2$ for $|w| \leq 1/2$ and $1 + w \leq \exp(w)$ for $w \geq 0$, we have

$$\mathbb{E}[\exp(W)] \leq 1 + \mathbb{E}[W] + \mathbb{E}[W^2] \leq \exp(\mathbb{E}[W^2]) \quad (21)$$

for a random variable W satisfying $|W| \leq 1/2$ and $\mathbb{E}[W] = 0$. The Markov inequality for any positive sequence a_n : $P(W > \eta_n) \leq \mathbb{E}[\exp(a_n W)] / \exp(a_n \eta_n)$.

First consider (18). For any $\eta_n > 0$, $P(\max_{1 \leq k \leq M_n} |\hat{g}(t_{k,n}) - \mathbb{E}[\hat{g}(t_{k,n})]| > \eta_n) \leq M_n \sup_{t \in \mathcal{T}_0} P(|\hat{g}(t) - \mathbb{E}[\hat{g}(t)]| > \eta_n, \mathcal{A}_n(C)) + \varepsilon$. We show that for $t \in \mathcal{T}_0$,

$$\begin{aligned} & P(|\hat{g}(t) - \mathbb{E}[\hat{g}(t)]| > \eta_n, \mathcal{A}_n(C)) \\ &= P\left(\left|n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} W_{i\ell}(t)\right| > \eta_n, \mathcal{A}_n(C)\right) \\ &= P\left(n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} W_{i\ell}(t) > \eta_n, \mathcal{A}_n(C)\right) + P\left(n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} W_{i\ell}(t) < -\eta_n, \mathcal{A}_n(C)\right) \\ &\leq \sum_{\ell=1}^L \sum_{i \in I_\ell} \{P(n^{-1} W_{i\ell}(t) > \eta_n, \mathcal{A}_{\ell n}(C)) + P(-n^{-1} W_{i\ell}(t) > \eta_n, \mathcal{A}_{\ell n}(C))\} \\ &\leq \sum_{\ell=1}^L \sum_{i \in I_\ell} \mathbb{E}[\mathbf{1}\{\mathcal{A}_{\ell n}(C)\} \mathbb{E}[\exp(a_n n^{-1} W_{i\ell}(t)) + \exp(-a_n n^{-1} W_{i\ell}(t)) | Z_\ell^c]] / \exp(a_n \eta_n) \\ &\leq 2n \exp(-a_n \eta_n) \mathbb{E}[\exp(a_n^2 n^{-2} \mathbb{E}[W_{i\ell}(t)^2 | Z_\ell^c])]. \end{aligned} \quad (22)$$

The second inequality uses (20) and the Markov inequality. Due to cross-fitting, conditional on Z_ℓ^c , $\hat{\gamma}$ and $\hat{\lambda}$ are fixed functions. When $\mathbf{1}\{\mathcal{A}_{\ell n}(C)\} = 1$, $|a_n n^{-1} W_{i\ell}(t)| \leq 1/2$, for all t, i, ℓ and for n large enough by choosing $a_n = \sqrt{\ln(n) n h^{d_t}} / A_n$. So by (21), the last inequality holds.

We choose η_n such that $a_n \eta_n \rightarrow \infty$ and $a_n \eta_n \geq a_n^2 n^{-2} \mathbb{E}[W_{i\ell}(t)^2 | Z_\ell^c]$, so $\sup_{t \in \mathcal{T}_0} P(|\hat{g}(t) - \mathbb{E}[\hat{g}(t)]| > \eta_n, \mathcal{A}_{\ell n}(C)) = o_p(1)$. For n large enough, there exists a positive constant c_1 such that $\mathbb{E}[W_{i\ell}(t)^2 | Z_\ell^c] \leq c_1 h^{-d_t} A_n^2$. We can choose $a_n \eta_n = c_2 \ln(n)$ for some positive constant c_2 , so $\eta_n = c_2 \sqrt{\ln(n) / (n h^{d_t})} A_n$. Then we can choose M_n such that $P(\max_{1 \leq k \leq M_n} |\hat{g}(t_{k,n}) - \mathbb{E}[\hat{g}(t_{k,n})]| > \eta_n) \leq M_n 2n \exp(-c_2 \ln(n) + c_1 \ln(n)/n) + \varepsilon \leq 2M_n n^{-(c_2 - c_1 n^{-1} - 1)} + \varepsilon \leq 2\varepsilon$ for $c_2 \geq 2$ and n large enough. So $\max_{1 \leq k \leq M_n} |\hat{g}(t_{k,n}) - \mathbb{E}[\hat{g}(t_{k,n})]| = O_p(\eta_n) = o_p(\sqrt{\ln(n) / (n h^{d_t})})$.

For (17), the Lipschitz condition in Assumption 8(iii) implies $\sup_{t \in \mathcal{T}_0 \cap \mathcal{C}_{k,n}} |K_h(T_i - t) \Delta_{i\ell}(t) - K_h(T_i - t_{k,n}) \Delta_{i\ell}(t_{k,n})| \leq c_3 h^{-(d_t+1)} \sup_{t \in \mathcal{T}_0 \cap \mathcal{C}_{k,n}} \|t - t_{k,n}\| \leq c_3 h^{-(d_t+1)} m_n$, for some constant $c_3 > 0$ and the Euclidean norm of a vector $\|\cdot\|$. By choosing $m_n = o(\sqrt{\ln(n) h^{(d_t+2)}/n})$, $\max_{1 \leq k \leq M_n} \sup_{t \in \mathcal{T}_0 \cap \mathcal{C}_{k,n}} |\hat{g}(t) - \hat{g}(t_{k,n})| \leq c_3 h^{-(d_t+1)} m_n = o_p(\sqrt{\ln(n) / (n h^{d_t})})$.

By the same argument, we can show that for (19) $\max_{1 \leq k \leq M_n} \sup_{t \in \mathcal{T}_0 \cap \mathcal{C}_{k,n}} |\mathbb{E}[\hat{g}(t_{n,k})] - \mathbb{E}[\hat{g}(t)]| = o_p(\sqrt{\ln(n) / (n h^{d_t})})$.

The same arguments apply to (R1-1) and (R1-2) by defining $\Delta_{i\ell}(t)$ accordingly: let $\Delta_{i\ell}(t) =$

$\lambda_{ilt}(\hat{\gamma}_{ilt} - \gamma_{ilt})$ for (R1-1) and $\Delta_{il}(t) = (\hat{\lambda}_{ilt} - \lambda_{ilt})(Y_t - \gamma_{ilt})$ for (R1-2).

For (R1-DR), the argument for the pointwise convergence in the proof of Theorem 1 can be extended to uniform convergence by Assumption 8(ii).

(ii) The results in Theorem 2 can be extended to uniformity in t by the same argument in (i). \square

Proof of Theorem 6 The proof follows closely the proof of Theorem 5, so we only notice the difference to conserve space. The idea is that the derivations proceed conditional on U_i and using the law of iterated iterations. Let $\hat{g}(t) = n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} U_i K_h(T_i - t) \Delta_{il}(t)$. The main difference is in (18). We show that for $t \in \mathcal{T}_0$,

$$\begin{aligned}
& P(|\hat{g}(t) - \mathbb{E}[\hat{g}(t)]| > \eta_n, \mathcal{A}_n(C)) \\
&= P\left(\left|n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} U_i W_{il}(t)\right| > \eta_n, \mathcal{A}_n(C)\right) \\
&= P\left(n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} U_i W_{il}(t) > \eta_n, \mathcal{A}_n(C)\right) + P\left(n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} U_i W_{il}(t) < -\eta_n, \mathcal{A}_n(C)\right) \\
&\leq \sum_{\ell=1}^L \sum_{i \in I_\ell} \{P(n^{-1} U_i W_{il}(t) > \eta_n, \mathcal{A}_{\ell n}(C)) + P(-n^{-1} U_i W_{il}(t) > \eta_n, \mathcal{A}_{\ell n}(C))\} \\
&= \sum_{\ell=1}^L \sum_{i \in I_\ell} \mathbb{E}\left[P(n^{-1} W_{il}(t) > \eta_n/U_i, \mathcal{A}_{\ell n}(C)|U_i) \mathbf{1}\{U_i \geq 0\}\right. \\
&\quad + P(-n^{-1} W_{il}(t) > -\eta_n/U_i, \mathcal{A}_{\ell n}(C)|U_i) \mathbf{1}\{U_i < 0\} \\
&\quad \left. + P(n^{-1} W_{il}(t) > \eta_n/U_i, \mathcal{A}_{\ell n}(C)|U_i) \mathbf{1}\{U_i \geq 0\} + P(-n^{-1} W_{il}(t) > -\eta_n/U_i, \mathcal{A}_{\ell n}(C)|U_i) \mathbf{1}\{U_i < 0\}\right] \\
&\leq \sum_{\ell=1}^L \sum_{i \in I_\ell} \mathbb{E}\left[\mathbf{1}\{\mathcal{A}_{\ell n}(C)\} \mathbb{E}\left[\exp(a_n n^{-1} W_{il}(t)) + \exp(-a_n n^{-1} W_{il}(t)) \middle| Z_\ell^c, U_i\right] \exp(-a_n \eta_n/|U_i|)\right] \\
&\leq 2n \mathbb{E}[\exp(-a_n \eta_n/|U_i|)] \mathbb{E}[\exp(a_n^2 n^{-2} \mathbb{E}[W_{il}(t)^2 | Z_\ell^c])].
\end{aligned}$$

The same arguments following (22) are valid conditional on U_i . Due to U_i , here we choose $a_n \eta_n = c_2 \ln(n) \ln(n)$ for some positive constant c_2 . So $\eta_n = c_2 \sqrt{\ln(n)/(n h^{d_t})} \mathcal{A}_n \ln(n)$. Next we show that we can choose M_n and c_2 such that $P(\max_{1 \leq k \leq M_n} |\hat{g}(t_{k,n}) - \mathbb{E}[\hat{g}(t_{k,n})]| > \eta_n) \leq M_n 2n \mathbb{E}[\exp(-c_2 \ln(n) \ln(n)/|U_i| + c_1 \ln(n)/n)] + \varepsilon \leq 2M_n \mathbb{E}[n^{-c_2 \ln(n)/|U_i| + c_1/n+1}] + \varepsilon \leq 2\varepsilon$ for n large enough.

By Assumption 9, there exist some constants c_4, c_5 such that $P(-c_2 \ln(n)/|U_i| + c_1/n + 1 \geq 0) \leq P(|U_i| \geq c_2 \ln(n)/(c_1 + 1)) \leq c_4 \exp(-c_5 c_2 \ln(n)/(c_1 + 1)) \leq c_4 n^{-c_5 c_2/(c_1 + 1)}$. So $\mathbb{E}[n^{-c_2 \ln(n)/|U_i| + c_1/n+1}] \leq \mathbb{E}[n^{-c_2 \ln(n)/|U_i| + c_1/n+1} \mathbf{1}\{-c_2 \ln(n)/|U_i| + c_1/n + 1 \leq 0\}] + n^{c_1+1} P(-c_2 \ln(n)/|U_i| + c_1/n + 1 \geq 0)$, where the second term $\leq c_4 n^{c_1+1-c_5 c_2/(c_1+1)} = o(1)$ by choosing c_2 . Therefore we show that we can choose M_n and c_2 such that $P(\max_{1 \leq k \leq M_n} |\hat{g}(t_{k,n}) - \mathbb{E}[\hat{g}(t_{k,n})]| > \eta_n) \leq 2\varepsilon$ for n large enough.

For (17), $\sup_{t \in \mathcal{T}_0 \cap \mathcal{C}_{k,n}} |U_i K_h(T_i - t) \Delta_{il}(t) - U_i K_h(T_i - t_{k,n}) \Delta_{il}(t_{k,n})| \leq c_3 h^{-(d_t+1)} \sup_{t \in \mathcal{T}_0 \cap \mathcal{C}_{k,n}} \|t -$

$t_{k,n}||U_i| \leq c_3 h^{-(d_t+1)} m_n |U_i|$, for some constant $c_3 > 0$. We can choose $m_n = o(\sqrt{\ln(n)h^{(d_t+2)}/n})$ such that $\max_{1 \leq k \leq M_n} \sup_{t \in \mathcal{T}_0 \cap \mathcal{C}_{k,n}} |\hat{g}(t) - \hat{g}(t_{k,n})| \leq c_3 h^{-(d_t+1)} m_n n^{-1} \sum_{i=1}^n |U_i| = o_p(\sqrt{\ln(n)/(nh^{d_t})})$, since $n^{-1} \sum_{i=1}^n |U_i| = O_p(1)$.

The result for the partial effect follows the same argument. \square

Gateaux derivative Let the Dirac delta function $\delta_t(T) = \infty$ for $T = t$, $\delta_t(T) = 0$ for $T \neq t$, and $\int g(s)\delta_t(s)ds = g(t)$, for any continuous compactly supported function g .²⁰ For any $F \in \mathcal{F}$,

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

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

The influence function can be calculated as

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

In particular, we specify F_Z^h following equation (3.12) in Ichimura and Newey (2017). Let $K_h(Z) = \Pi_{j=1}^{d_z} k(Z_j/h)/h$, where $Z = (Z_1, \dots, Z_{d_z})'$ and k satisfies Assumption 2 and is continuously differentiable of all orders with bounded derivatives. Let $F^{\tau h} = (1 - \tau)F^0 + \tau F_Z^h$ with pdf with respect to a product measure given by $f^{\tau h}(z) = (1 - \tau)f^0(z) + \tau f^0(z)\delta_Z^h(z)$, where $\delta_Z^h(z) = K_h(Z - z)\mathbf{1}\{f^0(z) > h\}/f^0(z)$, a ratio of a sharply peaked pdf to the true density. Thus $f_{YTX}^h(y, t, x) = K_h(Y - y)K_h(T - t)K_h(X - x)\mathbf{1}\{f^0(z) > h\}$. It follows that $\lim_{h \rightarrow 0} f_T^h(t) = \lim_{h \rightarrow 0} K_h(T - t)$ and

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

²⁰Note that a nascent delta function to approximate the Dirac delta function is $K_h(T - t) = k((T - t)/h)/h$ such that $\delta_t(T) = \lim_{h \rightarrow 0} K_h(T - t)$.

$\mathbb{E}\left[\frac{d}{d\tau}\beta_t(F^{\tau h})\Big|_{\tau=0}\right] = \mathbb{E}\left[\gamma(t, X) - \beta_t + \frac{Y - \gamma(t, X)}{f_{T|X}(t|X)}K_h(T - t)\right] = O(h^2)$. So Neyman orthogonality holds as $h \rightarrow 0$.

References

- Aitchison, J. and C. G. G. Aitken (1976). Multivariate binary discrimination by the kernel method. *Biometrika* 63, 413–420.
- 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., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics* 47(2), 1148–1178.
- 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.
- Biau, G. and E. Scornet (2016). A random forest guided tour. *TEST* 25, 197–227.
- 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.
- Bravo, F., J. Escanciano, and I. van Keilegom (2020). Two-step semiparametric likelihood inference. *Annals of Statistics* 48, 1–26.
- 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. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 155(2), 138–154.
- Cattaneo, M. D., M. H. Farrell, and Y. Feng (2020). Large sample properties of partitioning-based series estimators. *Annals of Statistics* 48(3), 1718–1741.

- Cattaneo, M. D. and M. Jansson (2019). Average density estimators: Efficiency and bootstrap consistency. [arxiv:1904.09372v1](https://arxiv.org/abs/1904.09372).
- Cattaneo, M. D., M. Jansson, and X. Ma (2019). Two-step estimation and inference with possibly many included covariates. *Review of Economic Studies* 86(3), 1095–1122.
- 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). Chapter 76 Large sample sieve estimation of semi-nonparametric models. Volume 6 of *Handbook of Econometrics*, pp. 5549–5632. Elsevier.
- Chen, X., Z. Liao, and Y. Sun (2014). Sieve inference on possibly misspecified semi-nonparametric time series models. *Journal of Econometrics* 178, 639–658.
- Chen, X. and D. Pouzo (2015). Sieve wald and QLR inferences on semi/nonparametric conditional moment models. *Econometrica* 83(3), 1013–1079.
- 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](https://arxiv.org/abs/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](https://arxiv.org/abs/1802.08667).
- 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.
- Chiang, H. D., K. Kato, Y. Ma, and Y. Sasaki (2021). Multiway cluster robust double/debiased machine learning. *Journal of Business & Economic Statistics* 0(0), 1–11.
- Das, M., W. K. Newey, and F. Vella (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies* 70(1), 33–58.

- 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 (2021). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, forthcoming.
- 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 (2021a). Deep learning for individual heterogeneity: An automatic inference framework. arxiv:2010.14694.
- Farrell, M. H., T. Liang, and S. Misra (2021b). Deep neural networks for estimation and inference. *Econometrica* 89(1), 181–213.
- 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.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69, 383–393.
- Hansen, B. E. (2021). *Econometrics*. <https://www.ssc.wisc.edu/bhansen/econometrics/Econometrics.pdf>.
- 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., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4), 1161–1189.
- 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. Lettry (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics* 35(7), 814–840.

- Hsu, Y.-C., T.-C. Lai, and R. P. Lieli (2020). Estimation and inference for distribution and quantile functions in endogenous treatment effect models. *Econometric Reviews* 0(0), 1–38.
- Ichimura, H. and W. Newey (2017). The influence function of semiparametric estimators. Working paper.
- Imbens, G. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* 87(3), 706–710.
- Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*.
- 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.
- Khan, S. and E. Tamer (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica* 78(6), 2021–2042.
- 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. Uhlenborff, 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, D. S. (2009, 07). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies* 76(3), 1071–1102.
- 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. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79(1), 147–168.
- Newey, W. K. and J. R. Robins (2018). Cross-fitting and fast remainder rates for semiparametric estimation. arxiv:1801.09138.
- Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probability and Statistics* 213(57).
- Noack, C., T. Olma, and C. Rothe (2021). Flexible covariate adjustments in regression discontinuity designs. 2107.07942.
- Oprea, M., V. Syrgkanis, and Z. S. Wu (2019). Orthogonal random forest for causal inference. arxiv:1806.03467v3.
- Ouyang, D., Q. Li, and J. S. Racine (2009). Nonparametric estimation of regression functions with discrete regressors. *Econometric Theory* 25(1), 1–42.
- 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 (2019). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. *Econometric Theory* 35(5), 1048–1087.
- Sasaki, Y. and T. Ura (2021). Estimation and inference for policy relevant treatment effects. *Journal of Econometrics*, forthcoming.
- Sasaki, Y., T. Ura, and Y. Zhang (2021). Unconditional quantile regression with high dimensional data. arxiv:2007.13659.
- Semenova, V. and V. Chernozhukov (2020). Estimation and inference about conditional average treatment effect and other causal functions. arxiv:1702.06240v3.
- Su, L., T. Ura, and Y. Zhang (2019). Non-separable models with high-dimensional data. *Journal of Econometrics* 212(2), 646–677.
- 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.