Cross-Validated Conditional Density Estimation and Nonparametric Continuous Difference-in-Differences Models

Lucas Z. Zhang
Department of Economics
University of California, Los Angeles
lucaszz@g.ucla.edu

October 19, 2022

Abstract

In this paper, we study the conditional density estimation based on a representation using the orthonormal series expansion. In this series representation, each term takes the form of a multiplication of the basis term and its conditional expectation. For high-dimensional conditioning variables with suitable structures, these conditional expectations can be estimated using various machine learning methods. We propose a data-driven method of selecting the series terms based on a modified cross-validation procedure and we establish an oracle inequality on the estimation error of such estimator. The conditional densities have a wide range of applications in various fields in economics, and we add to this literature a new application of nonparametric difference-in-differences models with continuous treatments. For this application, we establish the identification, estimation, and inference results under the double/debiased machine learning framework, and we illustrate our methods by revisiting an empirical study (Duflo (2001)) of a large policy intervention in Indonesia.

Keywords: Cross-Validation; Nonparametric Estimation; Conditional Density; Oracle Inequality; Difference-in-Differences

JEL classification coes: C1; C13; C14

1 Introduction

Given random variables Y and X, researchers are often interested in the dependent relationship between these variables. The conditional density $f_{Y|X}$ is one of the most fundamental

statistical objects that summarize such relationship. Its role in economics is especially pronounced with a wide range of applications. For example, when studying the identification problems of structural economic models, conditional densities are used to establish the connection between what can be observed from the data and the assumptions made on the economic models (e.g. Matzkin (2007, 2013)). For another example, in the auction literature, conditional densities of the bids can be used to recover the private values of the bidders, and the conditional density of the private values are also a parameter of interest in its own right (e.g. Guerre et al. (2000); Perrigne and Vuong (2019)). Other recent examples where the conditional density plays a key role include but are not limited to: treatment effects with continuous treatment (e.g. Hirano and Imbens (2004); Kennedy et al. (2017); Su et al. (2019); Semenova and Chernozhukov (2021)), nonparametric estimation of nonseparable models (e.g. Altonji and Matzkin (2005); Matzkin (2015); Blundell et al. (2020)), and nonparametric estimation of counterfactual distributions (e.g. Fortin el al. (2011)). We will examine the role of the conditional density in these examples in detail in Section 2.

The literature on conditional density estimation is vast. The most well-known nonparametric method is perhaps the kernel method proposed in Rosenblatt (1969), which involves taking ratio $f_{Y,X}/f_X$ of kernel estimators, and many subsequent literature are devoted to the kernel bandwidth selection for such estimator, see for example, Hall et al. (1999, 2004) and the references therein. Other popular methods include those using locally polynomial regression studied in Fan et al. (1996) and Fan and Yim (2004), and more recently the ones using orthogonal series, see for example, Efromovich (2010); Izbicki and Lee (2016, 2017) and the references therein. However, each of the aforementioned estimators has drawbacks. Although kernel type of estimators have attractive theoretical properties, it becomes computationally intractable as the dimension of the conditioning variable grows. On the other hand, while the estimators studied Izbicki and Lee (2016, 2017) are designed for the setting with high-dimensional conditioning variables, they are not data-driven in the sense that the theoretical properties developed require knowledge of the unknown smoothness parameters.[1] Moreover, even the data-driven estimators from Hall et al. (2004), Fan and Yim (2004) and Efromovich (2010) have drawbacks: Hall et al. (2004) requires cross-validation searching over each covariates, which becomes computationally intractable as the dimension grows; similarly, the thresholding estimator from Efromovich (2010) requires tensor products of basis over each dimension; the cross-validated estimator proposed by Fan and Yim (2004) performs well in their simulations, but its theoretical properties have yet been studied.

To bridge the gap, we propose a conditional density estimator that is not only feasible

[1] Both papers propose cross-validation algorithms but the theoretical properties of the resulting estimators are not studied.

for high-dimensional conditioning variable settings but also data-driven. Suppose $\{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis in $L^2(\nu)$, we first show the series expansion,

$$f_{Y|X}(y|x) = \sum_{j=1}^{\infty} E[\phi_j(Y)|X = x]\phi_j(y)$$

holds under very general conditions. This series expansion motivates an estimator of the form considered in Izbicki and Lee (2017)

$$\hat{f}_J(y|x) = \sum_{j=1}^{J} \widehat{E}[\phi_j(Y)|X = x]\phi_j(y).$$

For high-dimensional conditioning variable X with suitable sparsity structures, the conditional expectation $E[\phi_j(Y)|X]$ in each series term can be estimated using many state-of-the-art machine learning estimators, such as the deep neural networks. To choose the tuning parameter J in a data-driven way, we resort to a cross-validation procedure, in which the series cutoff $\hat{J}$ is chosen by minimizing the empirical risk. Our final estimator takes the form of an average of sub-sample estimators using the training samples with this cutoff. Following the general strategy proposed by Lecué and Mitchell (2012), we establish an oracle inequality that provides an upper bound on the estimation error of our estimator. To the best of our knowledge, this is the first such result of a conditional density estimator that is both data-driven and feasible in high-dimensional settings. We recognize that there is an extensive literature on cross-validation, and due to space limitations, we refer the readers to Arlot and Celisse (2010) for a comprehensive survey.

To add to the growing literature in economics where the conditional density is of key interest, we study in detail an application in the context of difference-in-differences models. Difference-in-differences (DiD for short) is one of the most popular empirical research designs, and the theories have evolved to accommodate the rich complexities of the empirical research. From the early success of Card and Krueger (1994) in the study of minimum wage, to the generalizations of DiD to, for examples, semiparametric setting (Abadie (2005)), nonlinear setting (Athey and Imbens (2006)), multiple periods and/or staggered treatment timing settings (e.g. Callaway and Sant'Anna (2021); de Chaisemartin and D'Haultfoeuille (2020); Athey and Imbens (2022)). Although most of the literature has focused on DiD with binary or discrete treatments, recent studies, such as Callaway et al. (2021) and D'Haultfoeuille et al. (2021), have considered the DiD settings with continuous treatment. Our methods expand upon this new line of research by considering the setting similar to Abadie (2005) but with continuous treatments.

In particular, we identify the average treatment effect on the treated (ATT) at any continuous treatment intensities under the conditional parallel trend assumption allowing for covariates. In this setting, the conditional density of the treatment conditional on covariates enters as a generalized propensity score. The fully nonparametric estimator of ATT involves averaging over estimated infinite-dimensional nuisance parameters, which is known to incur large biases. To address this issue, in such DiD setting with binary or discrete treatment, Sant'Anna and Zhao (2020) proposes doubly robust estimators while Chang (2020) studies doubly/debiased machine learning estimators, in which case the latter allows for high-dimensional controls. Following Chang (2020), we adopt and extend the double/debiased machine learning (DML) framework from CCDDHNR (2018) to allow for continuous treatment. To illustrate the usefulness of our methods, we revisit Duflo (2001), in which the author studies the effect of a large policy intervention in Indonesia (INPRES) on educational outcomes. One of the results in Duflo (2001) is estimated using a binary DiD comparing high vs. low treatment intensity regions, in which the author finds a small positive effect (but not statistically significant). In contrast, we allow the high-intensity regions to have varied treatment intensities instead of being grouped into one category. The ATTs at different treatment intensities estimated using our methods vary widely, suggesting significant heterogeneity.

The rest of the paper is organized as follows. In section 2, we motivate by providing a more detailed review of the previously mentioned examples that involve conditional densities. In section 3, we first show the validity of the series representations of the conditional densities, then we discuss the construction of our cross-validated estimator in detail, and finally we establish the theoretical properties of our estimator. In section 4, we formally set up the DiD models with continuous treatment, show its identification, estimation, and inference under DML framework, and illustrate the usefulness of our results with an empirical application. We conclude in Section 5. All the proofs will be given in the appendix.

2 Applications

In this section, we discuss several empirical examples where the conditional densities naturally occurs and the estimation of which is crucial.

Example 2.1 (**First Price Auction**)**.** Consider the first price auction in the independent private values (IPV) setting studied in Guerre et al. (2000). $I \geq 2$ bidders have i.i.d. private values $\{V_i\}_{i=1}^I$ with $V_i \in [v_L, v_H] \subset \mathbf{R}$. Each bidder bids $B_i = s(V_i)$ that maximizes the expected value. If the equilibrium bid function s is monotonic, then using the first order

condition, the unobserved private value V_i can be written as

$$V_i = B_i + \frac{1}{I-1}\frac{G(B_i|I)}{g(B_i|I)}$$

where $G(\cdot|I), g(\cdot|I)$ denote the observed equilibrium bid distribution and the density respectively. This is the main identification equation that enables the researcher to identify and estimate the model primitives (V_i, F_V).

Guerre et al. (2000) formulates the nonparametric estimation problem as following: econometricians observe $\{(B_{il})_{i=1}^{I_l}, X_l, I_l\}_{l=1}^{L}$, B_{il} denotes the equilibrium bids of individual i in auction l, L denotes the number of auctions, I_l denotes the number of bidders for each auction l and is also assumed to capture unobserved heterogeneity, X_l denotes auction-level covariates. The pseudo values $\hat{V}_{il}$ are estimated using the first order condition

$$\hat{V}_{il} = B_{il} = \frac{1}{I_l - 1}\frac{\hat{G}(B_{il}|X_l, I_l)}{\hat{g}(B_{il}|X_l, I_l)}$$

and the (conditional) value density is then estimated as

$$\hat{f}(v|x, I) \equiv \hat{f}(\hat{v}|x, I).$$

While Guerre et al. (2000) estimates the conditional distributions/densities using kernel methods, as the dimension of the covariates X grows, the traditional kernel methods becomes intractable. Haile et al. (2006) assumes $V_i = X'\beta + \tilde{V}_i$ with $\tilde{V}_i$ independent of X, which essentially collects the effects of the potentially high-dimensional covariates X into a single index. Perrigne and Vuong (2019) suggests an alternative single-index restriction by assuming the value distribution takes the form $F(V|X'\beta)$. While these dimensional reduction methods are easy to implement, they can suffer from significant misspecification errors if the single index assumptions don't hold.

Example 2.2 (**Nonparametric Nonseparable Models**)**.** In many nonparametric nonseparable models, the parameters of interests can be constructively identified as functions of conditional densities of observed variables. For example, Altonji and Matzkin (2005) study the model of the form $Y = m(X, \epsilon_1, \cdots, \epsilon_K)$ where Y, X are observable, $(\epsilon_1, \cdots, \epsilon_K)$ are unobservable, and there exists an external observable Z such that $X \perp (\epsilon_1, \cdots, \epsilon_K)|Z$. Specifically, they consider the local average response $\beta(x)$, which is defined as the average derivative of m with respect to x over the distribution $f_{\epsilon_1,\cdots,\epsilon_K}|X = x$. They show that the identified $\beta(x)$ takes the form

$$\beta(x) = \int \frac{\partial E[Y|X = x, Z = z]}{\partial x} f_{Z|X=x}(z)dz.$$

A nonparametric estimator can be constructed based on this expression, which requires estimation of the conditional density $f_{Z|X}$. A recent related work by Blundell et al. (2020) that studies the individual counterfactuals also uses the external variables. Similarly, the identification and estimation results established in that study rely on the conditional density $f_{Y|X,Z}$ and its estimator.

Matzkin (2015) provides several examples in the context of nonparametric nonseparable samultaneous equations. The first model studied in Matzkin (2015) concerns the system of simultaneous equations,

$$\begin{aligned} Y_1 &= m^1(X_1, Y_{-1}, Z, \epsilon_1) \\ &\cdots \\ Y_G &= m^G(X_G, Y_{-G}, Z, \epsilon_G) \end{aligned}$$

with $(Y_1, \cdots, Y_G)$ being observable endogenous variables, $(Z, X_1, \cdots, X_G)$ being the obervable exogenous covariates, and $(\epsilon_1, \cdots, \epsilon_G)$ being the unobservable variables. Under certain conditions, Matzkin (2015) establishes the constructive identification of the derivatives of these structural functions (eg. $\partial m^g / \partial y_j$) as functionals of the conditional density $f_{Y|X}$. Matzkin (2015) proposes average derivative type of estimators for these derivatives that require nonparametric estimation of the conditional densities.

The second model Matzkin (2015) considers is the two equations one instrument model

$$\begin{aligned} Y_1 &= m^1(Y_2, \epsilon_1) \\ Y_2 &= m^2(Y_1, X, \epsilon_2) \end{aligned}$$

with unobservables (ϵ_1, ϵ_2) independent of instrument X. Matzkin (2015) establishes several constructive identification results on the quantity $\partial m^1(y_2, \epsilon_1)/\partial y_2$, all of which are functionals of the conditional density $f_{Y|X}$. As before, nonparametric estimators of such derivatives and be constructed based on these identification results with the estimated conditional densities.

Example 2.3 (Continuous Treatment). Hirano and Imbens (2004) introduces a generalization of the potential outcome framework to the continuous treatment case, i.e., $Y(t)$ for $t \in [t_0, t_1]$, which is referred to as the individual level "dose-response" function, and the parameter of interests is the average dose-response function $E[Y(t)]$. It is assumed that we observe an i.i.d. sample of $\{Y_i, X_i, T_i\}$, where $Y_i \equiv Y_i(T_i)$ denotes the observed potential outcome at the received treatment dose T_i, X_i is a vector of covariates, and $T_i \in [t_0, t_1]$ denotes the continuous treatment. Hirano and Imbens (2004) defines the conditional density

$f_{T|X}$ as the generalized propensity score. Under the weak unconfoundedness assumption that

$$Y(t) \perp T \mid X \quad \text{for all} \quad t \in [t_0, t_1],$$

it can be shown that the average potential outcome at $T = t$ satisfies

$$E[Y(t)] = E[E[Y|T = t, f_{T|X}(t|X)]]. \tag{1}$$

The estimation of $E[Y(t)]$ based on above expression requires the estimation of the conditional density $f_{T|X}$ as a first step. In Hirano and Imbens (2004), $f_{T|X}$ is estimated using a linear model.

Kennedy et al. (2017) proposes an alternative identification result of (1), which relies on a doubly robust signal $Y(\eta)$

$$Y(\eta) \equiv \frac{Y - E[Y|T, X]}{f_{T|X}} \int_{\mathcal{X}} f_{T|X=x} dP_X(x) + \int_{\mathcal{X}} f_{T|X=x} E[Y|T, X = x] dP_X(x) \tag{2}$$

where $\eta = (E[Y|T, X], f_{T|X})$ denotes the infinite-dimensional nuisance parameters such that

$$E[Y(t)] = E[Y(\eta)|T = t]. \tag{3}$$

In practice, the researchers would first estimate the pseudo outcome $Y(\eta)$ and then estimate the conditional expectation (3) using the estimated pseudo outcome $\hat{Y}(\hat{\eta})$, see for example, Kennedy et al. (2017) using the kernel methods and Semenova and Chernozhukov (2021) using series. A crucial step in the above procedure requires estimating the infinite-dimensional nuisance parameters including the conditional density $f_{T|X}$. Kennedy et al. (2017) assumes a model $T = \mu(X) + \sigma(X)\epsilon$ where $\epsilon|X$ has zero mean and unit variance. Then they use a suite of ML methods to estimate $\mu(X) = E[T|X]$ and $\sigma(X) = Var(T|X)$. In the final step, the conditional density, now effectively a unit variate density estimation problem, is estimated using the standard kernel method.

In related works, Kallus and Zhou (2018), Su et al. (2019), and Colangelo and Lee (2022) consider estimation (and inference in the latter two studies) of $E[Y(t)]$ using an alternative score

$$E[Y|T = t, X] + \frac{K_h(T - t)}{f_{T|X}(t)}(Y - E[Y|T = t, X]) \tag{4}$$

such that

$$E[Y(t)] = \lim_{h \to 0} E\left[E[Y|T = t, X] + \frac{K_h(T - t)}{f_{T|X}(t)}(Y - E[Y|T = t, X])\right].$$

Colangelo and Lee (2022) studies the estimator based on this expression in detail and they propose an estimator on the inverse $1/f_{T|X}(t)$ by taking the numerical derivative of the estimated conditional quantile.

Example 2.4 (**Conditional Average Partial Derivative**)**.** Let (D, Z, Y) be the observed varibles, with $D \in \mathbf{R}$ being the continuous treatment variable, $Y = Y^D$ the observed outcome, and Z a vector of controls. Let X be a subvector of Z. Semenova and Chernozhukov (2021) defines the conditional average partial derivative $\partial_d E[Y^D|X = x]$ as the parameter of interests. Under conditional independence $\{Y^d, d \in \mathbf{R}\} \perp D|Z$, Semenova and Chernozhukov (2021) shows that

$$\partial_d E[Y^D|X = x] = E[Y(\eta)|X = x] \tag{5}$$

with the signal

$$Y(\eta) := -\partial_d \log f_{D|Z}(Y - E[Y|D, Z]) + \partial_d E[Y|D, Z]$$

where $\eta = (E[Y|D, Z], f_{D|Z})$ denotes the infinite-dimensional nuisance parameters. The estimation based on (5) requires first estimating the nuisance parameters $\hat{\eta}$, particularly the conditional density $f_{D|Z}$. Semenova and Chernozhukov (2021) assumes $D = \mu(Z) + \epsilon$ with $\epsilon \perp Z$ and they estimate $\mu(Z)$ using LASSO. Then the conditional density estimation problem is reduced to a univariate density estimation problem.

Example 2.5 (**Counterfactual Distributions**)**.** The counterfactual distributions have been studied extensively in the inequality literature. For example, in the context of DiNardo et al. (1996), the parameter of interest is the counterfactual wage (Y) distribution of the non-unionized workers (group A) if their covariates/attributes had the same distribution of the unionized workers (group B). Then under the assumption of the invariance of counterfactual distributions (see Fortin el al. (2011) Assumption 6), the counterfactual density of group A can be identified as

$$f^c_{Y_A}(y) = \int f_{Y_A|X_A}(y|x) \frac{dF_{X_B}(x)}{dF_{X_A}(x)} dF_{X_A}(x) \tag{6}$$

where the ratio of densities can be estimated by

$$\frac{dF_{X_B}(X)}{dF_{X_A}(X)} = \frac{P(D_B = 1|X)}{P(D_A = 1|X)} \frac{P(D_A = 1)}{P(D_B = 1)}$$

(see Fortin el al. (2011) section 4.5-4.6 for details). A nonparametric estimator of the counterfactual density can be constructed using the expression in (6), which requires estimation of the conditional density $f_{Y_A|X_A}$.

3 Conditional Density Estimation

3.1 Series Representation

First, we state a formal result that the conditional densities admit a series expansion under fairly general conditions. We make the following assumptions

Assumption 3.1. *(i)* $\mathbf{Y}$ *and* $\mathbf{X}$ *are Polish spaces; (ii) random variables* $(Y, X) \in \mathbf{Y} \times \mathbf{X}$ *with a probability measure* μ *on Borel* σ*-algebra* $\mathcal{B} := \mathcal{B}_Y \otimes \mathcal{B}_X$*; (iii) there exist* σ*-finite measures* ν_Y *and* ν_X *on* $\mathcal{B}_Y$ *and* $\mathcal{B}_X$ *such that* $\mu \ll \nu := \nu_Y \otimes \nu_X$*.*

Assumption 3.1 (i) is fairly general and can be easily satisfied for most cases in economics. For example, economic variables Y and X typically take values in well-behaved[2] subsets $\mathbf{Y} \times \mathbf{X} \subseteq \mathbf{R} \times \mathbf{R}^d$, which together with (iii) ensures that $L^2(\nu_Y)$ is separable so that an orthonormal basis exists.

Moreover, (iii) also ensures that the Radon-Nikodym derivative exists, i.e. density $f_{Y,X}$ of μ w.r.t ν exists:

$$\int_B f_{Y,X}(y,x)d\nu(y,x) = \mu(B) \quad \text{for all} \quad B \in \mathcal{B}.$$

Then the conditional density is well-defined:

$$f_X(x) := \int_{\mathbf{Y}} f_{Y,X}(y,x)d\nu_Y(y) \quad f_{Y|X}(y|x) := \begin{cases} \frac{f_{Y,X}(y,x)}{f_X(x)} & \text{if } f_X(x) \neq 0 \\ f_{Y,X}(y,x) & \text{if } f_X(x) = 0 \end{cases}.$$

Finally, let μ_X be the projection of μ onto $\mathbf{X}$, that is, for any $B \in \mathcal{B}_X$, $\mu_X(B) = \mu(\mathbf{Y} \times B)$. Then, we have the following proposition

Proposition 3.1. *Suppose Assumption 3.1 is satisfied. Then the following results hold:*

(i) $L^2(\nu_Y)$ *is separable;*

(ii) If $f_{Y|X} \in L^2(\nu_Y \otimes \mu_X)$ *and* $\{\phi_j\}_{j=1}^{\infty}$ *is an orthonormal basis for* $L^2(\nu_Y)$*, then*

$$P(\lim_{J\to\infty} \int (f_{Y|X}(y|X) - \sum_{j=1}^{J} E[\phi_j(Y)|X]\phi_j(y))^2 d\nu_Y(y) = 0) = 1$$

[2] $\mathbf{Y}(\mathbf{X})$ can be either open/closed intervals(hypercubes) in $\mathbf{R}(\mathbf{R^d})$ or the entire real line $\mathbf{R}^d$ ($\mathbf{R}^d$).

(iii) If $f_{Y|X} \in L^2(\nu_Y \otimes \mu_X)$ *and* $\{\phi_j\}_{j=1}^{\infty}$ *is an orthonormal basis for* $L^2(\nu_Y)$*, then*

$$\lim_{J\to\infty} E[\int (f_{Y|X}(y|X) - \sum_{j=1}^{J} E[\phi_j(Y)|X]\phi_j(y))^2 d\nu_Y(y)] = 0$$

if and only if $\lim_{J\to\infty} \sum_{j=1}^{J} E[(E[\phi_j(Y)|X])^2] < \infty$.

The proposition formally states that if $f_{Y|X} \in L^2(\nu_Y \otimes \mu_X)$, the series expansion holds μ_X-a.e. (in the sense that for a.e. x, the series converges in $L^2(\nu_Y)$) as well as in $L^2(\nu_Y \otimes \mu_X)$. From now on, we will use the following representation whenever the convergence holds:

$$f_{Y|X}(y|x) = \sum_{j=1}^{\infty} E[\phi_j(Y)|X = x]\phi_j(y). \tag{7}$$

3.2 Cross-Validation

Given an i.i.d. sample $\{(Y_i, X_i)\}_{i=1}^n$, a natural estimator can be constructed by first picking a cutoff J and estimating $h_j(X) := E[\phi_j(Y)|X]$ for $j = 1, \cdots, J$, then forming

$$\hat{f}_J(y|x) = \sum_{j=1}^{J} \hat{h}_j(x)\phi_j(y). \tag{8}$$

The quality of this estimator depends on both the estimated $\hat{h}_j$'s and the series cutoff J. Given the product measure $\nu_Y \otimes \mu_X$, it is natural to define the following norm[3]

$$\|g\|_H^2 := E[\int g^2(y, X) d\nu_Y(y)]$$

for any function g of (y, x). With this norm, given $\{\hat{h}_j\}_{j=1}^J$, we can measure the distance between $\hat{f}_J$ and the true conditional density $f_{Y|X}$

$$\|\hat{f}_J - f_{Y|X}\|_H^2 = \sum_{j=1}^{J} E_X[(\hat{h}_j(X) - h_j(X))^2] + \sum_{j=J+1}^{\infty} E[h_j^2(X)]$$

which corresponds to the familiar variance-bias trade-off.

We propose a data-driven procedure of selecting the series cutoff J based on a modified cross-validation procedure introduced by Lecué and Mitchell (2012). First, we consider a

[3]This is essentially the L^2 norm w.r.t. the product measure $\nu_Y \otimes \mu_X$.

loss function $Q : (\mathbf{Y} \times \mathbf{X}, L^2(\nu_Y \otimes \mu_X)) \to \mathbf{R}$

$$Q((y,x),f) := \int f^2(y,x)d\nu_Y(y) - 2f(y,x). \tag{9}$$

Then for any function $f \in L^2(\nu_Y \otimes \mu_X)$, the risk of f under this loss Q takes the form

$$\begin{aligned} R(f) :=& E[Q((Y,X),f)] \\ =& E[\int f^2(y,X)d\nu_Y(y) - 2f(Y,X)] \\ =& E[\int (f(y,X) - f_{Y|X}(y)^2 d\nu_Y(y) - \int f^2_{Y|X}(y)d\nu_Y(y) \\ &+ 2\int f(y,X)f_{Y|X}(y)d\nu_Y(y) - 2f(Y,X)] \\ =& \|f - f_{Y|X}\|_H^2 - \|f_{Y|X}\|_H^2 \end{aligned} \tag{10}$$

where the last equality holds by law of iterated expectation and the definition of the norm $\|\cdot\|_H$. In particular, the risk is minimized at the true conditional density $f_{Y|X}$. This suggests that this loss Q and the associated risk R are suitable for cross-validation type of procedures.

We now introduce some notations and formally define our estimator[4]. Let n denote the sample size and without loss of generality suppose n is divisible by some fixed integer K. Then we randomly split the sample $D^{(n)} := \{(Y_i, X_i\}_{i=1}^n$ into K validating sets $D^{(n_V)}$ of equal size $n_V := n/K$. For each of these validating sets, use the remaining data $D^{(n_T)} := D^{(n)} \setminus D^{(n_V)}$ of size $n_T := n - n_V$ as the training set. Let $\{\hat{f}_1, \cdots, \hat{f}_p\}$ be a set of statistics such that its j-th element is $\hat{f}_j = \sum_{k=1}^j \hat{h}_k \phi_k$ (recall $\hat{h}_k := \hat{E}[\phi_k(Y)|X]$). Note that we define a statistic $\hat{f} = (\hat{f}^{(m)})_{m \in \mathbf{N}}$ as a sequence such that $\hat{f}^{(m)}$ is an estimator trained with sample $D^{(m)}$. Then the K-fold empirical risk of $\hat{f} \in \{\hat{f}_j\}_{j=1}^p$ is defined as

$$R_{n,K}(\hat{f}) := \frac{1}{K}\sum_{k=1}^K \frac{1}{n_V} \sum_{i \in D_k^{(n_V)}} Q((Y_i, X_i), \hat{f}^{(n_T)}(D_k^{(n_T)})) \tag{11}$$

where we use $D_k^{(n_V)}$ to denote the k-th validating sample and $D_k^{(n_T)} = D^{(n)} \setminus D_k^{(n_V)}$ to denote the corresponding training sample. The estimator we study takes the form

$$\bar{f}^{(n)} := \frac{1}{K}\sum_{k=1}^K \hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)}) \quad \text{with} \quad \hat{j}^* = \arg\min_{1 \le j \le p} R_{n,K}(\hat{f}_j). \tag{12}$$

[4]The notations are largely borrowed from Lecué and Mitchell (2012).

That is, after finding the $\hat{j}^*$ that minimizes the K-fold empirical risk, we average estimators $hat f_{\hat{j}^*}$ trained using sub-samples to obtain the final estimator $\bar{f}^{(n)}$. Although this estimator aggregates sub-sample estimators, it still a series estimator: $\bar{f}^{(n)}(y|x) = \sum_{j=1}^{\hat{j}^*} \tilde{h}_j(x)\phi_j(y)$ with $\tilde{h}_j := K^{-1}\sum_{k=1}^{K} \hat{h}_j(D_k^{(n_T)})$. That is, we first use CV procedure to select $\hat{j}^*$, and then we define a new estimator for each conditional expectation h_j by using the average of sub-sample $\hat{h}_j$'s. Note that this estimator differs from the typical V-fold CV estimator $\hat{f}_{VCV} := \hat{f}_{\hat{j}^*}^{(n)}$ that is trained by using the full sample $D^{(n)}$ after finding the $\hat{j}^*$ above. While we do not compare[5] between $\bar{f}^{(n)}$ and $\hat{f}_{VCV}$, we want to emphasize that $\bar{f}^{(n)}$ still effectively uses the full sample to train the estimator and does not require re-training after selecting $\hat{j}^*$.

One potential concern is that the empirical risk $R_{n,V}$ takes the form of an empirical average of loss Q, which requires integral calculations. However, the orthonormality of the basis allows us to avoid such integral calculations altogether. Note that the estimators we consider take the form $\hat{f}_j = \sum_{k=1}^{j} \hat{h}_k\phi_k$ with ϕ_k's being elements in an orthonormal basis. Then using orthonormality, the loss can be rewritten as $Q((y,x),\hat{f}_j) = \sum_{k=1}^{j}\hat{h}_k^2(x) - 2\hat{f}_j(y,x)$, which only requires calculating simple summations when computing the empirical risk.

Another potential issue is that the estimator may not be a proper conditional density, i.e. $\int \bar{f}(y|x)d\nu_Y(y)$ may not equal to 1 and the estimator may be negative. The former is easy to solve. If we assume the orthonormal basis $\{\phi_j\}$ of $L^2(\nu_Y)$ contains a constant term, without loss of generality, say ϕ_1, then $\int \phi_j(y)d\nu_Y(y) = \mathbf{1}\{j=1\}$, which implies that $\int \bar{f}(y|x)d\nu_Y(y) = 1$ always. To address the second potential problem, we consider the following set

$$C := \{c \in \ell^2 : \sum_{j=2}^{\infty} c_j\phi_j(y) \geq -\phi_1\}.$$

Let $\hat{h}_j = \hat{E}[\phi_j(Y)|X]$, and for any x, we consider the projection of $\{\hat{h}_j(x)\}_{j=2}^{\infty}$ onto C:

$$\{\tilde{h}_j(x)\}_{j=2}^{\infty} = \arg\min_{c\in C} \|\hat{h}(x) - c\|_{\ell^2}$$

which can be implemented either on the final estimator $\bar{f}$ or on the sub-sample estimators $\hat{f}_{\hat{j}^*}$. Moreover, C is convex, and any convex projection in Hilbert spaces are non-expansive so that this post-processing actually weakly reduces the estimation error. One simple method for such projection is the *p-algorithm* proposed by Gajek (1986). Therefore, our main results will be established for the pre-processed estimators, and researchers can decide what

[5] However, as commented in Lecué and Mitchell (2012), with additional regularity conditions, the estimation error of $\hat{f}_{VCV}$ can be bounded using the sub-sample estimator.

post-processing methods to use if they suspect the estimator might be negative.

3.3 Theoretical Results

We first establish an oracle inequality for our estimator and the proof follows from the general strategy laid out in Lecué and Mitchell (2012) with some modifications, which we defer to the appendix.

Theorem 3.1. *Suppose random sample* $\{(Y_i, X_i)\}_{i=1}^n \sim^{i.i.d.} (Y, X)$ *with* $(Y, X) \in (\mathbf{Y} \times \mathbf{X}, \mathcal{B}_Y \otimes \mathcal{B}_X, \nu_Y \otimes \mu_X)$. *Assume the conditional density* $f_{Y|X} \in L^2(\nu_Y \otimes \mu_X)$ *and the statistiscs* $\{\hat{f}_j\}_{j=1}^p$ *defined as in (8) are bounded above by some constant* M. *Let* $\bar{f}$ *be the estimator defined in (12). Then for any constant* $a > 0$, *there exists a constant* C *that only depends on* a *such that*

$$E[\|\bar{f}^{(n)} - f_{Y|X}\|_H^2] \leq (1+a) \min_{1 \leq j \leq p} E[\|\hat{f}_j^{(n_T)} - f_{Y|X}\|_H^2] + C\frac{\log p}{n_V}. \tag{13}$$

This theorem establishes an oracle inequality under very few assumptions. In fact, the main assumption in the theorem we rely on is that the true conditional density $f_{Y|X}$ and the dictionary of estimator $\{\hat{f}_j\}_{j=1}^p$ are uniformly bounded above by some constant. We can even modify the theorem to allow for this bound grow with p.[6] In the background, however, the convexity of the loss Q and the associated risk R defined in section 2 plays a major role. In particular, the convexity of the risk allows us to bound the expected difference $R(\bar{f}^{(n)}) - R(f_{Y|X})$ by two terms, one of which being the oracle and the other being a shifted empirical process. The shifted empirical process is then controlled by a maximal inequality modified from Lecué and Mitchell (2012) to suit our estimators, which gives rise to the $\log(p)/n$ term.

This oracle inequality essentially states that, the estimation error of our estimator $\bar{f}$ is bounded above (up to a constant) by the smallest achievable estimation error for a given dictionary of estimators $\{\hat{f}_j\}_{j=1}^p$. To obtain a concrete estimation error, additional structures on our estimator and on the true conditional density $f_{Y|X}$ are needed. Note that the performance of any estimator $\hat{f}_J$ under norm $\|\cdot\|_H$ can be examined through the variance-bias decomposition

$$E[\|\hat{f}_J - f_{Y|X}\|_H^2] = \sum_{j=1}^{J} E[(\hat{h}_j(X) - h_j(X))^2] + \sum_{j=J+1}^{\infty} E[h_j^2(X)]$$

[6] In the proof, we kept the bound M explicit throughout the proof and one can make assumptions on how fast M grows with p and obtain different bounds on the shifted empirical process.

which suggests that this estimation error should be minimized at some J under suitable regularity conditions. Moreover, as long as K (as in K-fold cross-validation) is fixed, n_T and n_V will be of the same order as the sample size n. Hence, for sufficiently large p, we should be able to establish the estimation error for our estimator by solving the minimization problem in the first term on the right-hand side of (13). In the next theorem, we achieve this goal under one possible set of regularity conditions.

Theorem 3.2. *Suppose conditions in Theorem 3.1 are satisfied. Moreover, assume that*

(i) for some constant $0 < \delta \leq 1$, $E[(\hat{h}_j(X) - h_j(X))^2] \asymp n^{-\delta}$ *for all* $j \geq 1$;

(ii) for some constant $\gamma > 0$, $\sum_{j=J+1}^{\infty} E[h_j^2(X)] \lesssim J^{-\gamma}$ *for all* $J \geq 0$.

Then, for $p \gtrsim n^{\delta/(\gamma+1)}$,

$$E[\|\bar{f} - f_{Y|X}\|_H^2] = O(n^{-\frac{\gamma}{\gamma+1}\delta} \vee \frac{\log p}{n}).$$

The condition (i) in Theorem 3.2 makes an assumption on the quality of the estimated conditional expectations $\hat{h}_j(X) = \hat{E}[\phi_j(Y)|X]$. In general, without further assumptions, e.g. linearity or sparsity, we should expect δ to be small for nonparametric estimators and high dimensional X. A growing literature in statistics and machine learning is actively studying the estimation error of various state-of-the-art machine learning estimators. For example, Chen et al. (2022) establishes the estimation error in the form of condition (i) for the deep ReLU neural networks for Hölder classes embedded in high-dimensional spaces. Similarly, Suzuki (2018) and Hayakawa and Suzuki (2020) establish estimation errors of deep neural networks for other function classes. See also section 4 in Izbicki and Lee (2017) for several other examples that satisfy (i). A common feature of these ML estimators, especially the deep neural networks, is that they are adaptive to intrinsic low-dimensional variables in high-dimensional settings, which is particularly appealing in our setting.

On the other hand, condition (ii) controls the rate of decay of the tail sum of the series and hence the bias. In particular, as shown in Proposition 3.1 (iii), the existence of the series expansion of the conditional density $f_{Y|X}$ requires that the tail sum satisfies $\lim_{J\to\infty} \sum_{j=J+1}^{\infty} E[h_j^2(X)] = 0$. In the context of the regression and density estimation, condition (ii) is closely related to the *full approximation set* discussed in Lorentz (1966) and Yang and Barron (1999), and such assumptions place restrictions on the smoothness of the function classes under consideration. For comparison, in the context of full approximation set, see Yang and Barron (1999), with $\delta = 1$ and $\gamma = 2\alpha$, we obtain the minimax rate $n^{-2\alpha/(2\alpha+1)}$. In general, however, it is difficult to compare our results to the minimax

optimal nonparametric estimation rates in $\mathbf{R}^{d+1}$ (eg. the minimax rate $n^{2\alpha/(2\alpha+d+1)}$ in Stone (1982)): in addition to the nonparametric regression problem $E[\phi_j(Y)|X]$ in $\mathbf{R}^d$, we also have the additional structure on how fast $E[(E[\phi_j(Y)|X])^2]$ decays with j.

We want to emphasize three appeals of our estimator. First, comparing to the existing methods, our estimator is relatively simple to construct. In its core, it is a series estimator in which the researchers can use the growing variety of ML estimators to estimate each term. The second appeal of our estimator is that it is practical in the setting where the conditioning variable X is high-dimensional. When the conditions are satisfied for fast convergence of ML estimators $\hat{h}_j$ in high-dimensional setting (e.g. sparsity or approximate sparsity), our estimator achieves fast rate of convergence. Last but not the least, our estimator is data-driven. The optimal series cutoff are selected by a data-driven cross validation type of procedure, which does not rely on the smoothness assumptions on the true conditional densities.

In some applications, the researchers may be interested in the conditional density at a point, i.e. $f_{Y|X}(y|X)$ at a specific y. The next result shows that the MISE rate in Theorem 3.2 can be achieved in this point-wise case under the proposed conditions.

Theorem 3.3. *Suppose conditions in Theorem 3.2 are satisfied. Moreover, assume that*

(i) the orthonormal basis is uniformly bounded;

(ii) for every $J \leq p$, $\overline{EIG}(\Sigma_J)/\underline{EIG}(\Sigma_J) = O(1)$, where $\overline{EIG}(\Sigma_J)$ and $\underline{EIG}(\Sigma_J)$ denote the largest and smallest eigenvalues of Σ_J respectively and $\Sigma_J \equiv E[B_J(X)B_J(X)']$ with $B_J(X)$ being the column vector $B_J(X) \equiv (E[\phi_j(Y)|X] - \hat{E}[\phi_j(Y)|X])_{j=1}^J$.

Then, for $p \gtrsim n^{\delta/(\gamma+1)}$,

$$\sup_{y \in \mathbf{Y}} E[\|\bar{f}^{(n)}(y) - f_{Y|X}(y)\|^2_{P_X,2}] = O(n^{-\frac{\gamma}{\gamma+1}\delta} \vee \frac{\log p}{n}).$$

The condition (i) is needed so that the bias can be bounded in the same way as the integrated case, which can be satisfied using trigonometric bases on closed intervals in $\mathbf{R}$ or Hermite basis on the whole $\mathbf{R}$ for example. This condition can be relaxed to allow for an unbounded basis potentially at the cost of a slower rate. Condition (ii) is a high-level assumption, which is determined by the quality of the estimators $\hat{E}[\phi_j(Y)|X]$'s. The diagonal entries of the matrix Σ_J measure the "variances" of each conditional mean estimators in the series, while the off-diagonal measure the cross-term correlations. Note that in the integrated case there is no such correlation due to the orthonormality of ϕ_j's.

Remark 3.1. *A possible modification of our estimator in the point-wise case is to consider estimators of the form*

$$\check{f}_J(y|X) := \frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{j}\hat{E}[\phi_k(Y)|X]\phi_k(y).$$

The rationale of such modification is that the modified series takes the form of a Cesàro sum, which is known to have many good convergence properties in the context of Fourier analysis. For example, when $\{\phi_j\}$ is the cosine/trigonometric basis, $\check{f}_J$ resembles a kernel estimator (using the so-called Fejér kernel and estimated as $\hat{E}[J^{-1}\sum_{j=1}^{J}\sum_{k=1}^{j}\phi_k(Y)\phi_k(y)|X]$). Although such estimators have been considered in the context of density estimator (e.g. Kronmal and Tarter (1968)), they have yet to be studied in the conditional density case and further research is needed.

Remark 3.2. *So far we have assumed Y is low-dimensional. In the case when $Y = (Y_1, \cdots, Y_G)$, the same techniques we discussed above can be applied using an orthonormal basis on $\mathbf{Y} \subseteq \mathbf{R}^G$ via a tensor product of one-dimensional orthonormal bases. The number of the basis terms formed through such tensor product grows quickly with G and can become intractable for large G. One can consider an alternative approach that relies on a fundamental property in probability theory, namely, the chain rule:*

$$\begin{aligned}
&f(Y_1, \cdots, Y_G|X_1, \cdots, X_K)\\
&= f(Y_1|Y_2, \cdots, Y_G, X_1, \cdots, X_k)\\
&\times f(Y_2|Y_3, \cdots, Y_G, X_1, \cdots, X_k)\\
&\cdots\\
&\times f(Y_G|X_1, \cdots, X_k).
\end{aligned}$$

Then using this expression, instead of having to deal with potentially large number of tensor products of orthonormal bases, we can apply our results on each term in the product and form the final estimator accordingly. A rigorous study of such estimator is left for future research.

In the next section, we provide a detailed example in the context of nonparametric difference-in-differences models. In this example, not only does the conditional density play a key role in identifying the parameter of interest, its series representation also guides the estimation and inference procedures.

4 Double Debiased Continuous Difference in Differences

Difference-in-Differences (DiD) is one of the most popular research designs in empirical research. While the more common DiD designs deal with binary or discrete multi-valued treatments, there has been an increasing amount of interests in cases where the treatments are continuous, see for example Callaway et al. (2021); D'Haultfoeuille et al. (2021); de Chaisemartin et al. (2022). In particular, Callaway et al. (2021) examine what the popular two-way fixed effect (TWFE) regressions can identify in the DiD setting with a continuous treatment. On the other hand, D'Haultfoeuille et al. (2021) generalizes the change-in-changes models studied in Athey and Imbens (2006) to the case of continuous treatment. In contrast to the aforementioned literature, our results generalize Abadie (2005) to the settings with continuous treatments. Specifically, we focus on the average treatment effect on the treated (ATT) at any given treatment intensity under a conditional parallel trend assumption that allows for covariates.

One way these covariates enter the idenfication result is via the conditional density of the continuous treatment conditioning on these covariates[7]. Therefore, in our setting, the low-dimensional target parameter (ATT) depends, among others, on the infinite dimensional conditional density. This motivates us to consider the estimation and inference of the ATT's under the framework of double/debiased machine learning (DML) studied in CCDDHNR (2018). Similar to Chang (2020) that studies the DiD with discrete treatments, we derive scores that enjoy the *Neyman orthogonality condition* for DiD with continuous treatment in both repeated outcomes (panel data) and repeated cross section settings. Nevertheless, we want to emphasize that our results require non-trivial modifications over those in the aforementioned literature. We illustrate the potential usefulness of our method using the data from Ashraf et al. (2020) and revisit the treatment effect on education of a large policy intervention in Indonesia studied in Duflo (2001).

4.1 Setup and Identification

In this section, we formally set up the difference-in-differences models with continuous treatment following Abadie (2005). First, using the potential outcome notation (e.g. Rubin (1974), Heckman (1990)), let $Y_{i,t}(0)$ denote the potential outcome of individual i in period t when receiving no treatment, and similarly let $Y_{i,t}(d)$ denote the potential outcome of individual i in period t when receiving treatment with intensity d.

[7]Such conditional density is commonly referred to as the "genearlized propensity score", see Hirano and Imbens (2004).

The treatment variable D is modeled as a random variable with a mixture distribution: a probability mass at 0 and a continuous distribution on an interval $[d_L, d_H]$ excluding 0. To formalize this mixture distribution, consider a measure $\mu = \delta_0 + \lambda$, with λ being the Lebesgue measure and δ_0 being the Dirac delta at 0. Suppose $D \sim F_D$, then we have $dF_D/d\mu := \mathbf{1}\{D = 0\}P(D = 0) + \mathbf{1}\{D > 0\}f_D$ with f_D being the probability density on $[d_L, d_H]$. In particular, $F_D(0) = \int \mathbf{1}\{D = 0\}\frac{dF}{d\mu}d\mu = P(D = 0)$ and for any measurable $A \in \mathcal{B}$ such that $0 \notin A$, $F_D(D \in A) = \int_A f_D \, d\lambda$. We also assume similar structures hold in the conditional cases.

We restrict our attention to the two-period $(t-1, t)$ models, and as in the typical DiD settings, no subject receives treatment at period 0, so we may supress the time notation in treatment D_i. Let X_i denote the set of individual level control variables. We consider the following set of assumptions:

Assumption 4.1 (Repeated Outcomes)**.** *The observed data $\{Y_{i,t-1}, Y_{i,t}, D_i, X_i\}_{i=1}^n$ are independently and identically distributed.*

Assumption 4.2 (Repeated Cross Sections)**.**

(i) For each individual i in the pooled sample, the researcher observe $\{Y_i, D_i, X_i, T_i\}$, where T_i is a time indicator $= 1$ if obervation i belongs to the post-treatment sample and $= 0$ otherwise, and $Y_i = (1 - T_i)Y_{i,t-1} + T_i Y_{i,t}$;

(ii) Conditional on $T = 0$, data are i.i.d. from the distribution of (Y_{t-1}, D, X); Conditional on $T = 1$, data are i.i.d. from the distribution of (Y_t, D, X).

Assumption 4.3 (Support)**.**

(i) No subject receives treatment in the pre-treatment period;

(ii) the support of treatment intensity D satisfies $supp(D) = \{0\} \cup [d_L, d_H]$ with $0 < d_L < d_H \leq \infty$;

(iii) $P(D = 0|X) > 0$ almost surely;

(iv) $1 > P(D = 0) > 0$ and D admits a strictly positive probability density f_D on (d_L, d_H).

Assumption 4.4 (Conditional Parallel Trend)**.** *For all $d \in [d_L, d_H]$, the following holds*

$$E[Y_t(0) - Y_{t-1}(0)|X, D = d] = E[Y_t(0) - Y_{t-1}(0)|X, D = 0].$$

Remark 4.1. *Assumptions 4.1 and 4.2 are standard in the DiD literature. We note that, in Assumption 4.1, while allowing for covariates, these covariates cannot vary across time. Moreover, Assumption 4.2(ii) requires that the sample is not stratified by the outcome, treatment, or covariates.*[8] *Assumption 4.3 describes the requirements on the support that the treatment must satisfy. In particular, in the continuous DiD setting, the control group ($D = 0$) must have a positive measure, and the treated group must have positive likelihood of being treated at any level $d \in (d_L, d_H)$. Finally, Assumption 4.4 is the conditional parallel trend condition that generalizes the discrete cases, which is the main identifying assumption.*

Next, we describe our target parameter. The parameter we are interested in is the average treatment effect on the treated at any given treatment intensity $d \in (d_L, d_H)$:

$$ATT(d) \equiv E[Y_t(d) - Y_t(0)|D = d]. \tag{14}$$

The interpretation of this parameter is analogous to the cases with discrete treatment: the expected treatment effect of a treatment with intensity d given the subjects are treated with intensity d. Note that ATT is a local measure, and in the absence of stronger assumptions, the average treatment effect $ATE(d) \equiv E[Y_t(d) - Y_t(0)]$ is not identified. The following two theorems are the main results of this section, in which we establish the identifications of $ATT(d)$ for both repeated outcomes and repeated cross sections settings.

Theorem 4.1 (Repeated Outcomes)**.** *If Assumptions 4.1, 4.3, and 4.4 hold, for any $d \in (d_L, d_H)$,*

$$ATT(d) = E[Y_t - Y_{t-1}|D = d] - E[(Y_t - Y_{t-1})\mathbf{1}\{D = 0\}\frac{f_{D|X}(d)}{f_D(d)P(D = 0|X)}]$$

Theorem 4.2 (Repeated Cross Sections)**.** *If Assumptions 4.2, 4.3, and 4.4 hold, for any $d \in (d_L, d_H)$,*

$$ATT(d) = E[\frac{T - \lambda}{\lambda(1 - \lambda)}Y|D = d] - E[\frac{T - \lambda}{\lambda(1 - \lambda)}Y\mathbf{1}\{D = 0\}\frac{f_{D|X}(d)}{f_D(d)P(D = 0|X)}]$$

where $\lambda \equiv P(T_i = 1)$.

The proofs are given in the appendix. The main identifying assumption is the conditional parallel trend that allows us to substitute the unobserved counterfactual trend $E[Y_t(0) - Y_{t-1}(0)|X, D = d]$ by the observed trend $E[Y_t(0) - Y_{t-1}(0)|X, D = 0]$ of the control group.

[8] However, as pointed out in Abadie (2005), in the case of stratified sampling, reweighting methods can be applied to establish similar results.

While we derive our results under this parallel trend assumption, it is possible to consider alternative parallel trend assumptions that can be used to identify other causal parameters of interests (see Callaway et al. (2021) Section 3.3 for example).

With Theorem 4.1 and 4.2, one can build estimators for $ATT(d)$ using the estimated sample analogues. However, with the control variables X potentially being high-dimensional, the nonparametric estimations of $f_{D|X}(d)$ and $P(D=0|X)$ likely require regularization procedures that introduce non-trivial first order biases (see CCDDHNR (2018) and references therein for a detailed discussion). One way to alleviate such regularization biases is to construct scores that satisfies the so-called *Neyman orthogonality* property. In the next section, we study how to construct orthogonal scores for our target parameters.

4.2 Orthogonal Scores

First we formally define the Neyman orthogonality following CCDDHNR (2018) and Chang (2020). In particular, we consider the scores that satisfy the orthogonality only with respect to the infinite dimensional nuisance parameter. Let $\theta_0 \in \Theta \subset \mathbf{R}$ be the low dimensional parameter of interest, which corresponds to $ATT(d)$ in our setting. Let ρ_0 denote the true low dimensional nuisance parameter(s). For example, $\rho_0 = (P(T=1), f_D(d))$ for a given d in the repeated cross-sectional case. Let $\eta_0 \in \mathcal{T}$ denote the true infinite dimensional nuisance parameters, which in our case include $f_{D|X}(d|X)$, $P(D=0|X)$, and other nuisance parameters created when constructing the orthogonal scores. Let the observable random vector $Z \in (\mathbf{Z}, \mathcal{B}_Z, P)$, which corresponds to $Z = (Y, X, D, T)$ in our example. Let $\psi(Z, \theta_0, \rho_0, \eta_0) \in \mathbf{R}$ be a score such that $E_P[\psi(Z, \theta_0, \rho_0, \eta_0)] = 0$.

The *Gateaux* (directional) derivative with respect to the infinite dimensional nuisance parameter is defined as, for any $r \in [0, 1)$ and $\eta \in \mathcal{T}$,

$$\partial_r E_P[\psi(Z, \theta_0, \rho_0, \eta_0 + r(\eta - \eta_0))].$$

Moreover, let $\mathcal{T}_n \subset \mathcal{T}$ be a nuisance realization set in which the estimated infinite dimensional nuisance parameters $\hat{\eta}$ takes values with high probability. With these notations, we formally define the Neyman orthogonality.

Definition 4.1 (Neyman Orthogonality)**.** *A score ψ satisfies the Neyman orthogonality at $(\theta_0, \rho_0, \eta_0)$ with respect to a nuisance realization set $\mathcal{T}_n \subset \mathcal{T}$ if*

(i) the score satisfy the moment condition $E_P[\psi(Z, \theta_0, \rho_0, \eta_0)] = 0$;

(ii) for all $r \in [0,1)$ *and* $\eta \in \mathcal{T}_n$,

$$\partial_r E_P[\psi(Z, \theta_0, \rho_0, \eta_0 + r(\eta - \eta_0))]|_{r=0} = 0.$$

The first requirement in the definition says that the score identifies θ_0 while (ii) ensures that the first order bias from estimating these nuisance parameters is zero. We will construct scores that satisfy this orthogonality condition with some modifications that we will clarify shortly. We use the repeated outcomes case as our main example for illustration. The discussion on repeated cross sectional case will be deferred to the supplementary material since it only requires minor modifications.

Recall that in the repeated outcomes case, let $\Delta Y \equiv Y_t - Y_{t-1}$,

$$ATT(d) = E[\Delta Y | D = d] - \underbrace{E[\Delta Y \mathbf{1}\{D = 0\} \frac{f_{D|X}(d)}{f_D(d) P(D = 0|X)}]}_{:=\varphi},$$

which has two important features. First, the nonparametric estimations of $E[\Delta Y|D = d]$ and density $f_D(d)$ make the root-N consistency impossible. This appears to be a common feature in the literature that involves continuous treatment variables, see for example, Kennedy et al. (2017), Semenova and Chernozhukov (2021), and Colangelo and Lee (2022). Second, one can verify that the score φ does not satisfy Neyman orthogonality, and an adjustment term should be added to φ to construct a new score. In general, the adjustment term is straightforward to construct if the nuisance parameters can be written as conditional expectations. However, in our case, the two infinite dimensional nuisance parameters are $f_{D|X}(d|X)$ and $P(D = 0|X)$. While $P(D = 0|X) = E[\mathbf{1}\{D = 0\}|X]$ can be expressed as a conditional expectation, $f_{D|X}(d|X)$ being the conditional density presents additional challenges.

To address this issue, we use a modified series representation of the conditional density introduced in Section 3 so that we can approximate the conditional density using a finite series of conditional expectations. Let $\{\phi_j\}_{j=1}^{\infty}$ be an orthonormal basis, and for a strictly positive $d \in (d_L, d_H)$, we can represent $f_{D|X}$ as

$$f_{D|X}(d) = \sum_{j=1}^{\infty} E[\phi_j(D)\mathbf{1}\{D > 0\}|X]\phi_j(d).$$

Then, under suitable regularity conditions,[9]

$$E[\Delta Y \mathbf{1}\{D=0\}\frac{f_{D|X}(d)}{f_D(d)P(D=0|X)}] = \lim_{J\to\infty} E[\underbrace{\Delta Y \mathbf{1}\{D=0\}\frac{f_J(d|X)}{f_D(d)P(D=0|X)}}_{\equiv\varphi_J}]$$

where $f_J(d|X) \equiv \sum_{j=1}^J E[\phi_j(D)\mathbf{1}\{D>0\}|X]\phi_j(d)$. This expression suggests that we can construct an orthogonal score for each fixed J instead. Let $\theta_{0,J} = E[\varphi_J]$ so that the true θ_0 satisfies $\theta_0 = \lim_{J\to\infty}\theta_{0,J}$ (and for simplicity, we use the same notation for the repeated cross sections case). We will work with a fixed J for the remainder of this section and we will discuss the effect on the asymptotic distributions of letting J grow with sample size in the next section.

To simplify the expressions, denote: $m_J^d(D) \equiv \sum_{j=1}^J \phi_j(D)\phi_j(d)\mathbf{1}\{D>0\}$; $g(X) \equiv P(D=0|X)$; $\mathcal{E}_{\Delta Y}(X) \equiv E[\Delta Y\mathbf{1}\{D=0\}|X]$; $\mathcal{E}_{\lambda Y}(X) \equiv E[\frac{T-\lambda}{\lambda(1-\lambda)}Y\mathbf{1}\{D=0\}|X]$ with $\lambda = P(T=1)$; $f_d \equiv f_D(d)$. The following lemma introduce scores that satisfy Neyman orthogonality.

Lemma 4.1. *The scores (15) and (16) satisfy Neyman orthogonality defined in (4.1), where*

(i) for the repeated outcomes setting,

$$\begin{aligned}\psi_J^1 \equiv \Delta Y\mathbf{1}\{D=0\}\frac{f_J(d|X)}{f_d\cdot g(X)} - \theta_{0,J} \\ + \frac{m_J^d(D)g(X) - \mathbf{1}\{D=0\}f_J(d|X)}{f_d\cdot g^2(X)}\mathcal{E}_{\Delta Y}(X);\end{aligned} \tag{15}$$

(ii) for the repeated cross sections setting,

$$\begin{aligned}\psi_J^2 \equiv \frac{T-\lambda}{\lambda(1-\lambda)}Y\mathbf{1}\{D=0\}\frac{f_J(d|X)}{f_d\cdot g(X)} - \theta_{0,J} \\ + \frac{m_J^d(D)g(X) - \mathbf{1}\{D=0\}f_J(d|X)}{f_d\cdot g^2(X)}\mathcal{E}_{\lambda Y}(X).\end{aligned} \tag{16}$$

The proof is given in the appendix in which we explain the construction of the adjustment term and verify the Neyman orthogonality conditions given in Definition 4.1. We note that in these new scores, the infinite dimensional nuisance parameters are $f_J(d|X)$, $g(X)$, $\mathcal{E}_{\Delta Y}(X)$, and $\mathcal{E}_{\lambda Y}(X)$, with the latter two being the new ones created when constructing the adjustment terms.

[9] For example, if we assume boundedness of the ΔY, f_D and $f_{D|X}$, we can apply bounded convergence theorem to establish this result.

4.3 Estimation and Inference

In this section, we focus our discussion on the repeated outcomes case and we provide the results for the repeated cross sections in the supplementary material with minor modifications. First, we construct an estimator using the orthogonal score (15) from previous section. In particular, we adopt the cross-fitting techniques considered in CCDDHNR (2018), which allow us to prove the asymptotic results without having to verify Donsker conditions.

Algorithm 4.1 (CDID Estimator, Repeated Outcomes)**.** *Let $\{I_k\}_{k=1}^K$ denote a random partition of a randome sample $\{(Y_{i,t-1}, Y_{i,t}, D_i, X_i\}_{i=1}^N$, each with equal size $n = N/K$, and for each $k \in \{1, \cdots, K\}$, let $I_k^c \equiv N \setminus I_k$ denote the complement.*

- *Step 1: for each k, construct*

$$\widehat{ATT(d)}_k \equiv \frac{1}{n}\sum_{i \in I_k} \hat{\mathcal{E}}^d_{\Delta Y,k} - \Delta Y_i \mathbf{1}\{D_i = 0\}\frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k} \cdot \hat{g}_k(X_i)} - \frac{m_J^d(D_i)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k} \cdot \hat{g}_k^2(X_i)}\hat{\mathcal{E}}_{\Delta Y,k}(X_i)$$

  where $\hat{f}_{d,k}, \hat{\mathcal{E}}^d_{\Delta Y,k}, \hat{f}_{J,k}, \hat{g}_k, \hat{\mathcal{E}}_{\Delta Y,k}$ are the estimators of $f_d, E[\Delta Y|D = d], f_J(d|X), g(X)$ and $\mathcal{E}_{\Delta Y}(X)$ respectively using the rest of the sample I_k^c.

- *Step 2: average through k to obtain the final estimator*

$$\widehat{ATT(d)} \equiv \frac{1}{K}\sum_{k=1}^K \widehat{ATT(d)}_k.$$

Next, we state the regularity conditions that allow us to prove the asymptotic normality of our estimator for the repeated outcomes case. The corresponding conditions for the repeated cross sections are provided in the supplementary material.

Assumption 4.5 (Bounds)**.**

(i) $m_J^d(D) = \sum_{j=1}^J \phi_j(D)\phi_j(d)\mathbf{1}\{D > 0\}$ *are bounded by some constant M_J that grows with J such that $M_J/\sqrt{N} = o(1)$;*

(ii) for some constants $0 < c < 1$ and $0 < C < \infty$, $f_d > c$, $|E[\Delta Y|D = d]| < C$, and $|\mathcal{E}_{\Delta Y}(X)| < C$ almost surely;

(iii) for some constants $0 < \kappa < \frac{1}{2}$ and for all $J \geq 1$, $\kappa < f_J(d|X), g(X) < 1 - \kappa$ almost surely;

(iv) f_d *and* $E[\Delta Y|D=d]$ *are twice continuously differentiable at* $D=d\in(d_L,d_H)$ *with bounded second derivatives.*

Assumption 4.6 (Kernel)**.** *The kernel* K *satisfies:*

(i) K *is bounded and differentiable;*

(ii) $\int K(u)du=1$, $\int uK(u)du=0$, $0<\int u^2K(u)du<\infty$.

Assumption 4.7 (Rates)**.**

(i) kernel bandwidth satisfies $Nh\to\infty$ *and* $\sqrt{Nh^5}=o(1)$*and*

$$\frac{\sqrt{N}}{\max\{M_J,h^{-\frac{1}{2}}\}}E[|\sum_{j=J+1}^{\infty}E[\phi_j(D)\mathbf{1}\{D>0\}|X]\phi_j(d)|]=o(1);$$

(ii) with probability tending to 1, $\|\hat{f}_J-f_J(d|X)\|_{P,2}\leq M_J\varepsilon_N$, $\|\hat{g}(X)-g(X)\|_{P,2}\leq\varepsilon_N$, $\|\hat{\mathcal{E}}_{\Delta Y}(X)-\mathcal{E}_{\Delta Y}(X)\|_{P,2}\leq\varepsilon_N$;

(iii) with probability tending to 1, $\|\hat{\mathcal{E}}_{\Delta Y}(X)\|_{P,\infty}<C$, $\kappa<\|\hat{f}_J(X)\|_{P,\infty}<1-\kappa$, *and* $\kappa<\|\hat{g}(X)\|_{P,\infty}<1-\kappa$.

Assumption 4.5 is a mild one: (i) is stated in very general terms and can usually be verified by the choice of orthonormal basis, and in fact, $m_J^d(D)$ is a special case of a *delta sequence* that approximates the dirac delta at $D=d$, see Walter and Blum (1979) for an introduction. Given its well-established theoretical properties, we use kernel to estimate the low-dimensional parameters $f_D(d)$ and $E[\Delta Y|D=d]$, and assumption 4.6 is standard for kernel estimators. Assumption 4.7 concerns the quality of the nonparametric estimators: (i) requires the tuning parameters to undersmooth so that the bias vanishes asymptotically (otherwise asymptotic normality still holds but centered at $\theta_{0,J}$); (ii) is the standard rates conditions in the double/debiased ML literature. We remark that while $N^{-1/4}$ rate are needed for some nuisance estimators, the conditional density $\hat{f}_J$ can converge at a slower rate of $M_JN^{-1/4}$. This does not contradict the existing literature, as in the continuous treatment setting the nonparametric estimators for $ATT(d)$ can not achieve $\sqrt{N}$ rate.

Theorem 4.3 (Repeated Outcomes)**.** *Suppose assumptions 4.1, 4.3, 4.4, 4.5, 4.6, and 4.7 hold. Then for* $\varepsilon_N=o(N^{-1/4})$,

$$\frac{\widehat{ATT}(d)-ATT(d)}{\sigma_N/\sqrt{N}}\quad\to^d\quad N(0,1)$$

where

$$\sigma_N^2 \equiv E\Big[\Big(\frac{1}{f_d}(K_h(D-d)\Delta Y - E[K_h(D-d)\Delta Y]) \\ - \psi_J(Z,\theta_J,f_d,\eta) + (\frac{\theta_J}{f_d} - \frac{\mathcal{E}^d_{\Delta Y}}{f_d})(K_h(D-d) - E[K_h(D-d)])\Big)^2\Big].$$

and ψ_J is defined as in (15)

The proof follows the general framework for DML procedures studied in CCDDHNR (2018). The asymptotic variance roughly consists of two parts that contribute to the slower than $\sqrt{N}$ rate: the part from the orthogonal score ψ_J that grows with J and the part from the kernels used to nonparametrically estimate the density $f_D(d)$ and conditional mean $E[\Delta Y|D=d]$. We intentionally left the expression of the asymptotic variance in this way to avoid making further assumptions between the magnitudes of the kernel bandwidth h and M_J (through series cutoff J). A similar result for repeated cross sections is shown in the supplementary material with only minor modifications.

With a consistent estimator $\hat{\sigma}_N^2$ based on the expression in the theorem, one can establish a pointwise confidence interval for $ATT(d)$. Following CCDDHNR (2018) and Chang (2020), we consider the following cross-fitted variance estimator, while deferring the study of the theoretical properties of such estimator to future works:

$$\hat{\sigma}_N^2 := \frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\Big(\frac{1}{\hat{f}_{d,k}}(K_h(D-d)\Delta Y - E_{n^c,k}[K_h(D-d)\Delta Y]) \\ - \psi_J(Z,\hat{\theta}_J,\hat{f}_{d,k},\hat{\eta}_k) \\ + (\frac{\hat{\theta}_J}{\hat{f}_{d,k}} - \frac{\hat{\mathcal{E}}^d_{\Delta Y,k}}{\hat{f}_{d,k}})(K_h(D-d) - E_{n^c,k}[K_h(D-d)])\Big)^2]$$

where

$$\hat{\theta}_J \equiv \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k} \Delta Y_i \mathbf{1}\{D_i=0\}\frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k(X_i)} \\ + \frac{m_J^d(D_i)\hat{g}_k(X_i) - \mathbf{1}\{D_i=0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k^2(X_i)}\hat{\mathcal{E}}_{\Delta Y,k}(X_i)$$

and $E_{n^c,k}$ denotes the empirical average using the auxiliary sample I_k^c. Then, the $1-\alpha$ confidence interval can be constructed as $[\widehat{ATT}(d) - z_{1-\alpha/2}\hat{\sigma}_N/\sqrt{N}, \widehat{ATT}(d) + z_{1-\alpha/2}\hat{\sigma}_N/\sqrt{N}]$ where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal random variable.

Alternatively, one can use a multiplier bootstrap type of procedure to construct the

confidence interval for our estimator. We postpone the study of the theoretical properties of such bootstrap procedures to future iterations of this project, while noting the usage of such procedures in recent studies, see e.g. Belloni et al. (2017), Su et al. (2019), Cattaneo and Jansson (2021), Colangelo and Lee (2022), and Fan et al. (2022). Specifically, let $\{\xi_i\}_{i=1}^N$ be an i.i.d. sequence of sub-exponential random variables independent of $\{Y_{i,t-1}, Y_{i,t}, D_i, X_i\}_{i=1}^N$ such that $E[\xi_i] = E[\xi_i^2] = 1$. Then for each $b = 1, \cdots, B$, we draw such a sequence $\{\xi_i\}_{i=1}^N$ and construct

$$\widehat{ATT(d)}_b^* \equiv \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\xi_i\Big(\hat{\mathcal{E}}^d_{\Delta Y,k} - \Delta Y_i \mathbf{1}\{D_i = 0\}\frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k(X_i)} - \frac{m_J^d(D_i)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k^2(X_i)}\hat{\mathcal{E}}_{\Delta Y,k}(X_i)\Big)$$

Let $\hat{c}_\alpha$ be the α's quantile of $\{\widehat{ATT(d)}_b^* - \widehat{ATT(d)}\}_{b=1}^B$, and we construct the confidence interval as $[\widehat{ATT(d)} - \hat{c}_{1-\alpha/2}, \widehat{ATT(d)} - \hat{c}_{\alpha/2}]$.

4.4 Application: Revisit Duflo (2001)

Duflo (2001) studies the impact of a large policy intervention (INPRES program) taken place in Indonesia between 1973 and 1978. During this period, more than 60 thousands elementary schools were constructed in various regions in Indonesia, which is equivalent to about 2 schools per one thousand school-age children (see Duflo (2001) and Ashraf et al. (2020) for additional background details). Nevertheless, the "intensity" of this policy intervention was not uniform across Indonesia. In particular, Duflo (2001) models the treatment intensity as the number of schools constructed per 1000 children under this policy in each region. In the data set we consider, there are 161 regions, and the program intensity varies widely across the regions. Therefore, we model the treatment intensity as a continuous variable.

One of the main questions explored in Duflo (2001) is the effect of this policy on the educational attainment. As pointed out in the study, there is another dimension of variation in the treatment intensity: the cohort of children aged 12-17 in 1974 (cohort 0) would have already passed the elementary school age when the policy first started so that this cohort should not have benefited from the policy at all; on the other hand, the cohort aged 2-6 in 1974 (cohort 1) should have fully experienced the treatment. Moreover, based on the treatment intensity, the author divides the regions into two groups (low intensity group and high intensity group). Exploring the two-dimensional variations in treatment intensity across regions and cohorts, Duflo (2001) initially attempts to estimate the causal effect of

this policy on the educational attainment by using a simple difference-in-differences design under the usual parallel trend assumption (see Table 3 in Duflo (2001)).[10]

To study the treatment effect of this policy in our setting, we still consider the same two cohorts, which will be our repeated cross-sections. On the other hand, while we also use the low intensity regions as the control group, we will allow the treatment intensity to vary at the district level in the high intensity group (treatment group). We consider the following setup:

- let $T_i = 1$ if individual i belongs to the cohort 1 (age 2-6 in 1974) and $T_i = 0$ other wise (age 12-17 in 1974);
- the district level treatment intensity is defined as the schools constructed under this policy per 1000 school-aged children in a given birth district (importantly, this ensures the validity for the repeated cross section setup, as the treatment intensities are known to both cohorts);
- we define the regions with treatment intensities at or below 40 percentile on the distribution as the "low" group, and the regions with treatment intensities at or above 60 percentile on the distribution as the "high" group;
- we normalize the "low" group to have treatment intensity $D = 0$;
- for the "high" group, we re-define the treatment intensity by subtracting the 40 percentile value of the treatment intensity on the overall distribution; this ensures that the treatment intensities $D = d$ for the high group fall under an interval $[d_L, d_H]$ with $d_L > 0$;
- let Y_i denote the educational level of individual i;
- we include the following covariates X_i: gender, religion, land ownership (as an approxy for family wealth), community size, urban/rural residency;
- finally, for our sample, we consider all individuals who had stayed in the regions they were born, which is in contrast with Duflo (2001) in which the author considers the sample of males with valid wage data.

Remark 4.2. *The choice of using* 40/60 *percentiles as cutoffs for low and high intensity regions is rather arbitrary as we aren't able to find the exact criteria used in Duflo (2001)*

[10] We want to emphasize that besides the simple DiD design mentioned here, Duflo (2001) explores the effects of this policy on education and wage in various other research designs in great details. We only intend to use this exercise as an illustration on how to apply the continuous DiD design and our nonparametric estimator in an empirical setting and hopefully to showcase the potential usefulness of our methods.

to define the "low" vs. "high" regions. Nevertheless, with this 40/60 *cutoff, the mean difference in treatment intensities between "low" and "high" regions in our setting roughly matches those in Duflo (2001).*

Remark 4.3. *In our setting, we do not include the district level covariates, district fixed effects, and birth-year fixed effects. In particular, since the treatment intensity (and hence the treatment status) is defined at district level, the nonparametric machine learning methods such as Random Forest and deep neural networks can often perfectly predict the treatment status with such district level covariates, which creates issues for estimations due to the zeros in the denominators. Moreover, since the cohorts are defined by the birth-year, including birth-year fixed effects in the covariates will make the cohorts T and covariates X correlated, which violates the sampling assumption in the repeated cross sections setting.*

Due to the discrepancy in the data, for comparison purposes, we first replicate the baseline diff-in-diff result between low and high intensity regions ($D_i \in \{0, 1\}$ in this case) between cohorts ($T_i \in \{0, 1\}$) in Duflo (2001), using the following regression specification:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 D_i + \beta_3 (T_i \times D_i) + \epsilon_i$$

and we report the estimated ATT in the first row in Table 1. Similar to the results in table 3 in Duflo (2001), our replication results suggest that the treatment effect is positive but not statistically significant. We also estimate the double/debiased nonparametric DiD estimator (with binary $D_i \in \{0, 1\}$) proposed in Chang (2020) with the same covariates we considered for our continuous DiD estimator, and we reports the estimated ATT in the second row in Table 1. We note that the nonparametric estimator from Chang (2020) with covariates shows a much larger treatment effect and has statistical significance.

Table 1: Diff-in-Diff with Binary Treatment

dep var: educ	ATT($D = 1$)	std. err	sample size (N)	covariates
Duflo (Basline)	0.0876	0.0710	41240	–
Chang (Nonparametric)	0.5237	0.1759	41240	✓

For our continuous DiD estimator, we consider 9 different treatment intensities ranging from 10-percentile to 90-percentile of the empirical distribution of the intensities in the treatment group. We present the results for our estimator in Table 2 with visualization in Figure 1 (since the results using either machine learning methods are relatively close, we only present the graph with results using the Random Forest). In contrast to the results using

Table 2: Diff-in-Diff with Continuous Treatment

	(Random Forest) ATT($D = d$)	(Neural Network) ATT($D = d$)	sample size (N)	covariates
CDID ($\alpha = 0.1$)	0.5682	0.6260	41240	✓
	(0.2134)	(0.2119)		
CDID ($\alpha = 0.2$)	0.0693	0.1284	41240	✓
	(0.1992)	(0.1986)		
CDID ($\alpha = 0.3$)	0.2490	0.3764	41240	✓
	(0.2362)	(0.2335)		
CDID ($\alpha = 0.4$)	2.0328	1.9631	41240	✓
	(0.2396)	(0.2386)		
CDID ($\alpha = 0.5$)	1.1096	1.1523	41240	✓
	(0.2377)	(0.2363)		
CDID ($\alpha = 0.6$)	0.3405	0.3925	41240	✓
	(0.2519)	(0.2502)		
CDID ($\alpha = 0.7$)	0.5238	0.5517	41240	✓
	(0.2555)	(0.2525)		
CDID ($\alpha = 0.8$)	0.1778	-0.0612	41240	✓
	(0.4074)	(0.4043)		
CDID ($\alpha = 0.9$)	0.3857	0.4481	41240	✓
	(0.4352)	(0.4356)		

Notes: (i) α indicates the treatment intensity d being the corresponding percentile values, with standard errors in parentheses; (ii) in column 2, all the nuisance parameters are estimated using the Random Forest (RF) methods; (iii) in column 3, all the nuisance parameters are estimated using the deep neural network of multi-layer perceptron (MLP) class with ReLU activation; (iv) see the supplementary material for the other implementation details.

binary treatment, our results suggest that the ATT varies widely across different treatment intensities. In particular, for the nuisance parameters estimated using either the Random Forest (column 2) or deep neural network (column 3), we have large positive ATTs at some intensities (40 and 50 percentile values) but not others. We also want to emphasize that, echoing Callaway et al. (2021), each of these ATTs is local in nature (i.e. on its own dose-response curve), and the differences between ATTs, say $ATT(d_1) - ATT(d_2)$, can not be interpreted as the average causal response without further assumptions. Nevertheless, our estimation results show significant heterogeneity in treatment effects, which suggests that in practice, the researchers should fully explore the continuous nature of the treatments, and our framework offers one avenue to achieve this.

5 Conclusion

In this paper, we have proposed a data-driven conditional density estimator that is feasible for potentially high-dimensional conditioning variables. This estimator is based on a cross-

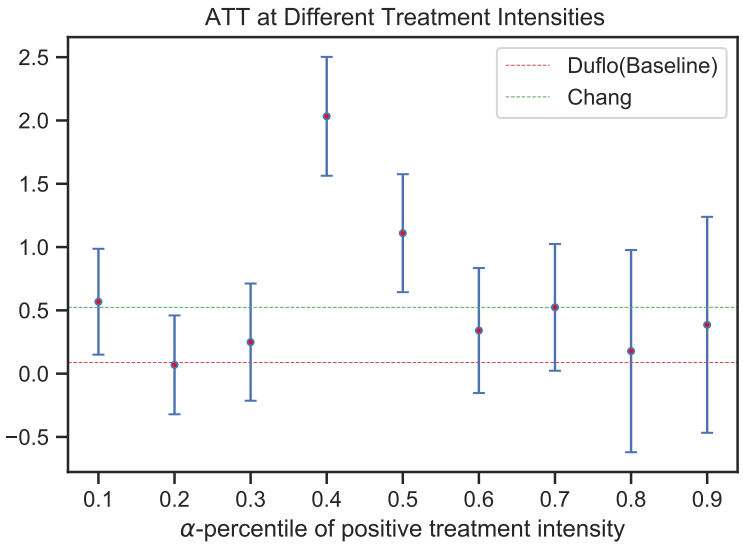

Figure 1: Diff-in-Diff with Continuous Treatment (Random Forest)

validation procedure and we have established an oracle inequality on its estimation error. Importantly, this data-driven conditional density estimator has the potential to accommodate any new machine learning methods (to estimate the conditional expectation in each of the series terms). Thus our estimator can facilitate better understanding of the dependence relationships between the economic variables, albeit the richer data sources and the increasing complexity of the economic models. Adding onto the growing list of economics applications where the conditional densities play a key role, we study the nonparametric difference-in-differences models with continuous treatments in detail. Such models have important empirical research implications and we hope our methods can aid future researches to further explore the effects of continuous treatment variables.

6 Simulations (WIP, ignore for now)

- code is ready, just need to implement it;
- can potentially also replicate and compare with the empirical example in Semenova and Chernozhukov (2021) if time permits.

A Proofs

A.1 Proof of Theorem 3.1

The proof consists of three main parts. In the first part, we show the loss Q and risk R are convex. Then we apply Lecué and Mitchell (2012) to upper bound the expected loss in $\|\cdot\|_H$ norm by the sum of the "oracle" and a shifted empirical process. Finally, we use boundedness of the true conditional density and of the estimators to control the shifted empirical process.

Step 1: Convexity of Loss

We first show the loss $Q((y,x),f) := \int f^2(y,x)d\nu_Y(y) - 2f(y,x)$ is convex in f. Take any $\lambda \in (0,1)$ and $f_1, f_2 \in L^2(\nu_Y \otimes \mu_X)$, supressing (y,x) in Q for notation simplicity, we have

$$\begin{aligned} Q(\lambda f_1 + (1-\lambda)f_2) &= \int (\lambda f_1 + (1-\lambda)f_2)^2 d\nu_Y(y) - 2(\lambda f_1 + (1-\lambda)f_2) \\ &\leq \int \lambda f_1^2 + (1-\lambda)f_2^2 d\nu_Y(y) - 2(\lambda f_1 + (1-\lambda)f_2) \\ &= \lambda Q(f_1) + (1-\lambda)Q(f_2) \end{aligned}$$

which proves the convexity of Q in f for any $(y,x) \in \mathbf{Y} \times \mathbf{X}$. Then the convexity of risk $R(f) := E[Q((Y,X),f)]$ follows from the monotonicity and linearity of expectation:

$$\begin{aligned} R(\lambda f_1 + (1-\lambda)f_2) &= E[Q((Y,X); \lambda f_1 + (1-\lambda)f_2)] \\ &\leq E[\lambda Q((Y,X), f_1) + (1-\lambda)Q((Y,X), f_2)] \\ &= \lambda R(f_1) + (1-\lambda)R(f_2). \end{aligned}$$

Using the convexity, next we are going to bound the risk.

Step 2: Bound on the Risk

This part of the proof is adapted from Lecué and Mitchell (2012), which we replicate here for the sake of completeness. Since $\hat{j}^*$ is the index that minimizes $R_{n,V}(\hat{f}_j)$, we define $R^*_{n,V}$ as the minimized empirical risk, that is,

$$R^*_{n,V} = \frac{1}{V}\sum_{k=1}^{V} \frac{1}{n_V} \sum_{i \in D_k^{(n_V)}} Q((Y_i, X_i), \hat{f}^{(n_T)}_{\hat{j}^*}(D_k^{(n_T)})).$$

Then, for all $1 \leq j \leq p$, the difference in the risk of our estimator and the risk at the true

conditional density can be rewritten as

$$
\begin{aligned}
&R(\bar{f}^{(n)}) - R(f_{Y|X}) \\
=&(1+a)(R^*_{n,V} - R_{n,V}(f_{Y|X})) + (R(\bar{f}^{(n)} - R(f_{Y|X})) - (1+a)(R^*_{n,V} - R_{n,V}(f_{Y|X})) \\
\leq&(1+a)(R_{n,V}(\hat{f}_j) - R_{n,V}(f_{Y|X})) + (R(\bar{f}^{(n)}) - R(f_{Y|X})) - (1+a)(R^*_{n,V} - R_{n,V}(f_{Y|X})).
\end{aligned}
$$

Taking expectation of $R_{n,V}(\hat{f}_j) - R_{n,V}(f_{Y|X})$ with respect to the full data, we have

$$
\begin{aligned}
&E[R_{n,V}(\hat{f}_j) - R_{n,V}(f_{Y|X})] \\
=&E[\frac{1}{V}\sum_{k=1}^{V}\frac{1}{n_V}\sum_{i\in D_k^{(n_V)}} Q((Y_i,X_i),\hat{f}_j^{(n_T)}(D_k^{(n_T)})) - Q((Y_i,X_i),f_{Y|X})] \\
=&\frac{1}{V}\sum_{k=1}^{V}\frac{1}{n_V}\sum_{i\in D_k^{(n_V)}} E[Q((Y_i,X_i),\hat{f}_j^{(n_T)}(D_k^{(n_T)}))] - E[Q((Y_i,X_i),f_{Y|X})] \\
=&E_{D^{(n_T)}}[R(\hat{f}_j^{(n_T)}(D^{(n_T)}))] - R(f_{Y|X})
\end{aligned}
$$

where the second equality holds since $\{(Y_i,X_i)\}_{i=1}^n$ are i.i.d. and the last equality holds by law of iterated expectation. Moreover, by convexity of R, we have

$$
\begin{aligned}
R(\bar{f}^{(n)}) &= R(\frac{1}{V}\sum_{k=1}^{V}\hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)})) \\
&\leq \frac{1}{V}\sum_{k=1}^{V} R(\hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)})) := \frac{1}{V}\sum_{k=1}^{V} PQ((Y,X),\hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)}))
\end{aligned}
$$

where P denotes the probability measure with respect to (Y,X). Then

$$
\begin{aligned}
&E[(R(\bar{f}^{(n)}) - R(f_{Y|X})) - (1+a)(R^*_{n,V} - R_{n,V}(f_{Y|X}))] \\
\leq&E[\frac{1}{V}\sum_{k=1}^{V} PQ((Y,X),\hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)})) - PQ((Y,X),f_{Y|X}) \\
&- (1+a)(\frac{1}{V}\sum_{k=1}^{V}\frac{1}{n_V}\sum_{i\in D_k^{(n_V)}} Q((Y_i,X_i),\hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)})) - Q((Y_i,X_i),f_{Y|X})))] \\
=&\frac{1}{V}\sum_{k=1}^{V} E[P(Q((Y,X),\hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)})) - Q((Y,X),f_{Y|X})) \\
&- \frac{1+a}{n_V}\sum_{i\in D_k^{(n_V)}} Q((Y_i,X_i),\hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)})) - Q((Y_i,X_i),f_{Y|X})))]
\end{aligned}
$$

$$\leq E[\max_{1\leq j\leq p}(P-(1+a)P_{n_V})(Q((Y,X),\hat{f}^{(n_T)}_{\hat{j}^*}(D^{(n_T)}))-Q((Y,X),f_{Y|X}))].$$

Combine above results together, and use the definition that $R(\bar{f}^{(n)}) = \|\bar{f}^{(n)} - f_{Y|X}\|_H^2 - \|f_{Y|X}\|_H^2$ and $R(f_{Y|X}) = -\|f_{Y|X}\|_H^2$, we have

$$\begin{aligned}
&E[\|\bar{f}^{(n)} - f_{Y|X}\|_H^2] \\
\leq &\min_{1\leq j\leq p}(1+a)E[\|\hat{f}_j^{(n_T)} - f_{Y|X}\|_H^2] \\
&+ E[\max_{1\leq j\leq p}(P-(1+a)P_{n_V})(Q((Y,X),\hat{f}^{(n_T)}_{\hat{j}^*}(D^{(n_T)}))-Q((Y,X),f_{Y|X}))].
\end{aligned} \tag{17}$$

In the next section, we bound the the max of the shifted empirical process using a modified maximal inequality inspired by Lecué and Mitchell (2012) Lemma 5.3.

Step 3: A Maximal Inequality on Shifted Empirical Process

We first show a maximal inequality. Let $\{G_1, \cdots, G_p\}$ be a set of measurable functions on $(\mathbf{Z}, \mathcal{B}_Z)$, and $\{Z_i\}_{i=1}^n \sim Z$ a sequence of i.i.d. random variables with $Z \in (\mathbf{Z}, \mathcal{B}_Z)$. Moreover, we assume that, for all $1 \leq j \leq p$, (i) $E[G_j(Z)] \geq 0$; (ii) $\|G_j(Z)\|_{L_2} \leq C(E[G_j(Z)])^{1/2}$ for some contant C; (iii) $\|G_j\|_\infty \leq \tilde{M}$ for some constant $\tilde{M}$.

Consider any $x > 0$,

$$\begin{aligned}
&P\left[\max_{1\leq j\leq p} E[G_j(Z)] - (1+a)\frac{1}{n}\sum_{i=1}^n G_j(Z_i) \geq x\right] \\
\leq &\sum_{j=1}^p P\left[E[G_j(Z)] - (1+a)\frac{1}{n}\sum_{i=1}^n G_j(Z_i) \geq x\right] \\
= &\sum_{j=1}^p P\left[E[G_j(Z)] - \frac{1}{n}\sum_{i=1}^n G_j(Z_i) \geq \frac{x + aE[G_j(Z)]}{1+a}\right]
\end{aligned}$$

where the second inequality holds by union bound. Then, for each term in the sum, we have for some constants c_1, c_2, c_3, c_4,

$$\begin{aligned}
&P\left[E[G_j(Z)] - \frac{1}{n}\sum_{i=1}^n G_j(Z_i) \geq \frac{x + aE[G_j(Z)]}{1+a}\right] \\
&\leq \exp\left(-c_1 n \frac{(\frac{x+aE[G_j(Z)]}{1+a})^2}{E[G_j^2(Z)] + \tilde{M}\frac{x+aE[G_j(Z)]}{1+a}}\right) \\
&\leq \exp\left(-c_2 n \frac{(\frac{x+aE[G_j(Z)]}{1+a})^2}{E[G_j^2(Z)]} \wedge \frac{x+aE[G_j(Z)]}{1+a}{\tilde{M}}\right)
\end{aligned}$$

$$\leq \exp\left(-c_3 n \frac{(x + aE[G_j(Z)])^2}{E[G_j^2(Z)]} \wedge \frac{x + aE[G_j(Z)]}{\tilde{M}}\right)$$
$$\leq \exp\left(-c_4 n \left(\frac{x + aE[G_j(Z)]}{(E[G_j(Z)])^{1/2}}\right)^2 \wedge \frac{x + aE[G_j(Z)]}{\tilde{M}}\right)$$

where the first inequality holds by Bernstein's inequality (see, for example, van der Vaart and Wellner (1996) Lemma 2.2.9), the second inequality holds by definition ($\wedge$ is the minimum operator), and the last inequality holds by the condition that $\|G_j(Z)\|_{L_2} \leq C(E[G_j(Z)])^{1/2}$. Then comparing x with $E[G_j(Z)]$, we have

$$\left(\frac{x + aE[G_j(Z)]}{(E[G_j(Z)])^{1/2}}\right)^2 \wedge \frac{x + aE[G_j(Z)]}{\tilde{M}} \quad \gtrsim \quad \frac{x}{\tilde{M}}$$

which implies that

$$P\left[\max_{1\leq j\leq p} E[G_j(Z)] - (1+a)\frac{1}{n}\sum_{i=1}^n G_j(Z_i) \geq x\right] \leq p\exp(-C_1 n \frac{x}{\tilde{M}}).$$

Then, for any $u > 0$, we have

$$\begin{aligned}
&E\left[\max_{1\leq j\leq p} E[G_j(Z)] - (1+a)\frac{1}{n}\sum_{i=1}^n G_j(Z_i)\right] \\
\leq &\int_0^\infty P\left[\max_{1\leq j\leq p} E[G_j(Z)] - (1+a)\frac{1}{n}\sum_{i=1}^n G_j(Z_i) \geq x\right] d\mu(x) \\
\leq &\int_0^\infty p\exp(-C_1 n\frac{x}{\tilde{M}}) d\mu(x) \\
\leq &u + p\int_u^\infty \exp(-C_1 n\frac{x}{\tilde{M}}) d\mu(x) \\
\leq &u + p\frac{\exp(-C_1 nu/\tilde{M})}{C_1 n/\tilde{M}}
\end{aligned}$$

where the first inequality holds since $E[X] = \int_{\mathbf{R}} 1_{x\geq 0}(x) - F_X(x) d\mu(x)$ and the last inequality holds using the fact that $\int_u^\infty \exp(-Bt)dt \leq \exp(-Bu)/B$ (see, for example, Lecué and Mitchell (2012) Lemma 5.3). Define $x(p)$ to be the unique solution of $x = p\exp(-x)$, which satisfies $x(p) \leq \log(ep)$. Let $u = \tilde{M}x(p)/(nC_1)$, we have

$$u + p\frac{\exp(-C_1 nu/\tilde{M})}{C_1 n/\tilde{M}} = \frac{2\tilde{M}x(p)}{nC_1} \leq \frac{2\tilde{M}\log(ep)}{C_1 n}.$$

Therefore, we conclude that, for some constant C_2 that only depends on a and C,

$$E\left[\max_{1\leq j\leq p} E[G_j(Z)] - (1+a)\frac{1}{n}\sum_{i=1}^{n} G_j(Z_i)\right] \leq C_2\frac{\tilde{M}\log(p)}{n}.$$

Note that throughout the derivation, we kept the constant $\tilde{M}$ explicit to accommodate the possibility of $\tilde{M}$ potentially growing with p.[11]

Step 4: Bound on Shifted Empirical Process

Now we apply this maximal inequality in our case. We need to first verify the assumptions used in *Step 3*. Conditional on $\{\hat{f}_j\}_{j=1}^p$, let $Z := (Y, X)$ and define

$$G_j(Z) := Q(Z, \hat{f}_j) - Q(Z, f_{Y|X})$$

where Q is the loss defined in 9. First, by definition,

$$\begin{aligned} E[G_j(Z)] &= E[Q(Z,\hat{f}_j) - Q(Z, f_{Y|X})] \\ &= \|\hat{f}_j - f_{Y|X}\|_H^2 - \|f_{Y|X}\|_H^2 - (-\|f_{Y|X}\|_H^2) \\ &= \|\hat{f}_j - f_{Y|X}\|_H^2 \\ &\geq 0. \end{aligned}$$

Next, we check $\|G_j(Z)\|_{L_2} \leq C(E[G_j(Z)])^{1/2}$ for some constant C. Plug in the definition of the loss Q, we have

$$\begin{aligned} \|G_j\|_{L_2} &= \|Q(Z,\hat{f}_j) - Q(Z, f_{Y|X})\|_{L_2} \\ &= \|\int \hat{f}_j(y|X)^2 d\nu(y) - 2\hat{f}_j(Y|X) - \int f_{Y|X}(y)^2 d\nu(y) - 2f_{Y|X}\|_{L_2} \\ &= \|\int (\hat{f}_j(y|X) - f_{Y|X}(y))(\hat{f}_j(y|X) + f_{Y|X}(y))d\nu(y) - 2(\hat{f}_j(Y|X) - f_{Y|X})\|_{L_2} \\ &\leq \|\int (\hat{f}_j(y|X) - f_{Y|X}(y))(\hat{f}_j(y|X) + f_{Y|X}(y))d\nu(y)\|_{L_2} + 2\|(\hat{f}_j(Y|X) - f_{Y|X})\|_{L_2} \end{aligned}$$

where the last line holds by triangle inequality. For the first term above, we have

$$\begin{aligned} &\|\int (\hat{f}_j(y|X) - f_{Y|X}(y))(\hat{f}_j(y|X) + f_{Y|X}(y))d\nu(y)\|_{L_2}^2 \\ =&E\left[\left(\int (\hat{f}_j(y|X) - f_{Y|X}(y))(\hat{f}_j(y|X) + f_{Y|X}(y))d\nu(y)\right)^2\right] \end{aligned}$$

[11] The constant C in assumption (ii), that $\|G_j(Z)\|_{L_2} \leq C(E[G_j(Z)])^{1/2}$, can also depend on $\tilde{M}$. The proofs can be modified accordingly to accommodate this possibility.

$$
\begin{aligned}
&\leq E\left[\int (\hat{f}_j(y|X) - f_{Y|X}(y))^2 d\nu(y) \int (\hat{f}_j(y|X) + f_{Y|X}(y))^2 d\nu(y)\right] \\
&\leq E\left[\int (\hat{f}_j(y|X) - f_{Y|X}(y))^2 d\nu(y)(4M) \int \frac{(\hat{f}_j(y|X) + f_{Y|X}(y))}{2} d\nu(y)\right] \\
&\leq 4ME\left[\int (\hat{f}_j(y|X) - f_{Y|X}(y))^2 d\nu(y)\right] \\
&= 4M\|\hat{f}_j - f_{Y|X}\|_H^2 \\
&= 4ME[G_j]
\end{aligned}
$$

where the first line holds by definition, the second line holds by Cauchy-Schwarz, the third line holds by our assumption that $\{\hat{f}_j\}_{j=1}^p$ and $f_{Y|X}$ are uniformly bounded by some constant M, the fourth line holds since $(\hat{f}_j + f_{Y|X})/2$ is still a density that integrates to 1, and the last line holds by definition of $E[G_j] = E[Q(\hat{f}_j) - Q(f_{Y|X})] = \|\hat{f}_j - f_{Y|X}\|_H^2$. For the second term, note that

$$
\begin{aligned}
\|(\hat{f}_j(Y|X) - f_{Y|X})\|_{L_2}^2 &= E[(\hat{f}_j(Y|X) - f_{Y|X})^2] \\
&= E_X E_{Y|X}[(\hat{f}_j(Y|X) - f_{Y|X})^2] \\
&= E_X[\int (\hat{f}_j(Y|X) - f_{Y|X})^2 f_{Y|X}(y) d\nu(y)] \\
&\leq 2ME_X[\int (\hat{f}_j(Y|X) - f_{Y|X})^2 \nu(y)] \\
&= 2M\|\hat{f}_j - f_{Y|X}\|_H^2 \\
&= 2ME[G_j]
\end{aligned}
$$

where the second line holds by law of iterated expectation and the fourth line holds by boundedness of $f_{Y|X}$. Therefore, combine above results together, we have shown that

$$
\|G_j\|_{L_2} \leq 2M^{\frac{1}{2}} (E[G_j])^{\frac{1}{2}}
$$

so we can take the constant $C := 2M^{1/2}$.

Finally, we check $\|G_j\|_\infty \leq \tilde{M}$ for some constant $\tilde{M}$. By definition

$$
\|G_j\|_\infty = \| \int \hat{f}_j(y|x)^2 d\nu(y) - 2\hat{f}_j(y|x) - \int f_{Y|X=x}(y)^2 d\nu(y) - 2f_{y|x}\|_\infty \leq 6M
$$

where the inequality holds by boundedness of $\hat{f}_j$ and $f_{Y|X}$, so we can take $\tilde{M} = 6M$.

Then we apply *Step 3* conditional on $\{\hat{f}_j\}_{j=1}^p$ and use law of iterated expectation and monotonicity of expectation to conclude. We want to emphasize on the fact that we can

allow the bound on the dictionary $\{\hat{f}_j\}_{j=1}^p$ to potentially grow with p. For example, if the bound $M = O(\log(p))$, then there is one extra $\log(p)$ term (or some polynomial power of it) showing up in the rate in the theorem. ■

A.2 Proof of Theorem 3.2

First, given that V is fixed, the training sample size n_T and testing/validating sample size n_V are on the same order as n, so we will drop the supscripts. Let $\{\phi_j\}_{j=1}^\infty$ be an orthonormal basis on $L^2(\nu_Y)$ and let's denote $h_j = E[\phi_j(Y)|X]$ and $\hat{h}_j$ the corresponding estimator. Then by definition, for a given $j \in \{1, \cdots, p\}$, we have

$$\begin{aligned}
E[\|\hat{f}_j - f_{Y|X}\|_H^2] &= E[\|\sum_{k=1}^{j} \hat{h}_k\phi_k - \sum_{k=1}^{\infty} h_k\phi_k\|_H^2] \\
&= E[\|\sum_{k=1}^{j} (\hat{h}_k - h_k)\phi_k - \sum_{k=j+1}^{\infty} h_k\phi_k\|_H^2] \\
&= E[E_X[\int \left(\sum_{k=1}^{j} (\hat{h}_k(X) - h_k(X))\phi_k(y) - \sum_{k=j+1}^{\infty} h_k(X)\phi_k(y)\right)^2 d\nu(y)]] \\
&= E[E_X[\sum_{k=1}^{j} \left(\hat{h}_k(X) - h_k(X)\right)^2 + \sum_{k=j+1}^{\infty} h_k^2(X)]] \\
&= \sum_{k=1}^{j} E[(\hat{h}_k(X) - h_k(X))^2] + \sum_{k=j+1}^{\infty} E[h_k^2(X)]
\end{aligned}$$

where the second to last equality holds by orthonormality of the basis $\{\phi_j\}_{j=1}^\infty$. By assumption, for some constants $\delta, \gamma > 0$, we have the variance $E[(\hat{h}_k(X) - h_k(X))^2] \asymp n^{-\delta}$ and bias $\sum_{k=j+1}^\infty E[h_k^2(X)] \lesssim j^{-\gamma}$, which implies

$$E[\|\hat{f}_j - f_{Y|X}\|_H^2] \lesssim jn^{-\delta} + j^{-\gamma}.$$

Then minimizing over j, we have the minimizer $j^* = n^{\delta/(\gamma+1)}$. Given the assumption on p, this minimizer can be attained in our dictionary of estimators, which gives us

$$\min_{1 \le j \le p} E[\|\hat{f}_j - f_{Y|X}\|_H^2] \lesssim n^{-\frac{\gamma}{\gamma+1}\delta}.$$

Combine this result with the oracle inequality in 3.1, we have the desired result. ■

A.3 Proof of Theorem 3.3

Let $h_j(x) := E[\phi_j(Y)|X = x]$ and $\hat{h}_j(x)$ being its estimator. Then for any $J \geq 1$ and $y \in \mathbf{Y}$,

$$\begin{aligned}
&E[\|\hat{f}_J(y|X) - f_{Y|X}(y|X)\|^2_{P_X}] \\
=&E[\int (\sum_{j=1}^{J} h_j(x)\phi_j(y) - f_{Y|X}(y|x))^2 dP_X(x)] \\
\leq&E[\int \sum_{j=1}^{J} (h_j(x) - \hat{h}_j(x))^2 \phi_j^2(y) dP_X(x)] + \int (\sum_{j=J+1}^{\infty} h_j(x)\phi_j(y))^2 dP_X(x).
\end{aligned}$$

First, we focus on the second term

$$\begin{aligned}
&\int (\sum_{j=J+1}^{\infty} h_j(x)\phi_j(y))^2 dP_X(x) \\
=&\int \sum_{j=J+1}^{\infty} (h_j(x)\phi_j(y))^2 dP_X(x) + \sum_{\substack{j,k=J+1 \\ j\neq k}}^{\infty} h_j(x)h_k(x)\phi_j(y)\phi_k(y) dP_X(x) \\
\leq& \sum_{j=J+1}^{\infty} E[h_j^2(X)]\phi_j^2(y) + \sum_{\substack{k=J+1 \\ k\neq j}}^{\infty} \sum_{j=J+1}^{\infty} (E[h_j^2(X)]\phi_j^2(y))^{\frac{1}{2}} (E[h_k^2(X)]\phi_k^2(y))^{\frac{1}{2}} \\
\leq& C_1 \sum_{j=J+1}^{\infty} E[h_j^2(X)] + C_2 \sum_{\substack{k=J+1 \\ k\neq j}}^{\infty} \sum_{j=J+1}^{\infty} (E[h_j^2(X)])^{\frac{1}{2}} (E[h_k^2(X)])^{\frac{1}{2}} \\
\leq& C_3 \sum_{j=J+1}^{\infty} E[h_j^2(X)]
\end{aligned}$$

where the first inequality holds by triangle inequality and Cauchy-Schwarz, and the second inequality holds by the boundedness of the orthonormal basis.

Now consider the first term $E[\int \sum_{j=1}^{J} (h_j(x) - \hat{h}_j(x))^2 \phi_j^2(y) dP_X(x)]$. Define the column vector $B_J(X) := (h_j(X) - \hat{h}_j(X))_{j=1}^J$, $P_J(y) := (\phi_j(y))_{j=1}^J$, $\Sigma_J := E[B_J(X)B_J(X)']$, and rewrite

$$E[\int \sum_{j=1}^{J} (h_j(x) - \hat{h}_j(x))^2 \phi_j^2(y) dP_X(x)] = E[(P_J(y)'B_J(X))^2] = P_J(y)'\Sigma_J P_J(y).$$

Moreover, let $\overline{EIG}$ and $\underline{EIG}$ denote the largest and smallest eigenvalues of Σ_J respectively.

Then

$$P_J(y)'\Sigma_J P_J(y) \leq \overline{EIG} \cdot \|P_J(y)\|_2^2$$
$$= \frac{\|P_J(y)\|_2^2}{\int \|P_J(y)\|_2^2 \, d\nu_Y(y)} \times \frac{\overline{EIG}}{\underline{EIG}} \times \underline{EIG} \int \|P_J(y)\|_2^2 \, d\nu_Y(y).$$

Note that $\|P_J(y)\|_2^2 / \int \|P_J(y)\|_2^2 \, d\nu_Y(y) = O(1)$ by orthonormality, $\overline{EIG}/\underline{EIG} = O(1)$ by assumption, and the last term is bounded by

$$\underline{EIG} \int \|P_J(y)\|_2^2 \, d\nu_Y(y) \leq \int P_J'(y)\Sigma P_J(y) d\nu_Y(y) = \sum_{j=1}^{J} E[(\hat{h}_j(X) - h_j(X))^2].$$

where the last equality holds by orthonormality. Combining above results, we have

$$E[\|\hat{f}_J(y|X) - f_{Y|X}(y|X)\|_{P_X}^2] \lesssim \sum_{j=1}^{J} E[(\hat{h}_j(X) - h_j(X))^2] + \sum_{j=J+1}^{\infty} E[h_j^2(X)]$$

which is the same bound as in the MISE case. Then use the cross-validated $\hat{J}^*$, Theorem 3.2, and the fact that the upper bound does not depend on y, we conclude that

$$\sup_y E[\|\hat{f}_J(y|X) - f_{Y|X}(y|X)\|_{P_X}^2] \lesssim n^{-\frac{\gamma}{\gamma+1}\delta} \vee \frac{\log p}{n}.$$

■

A.4 Proof of Theorem 4.1

By definition, $ATT(d) = E[Y_t(d) - Y_t(0)|D = d]$. First,

$$E[Y_t - Y_{t-1}|D = d] = E[Y_t(d) - Y_{t-1}(0)|D = d]$$

the fact that $Y_t|_{D=d} = Y_t(d)$ and $Y_{t-1}|_{D=d} = Y_{t-1} = Y_{t-1}(0)$.

Second,

$$E[(Y_t - Y_{t-1})\mathbf{1}\{D = 0\}\frac{f_{D|X}(d)}{f_D(d)P(D = 0|X)}]$$
$$= E[(Y_t - Y_{t-1})\frac{f_{D|X}(d)}{f_D(d)P(D = 0|X)}|D = 0]P(D = 0)$$
$$= E[E[(Y_t - Y_{t-1})|X = x, D = 0]\frac{f_{D|X=x}(d)P(D = 0)}{f_D(d)P(D = 0|X = x)}|D = 0]$$

$$
\begin{aligned}
&= \int E[(Y_t(0) - Y_{t-1}(0))|X = x, D = 0] \frac{f_{D|X=x}(d)P(D=0)}{f_D(d)P(D=0|X=x)} f_{X|D=0}(x)d\mu(x) \\
&= \int E[(Y_t(0) - Y_{t-1}(0))|X = x, D = d] \\
&\times \frac{f_{D|X=x}(d)P(D=0)}{f_D(d)P(D=0|X=x)} \frac{P(D=0|X=x)f_X(x)}{P(D=0)} d\mu(x) \\
&= \int E[(Y_t(0) - Y_{t-1}(0))|X = x, D = d] f_{X|D=d}(x)d\mu(x) \\
&= E[(Y_t(0) - Y_{t-1}(0))|D = d]
\end{aligned}
$$

where the first equality holds by the law total probability, second equality holds by law of iterated expectation, the third equality holds by that $Y_t|_{D=0} = Y_t(0)$ and $Y_{t-1}|_{D=0} = Y_t(0)$, the fourth equality holds by Bayes' rule and conditional parallel trend, and the fifth equality holds by Bayes rule.

Then combining above results, we have

$$
\begin{aligned}
&E[(Y_t - Y_{t-1}|D = d] - E[(Y_t - Y_{t-1})\mathbf{1}\{D = 1\} \frac{f_{D|X}(d)}{f_D(d)P(D=0|X)}] \\
&= E[Y_t(d) - Y_{t-1}(0)|D = d] - E[Y_t(0) - Y_{t-1}(0)|D = d] \\
&= E[Y_t(d) - Y_t(0)|D = d] \\
&= ATT(d)
\end{aligned}
$$

■

A.5 Proof of Theorem 4.2

We show that the expression in the theorem can be expressed as that of 4.1 and the proof follows. First, note that

$$
\begin{aligned}
&E[\frac{T-\lambda}{\lambda(1-\lambda)} Y|D = d] \\
&= E[E[\frac{T-\lambda}{\lambda(1-\lambda)} Y|D = d, T]] \\
&= E[\frac{T-\lambda}{\lambda(1-\lambda)} Y|D = d, T = 1]P(T = 1|D = d) \\
&+ E[\frac{T-\lambda}{\lambda(1-\lambda)} Y|D = d, T = 0]P(T = 0|D = d) \\
&= E[\frac{1-\lambda}{\lambda(1-\lambda)} Y|D = d, T = 1]\lambda + E[\frac{0-\lambda}{\lambda(1-\lambda)} Y|D = d, T = 0](1-\lambda)
\end{aligned}
$$

$$= E[Y_t|D=d] - E[Y_{t-1}|D=d]$$
$$= E[Y_t - Y_{t-1}|D=d]$$

where the first equality holds by law of iterated expectation, the second equality holds by definition, and the last two equalities hold by assumption 4.2.

Similarly, by law of iterated expectation and assumption 4.2

$$\begin{aligned}
&E[\frac{T-\lambda}{\lambda(1-\lambda)} Y \mathbf{1}\{D=0\} \frac{f_{D|X}(d)}{f_D(d)P(D=0|X)}] \\
&= E[\frac{1-\lambda}{\lambda(1-\lambda)} Y \mathbf{1}\{D=0\} \frac{f_{D|X}(d)}{f_D(d)P(D=0|X)} | T=1] P(T=1) \\
&+ E[\frac{0-\lambda}{\lambda(1-\lambda)} Y \mathbf{1}\{D=0\} \frac{f_{D|X}(d)}{f_D(d)P(D=0|X)} | T=0] P(T=0) \\
&= E[\frac{1-\lambda}{\lambda(1-\lambda)} Y_t \mathbf{1}\{D=0\} \frac{f_{D|X}(d)}{f_D(d)P(D=0|X)} | T=1] \lambda \\
&+ E[\frac{0-\lambda}{\lambda(1-\lambda)} Y_{t-1} \mathbf{1}\{D=0\} \frac{f_{D|X}(d)}{f_D(d)P(D=0|X)} | T=0] (1-\lambda) \\
&= E[(Y_t - Y_{t-1}) \mathbf{1}\{D=0\} \frac{f_{D|X}(d)}{f_D(d)P(D=0|X)}]
\end{aligned}$$

and the claim of the theorem follows from Theorem 4.1.■

A.6 Proof of Lemma 4.1

First consider the repeated outcomes case. Recall that the unadjusted score φ_J takes the form

$$\varphi_J(Z, \theta_J, f_d^0, f_J^0(d|X), g_0) \equiv \Delta Y \mathbf{1}\{D=0\} \frac{f_J^0(d|X)}{f_d^0 \cdot g_0(X)} - \theta_{0J}$$

where $\Delta Y = Y_t - Y_{t-1}$, $f_d^0 \equiv f_{\tilde{D}}(d)$, $f_J^0(d|X) \equiv f_{D|X}(d)$, $g_0(X) \equiv P(D=0|X)$. We will add an adjustment term to the original score so that the new score satisfies the Neyman orthogonality wrt the infinite dimensional parameters. Let $m_J^d(D) \equiv \sum_{j=1}^J \phi_j(D)\phi_j(d)\mathbf{1}\{D>0\}$.

The two infinite dimensional nuisance parameters are $f_J^0(X)$ and $g_0(X)$, and in particular, they satisfy $f_J^0(d|X) = E[m_J^d(D)|X]$ and $g_0(X) = E[\mathbf{1}\{D=0\}|X]$. Then the adjustment term c_J takes the form

$$c_J \equiv (m_J^d(D) - f_J^0(d|X)) E[\partial_1 \varphi_J | X] + (\mathbf{1}\{D=0\} - g_0(X)) E[\partial_2 \varphi_J | X]$$

where ∂_1 and ∂_2 denotes the partial derivatives wrt the positions of $f_J^0(d|X)$ and $g_0(X)$

respectively. Then, we have

$$
\begin{aligned}
c_J &= (m_J^d(D) - f_J^0(d|X))\frac{1}{f_d^0 \cdot g_0(X)} \underbrace{E[\Delta Y \mathbf{1}\{D=0\}|X]}_{\equiv \mathcal{E}_{\Delta Y}^0(X)} \\
&\quad - (\mathbf{1}\{D=0\} - g_0(X))\frac{f_J^0(d|X)}{f_d^0 \cdot g_0^2(X)}\mathcal{E}_{\Delta Y}^0(X) \\
&= \frac{[m_J^d(D) - f_J^0(d|X)]g_0(X) - [\mathbf{1}\{D=0\} - g_0(X)]f_J^0(d|X)}{f_d^0 \cdot g_0^2(X)}\mathcal{E}_{\Delta Y}^0(X) \\
&= \frac{m_J^d(D)g_0(X) - \mathbf{1}\{D=0\}f_J^0(d|X)}{f_d^0 \cdot g_0^2(X)}\mathcal{E}_{\Delta Y}^0(X)
\end{aligned}
$$

Now it remains to show the new score $\psi_J \equiv \varphi_J + c_J$ satisfies Neyman orthogonality wrt the nuisance parameters, $f_J^0(d|X)$, $g_0(X)$, and $\mathcal{E}_{\Delta Y}^0(X)$. First, we need to check $E[\psi_J] = 0$. Since $E[\varphi_J] = 0$, we only need to check $E[c_J] = 0$. Then we have

$$
\begin{aligned}
E[c_J] &= E[\frac{m_J^d(D)g_0(X) - \mathbf{1}\{D=0\}f_J^0(d|X)}{f_d^0 \cdot g_0^2(X)}\mathcal{E}_{\Delta Y}^0(X)] \\
&= E[\frac{E[m_J^d(D)|X]g_0(X) - E[\mathbf{1}\{D=0\}|X]f_J^0(d|X)}{f_d^0 \cdot g_0^2(X)}\mathcal{E}_{\Delta Y}^0(X)] \\
&= E[\frac{f_J^0(d|X)g_0(X) - g_0(X)f_J^0(d|X)}{f_d^0 \cdot g_0^2(X)}\mathcal{E}_{\Delta Y}^0(X)] \\
&= 0
\end{aligned}
$$

where the second equality holds by law of iterated expectation and the third equality holds by the fact that $E[m_J^d(D)|X] = f_J^0(d|X)$ and $E[\mathbf{1}\{D=0\}|X] = g_0(X)$.

Second, we need to show the Gateaux derivative of the score wrt the nuisance parameters $\eta_0 \equiv (f_J^0(d|X), g_0(X), \mathcal{E}_{\Delta Y}^0(X))$ vanishes at zero, that is, we need to show

$$
\partial_r E[\psi_J(\eta_0 + r(\eta - \eta_0))]|_{r=0} = 0.
$$

By the definition of Gateaux derivative, it suffices to show the partial derivative is zero wrt each nuisance parameter separately.

w.r.t $f_J(d|X)$:

$$
\begin{aligned}
&\partial_r E[\psi_J(f_J^0(d|X) + r(f_J(d|X) - f_J^0(d|X)))]|_{r=0} \\
&= E[(\Delta Y \mathbf{1}\{D=0\}\frac{1}{f_d^0 \cdot g_0(X)} - \frac{\mathbf{1}\{D=0\}}{f_d^0 \cdot g_0^2(X)}\mathcal{E}_{\Delta Y}^0(X))\Delta f_J]
\end{aligned}
$$

$$
\begin{aligned}
&= E[(E[\Delta Y \mathbf{1}\{D=0\}|X]\frac{1}{f_d^0 \cdot g_0(X)} - \frac{E[\mathbf{1}\{D=0\}|X]}{f_d^0 \cdot g_0^2(X)}\mathcal{E}^0_{\Delta Y}(X))\Delta f_J] \\
&= E[(\mathcal{E}^0_{\Delta Y}(X)\frac{1}{f_d^0 \cdot g_0(X)} - \frac{g_0(X)}{f_d^0 \cdot g_0^2(X)}\mathcal{E}^0_{\Delta Y}(X))\Delta f_J] \\
&= 0
\end{aligned}
$$

where the first equality holds by definition with $\Delta f_J \equiv f_J(d|X) - f_J^0(d|X)$, second equality holds by law of iterated expectation, and the third equality holds by the fact that $E[\Delta Y \mathbf{1}\{D=0\}|X] = \mathcal{E}^0_{\Delta Y}(X)$ and $E[\mathbf{1}\{D=0\}|X] = g_0(X)$.

w.r.t $g(X)$:

$$
\begin{aligned}
&\partial_r E[\psi_J(g_0(X) + r(g(X) - g_0(X)))]|_{r=0} \\
&= E[(-\Delta Y \mathbf{1}\{D=0\}\frac{f_J^0(d|X)}{f_d^0 \cdot g_0^2(X)} - (\frac{m_J^d(D)}{f_d^0 \cdot g_0^2(X)} - 2 \cdot \frac{\mathbf{1}\{D=0\}f_J^0(d|X)}{f_d^0 \cdot g_0^3(X)})\mathcal{E}^0_{\Delta Y}(X))\Delta g] \\
&= E[(-E[\Delta Y \mathbf{1}\{D=0\}|X]\frac{f_J^0(d|X)}{f_d^0 \cdot g_0^2(X)} - (\frac{E[m_J^d(D)|X]}{f_d^0 \cdot g_0^2(X)} \\
&- 2 \cdot \frac{E[\mathbf{1}\{D=0\}|X]f_J^0(d|X)}{f_d^0 \cdot g_0^3(X)})\mathcal{E}^0_{\Delta Y}(X))\Delta g] \\
&= 0
\end{aligned}
$$

where the first equality holds by definition with $\Delta g \equiv g(X) - g_0(X)$, second equality holds by law of iterated expectation, and the last equality holds by that $E[\Delta Y \mathbf{1}\{D=0\}|X] = \mathcal{E}^0_{\Delta Y}(X)$, $E[m_J^d(D)|X] = f_J^0(d|X)$, and $E[\mathbf{1}\{D=0\}|X] = g_0(X)$.

w.r.t $\mathcal{E}_{\Delta Y}(X)$:

$$
\begin{aligned}
&\partial_r E[\psi_J(\mathcal{E}^0_{\Delta Y}(X) + r(\mathcal{E}_{\Delta Y}(X) - \mathcal{E}^0_{\Delta Y}(X)))]|_{r=0} \\
&= E[\frac{m_J^d(D)g_0(X) - \mathbf{1}\{D=0\}f_J^0(d|X)}{f_d^0 \cdot g_0^2(X)}\Delta\mathcal{E}] \\
&= E[\frac{E[m_J^d(D)|X]g_0(X) - E[\mathbf{1}\{D=0\}|X]f_J^0(d|X)}{f_d^0 \cdot g_0^2(X)}\Delta\mathcal{E}] \\
&= 0
\end{aligned}
$$

where the first line holds by definition with $\Delta\mathcal{E} = \mathcal{E}_{\Delta Y}(X) - \mathcal{E}^0_{\Delta Y}(X)$, the second equality holds by law of iterated expectation, and the last equality holds by the definition that $E[m_J^d(D)|X] = f_J^0(d|X)$ and $E[\mathbf{1}\{D=0\}|X] = g_0(X)$.

This shows that the score ψ_J is Neyman orthogonal wrt the infinite dimensional nuisance parameters. Note that for the repeated cross section case, replace ΔY with $\frac{T-\lambda}{\lambda(1-\lambda)}Y$, the

identical arguments follows. ■

A.7 Proof of Theorem 4.3 (Repeated Outcomes)

Let T_N be the set of square integrable $\eta := (f_J, g(X), \mathcal{E}_{\Delta Y}(X))$ such that assumption 4.7 holds. Let F_N, E_N be the set of $f > 0$ and $\mathcal{E}^d_{\Delta Y}$ such that $|f - f^0_d| \leq (Nh)^{-1/2}$ and $|\mathcal{E}^d_{\Delta Y} - \mathcal{E}^d_{\Delta Y,0}| \leq (Nh)^{-1/2}$. Then assumption 4.7 implies that, with probability tending to 1, $\hat{\eta}_k \in T_N$, $\hat{f}_{d,k} \in F_N$, and $\hat{\mathcal{E}}^d_{\Delta Y} \in E_N$.

Recall that our estimator is $K^{-1}\sum_{k=1}^K \widehat{ATT(d)}_k$ where

$$\widehat{ATT(d)}_k \equiv \frac{1}{n}\sum_{i\in I_k} \underbrace{\hat{\mathcal{E}}^d_{\Delta Y,k}}_{(1)} - \underbrace{\Delta Y_i \mathbf{1}\{D_i = 0\}\frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k(X_i)}}_{(2)}$$
$$\underbrace{- \frac{m^d_J(D_i)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}^2_k(X_i)}\hat{\mathcal{E}}_{\Delta Y,k}(X_i)}_{(3)}$$

We present the proof in three subsections. We focus on (1) in the first part. The second part concerns (2) and (3), which contains majority of the proof. In the last subsection, we combine the previous results and conclude.

Part I: Kernel Regression Results

We first consider (1), $\hat{\mathcal{E}}^d_{\Delta Y,k} \equiv \hat{E}[\Delta Y|D = d]$, which is estimated using kernel (and the density f_d is estimated using the same bandwidth h):

$$\hat{\mathcal{E}}^d_{\Delta Y,k} = \frac{\frac{1}{n}\sum_{i\in I_k} K_h(D_i - d)\Delta Y_i}{\hat{f}_{d,k}}, \quad \text{where} \quad \hat{f}_{d,k} = \frac{1}{n}\sum_{i\in I_k} K_h(D_i - d)$$

Then, with the standard results for kernel regression, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{i=1}^K \hat{\mathcal{E}}^d_{\Delta Y,k} - \mathcal{E}^d_{\Delta Y} \\
&= \frac{1}{N}\sum_{i=1}^N \frac{K_h(D_i - d)\Delta Y_i - E[(K_h(D - d)\Delta Y])}{f_d} \\
&- \frac{E[\Delta Y|D = d]}{f_d}\frac{1}{N}\sum_{i=1}^N K_h(D_i - d) - E[K_h(D - d)]
\end{aligned}$$

$$+ o_p((Nh)^{-1/2}).$$

Part II: Orthogonal Scores

To save notation, let $\hat{\theta}_J$ be defined as

$$\hat{\theta}_J \equiv \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i\in I_k}\Delta Y_i \mathbf{1}\{D_i = 1\}\frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k(X_i)} + \frac{m_J^d(D_i)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k^2(X_i)}\hat{\mathcal{E}}_{\Delta Y,k}(X_i).$$

Then we can decompose the following difference as

$$\hat{\theta}_J - \theta_0 = \underbrace{\hat{\theta}_J - \theta_{0J}}_{(\dagger)} + \underbrace{\theta_{0J} - \theta_0}_{(\dagger\dagger)}$$

where (†) will be our main focus while the bias term (††) will be taken care of by under-smoothing assumption in assumption 4.7.

By definition,

$$\sqrt{N}(\hat{\theta}_J - \theta_{0J}) = \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\psi_J(Z_i, \theta_{0,J}, \hat{f}_{d,k}, \hat{\eta}_k)] \tag{18}$$

where ψ_J is defined as in (15), and $E_{n,k}(f) = \frac{1}{n}\sum_{i\in I_k} f(Z_i)$ denotes the empirical average. Then we have the following decomposition, using Taylor's theorem:

$$\sqrt{N}(\hat{\theta}_J - \theta_{0J}) = \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\psi_J(Z, \theta_{0J}, f_d^0, \hat{\eta}_k)] \tag{19}$$

$$+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_f\psi_J(Z, \theta_{0J}, f_d^0, \hat{\eta}_k)](\hat{f}_{d,k} - f_d^0) \tag{20}$$

$$+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_f^2\psi_J(Z, \theta_{0J}, \bar{f}_k, \hat{\eta}_k)](\hat{f}_{d,k} - f_d^0)^2 \tag{21}$$

where $\bar{f}_k \in (f_d^0, \hat{f}_{d,k})$. This decomposition provides a roadmap for the remaining of the proof in part II. There are roughly four steps. In the first step, we show the second-order term (21) vanishes rapidly and does not contribute to the asymptotic variance. In the second step, we bound first order term (20), which potentially contributes to the asymptotic variance. In step 3, we expand (19) around the nuisance parameter $\hat{\eta}_k$, in which the first order bias

disappears by Neyman orthogonality, and we show the second order terms have no impact on the asymptotics. In the final step, we verify the results used in the first two steps and conclude.

Step 1: Second Order Terms

First, we consider (21). By triangle inequality, we have

$$\begin{aligned}
&|E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},\bar{f}_k,\hat{\eta}_k)] - E[\partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0)]| \\
\leq &\underbrace{|E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},\bar{f}_k,\hat{\eta}_k)] - E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0)]|}_{J_{1k}} \\
+ &\underbrace{|E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0)] - E[\partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0)]|}_{J_{2k}}.
\end{aligned}$$

To bound J_{2k}, note that since f_d^0 is bounded away from zero and the score ψ is bounded by M_J,

$$\partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0) = \frac{2}{(f_d^0)^2}(\psi_J(Z,\theta_{0J},f_d^0,\eta_0)+\theta_{0J})$$

which implies that

$$E[J_{2k}^2] \leq \frac{1}{N}E[(\partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0))^2] \lesssim M_J^2/N$$

and by Markov's inequality, we have $J_{2k} \leq O_p(M_J/\sqrt{N})$. For J_{1k}, we have

$$\begin{aligned}
E[J_{1k}^2|I_k^c] &= E[|E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},\bar{f}_k,\hat{\eta}_k)] - E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0)]|^2|I_k^c] \\
&\leq \sup_{f\in F_N,\eta\in T_N} E[|\partial_f^2\psi_J(Z,\theta_{0J},f,\eta) - \partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0)|^2|I_k^c] \\
&\leq \sup_{f\in F_N,\eta\in T_N} E[|\partial_f^2\psi_J(Z,\theta_{0J},f,\eta) - \partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0)|^2] \\
&\lesssim M_J^2\varepsilon_N^2 \qquad \text{(a)}
\end{aligned}$$

Then by conditional Markov's inequality, $(\hat{f}_{d,k} - f_d^0)^2 \leq O_p((Nh)^{-1})$, and assumption 4.5, we conclude that (21) $= o_p(1)$. We will show (a) at the end of this section.

Step 2: First Order Terms

To bound (20), we use first the triangle inequality to obtain the decomposition

$$\begin{aligned}
&|E_{n,k}[\partial_f\psi_J(Z,\theta_{0J},f_d^0,\hat{\eta}_k)] - E[\partial_f\psi_J(Z,\theta_{0J},f_d^0,\eta_0)]| \\
\leq &\underbrace{|E_{n,k}[\partial_f\psi_J(Z,\theta_{0J},f_d^0,\hat{\eta}_k)] - E_{n,k}[\partial_f\psi_J(Z,\theta_{0J},f_d^0,\eta_0)]|}_{J_{3k}}
\end{aligned}$$

$$+\underbrace{|E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)] - E[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)]|}_{J_{4k}}.$$

We first bound J_{4k}. Note that since f_d^0 is bounded away from zero and the score ψ is bounded by M_J, we have

$$\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0) = -\frac{1}{f_d^0}(\psi_J(Z, \theta_{0J}, f_d^0, \eta_0) + \theta_{0J})$$

which implies that

$$E[J_{4k}^2] \leq \frac{1}{N} E[(\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0))^2] \lesssim M_J^2/N$$

and by Markov's inequality, we have $J_{4k} \leq O_p(M_J/\sqrt{N})$. With the assumption that $M_J/\sqrt{N} = o(1)$, we have $J_{4k} = o_p(1)$.

Second, to bound J_{3k}, note that

$$\begin{aligned}
E[J_{3k}^2 | I_k^c] &= E[|E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \hat{\eta}_k)] - E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)]|^2 | I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta) - \partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)|^2 | I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta) - \partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)|^2] \\
&\lesssim M_J^2 \varepsilon_N^2 \qquad \text{(b)}
\end{aligned}$$

where the first equation holds by definition, the second line holds by Cauchy-Schwarz and the third line holds by the construction that all the parameters are estimated using auxiliary sample I_k^c. Then we conclude with the conditional Markov's inequality that $J_{3k} = o_p(1)$. Therefore,

$$E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \hat{\eta}_k)] \to^p E[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)] \equiv S_f^0$$

Note that under the assumption, $(\hat{f}_{d,k} - f_d^0) = O_p((Nh)^{-1/2})$, we can rewrite (20) as

$$\begin{aligned}
(20) &= \sqrt{N} \frac{1}{K} \sum_{k=1}^{K} E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \hat{\eta}_k)](\hat{f}_{d,k} - f_d^0) \\
&= \sqrt{N} \frac{1}{K} \sum_{k=1}^{K} S_f^0(\hat{f}_{d,k} - f_d^0) + o_p(h^{-1/2}) \\
&= \sqrt{N} \frac{1}{N} \sum_{i=1}^{N} S_f^0(K_h(D_i - d) - E[K_h(D - d)]) + o_p(h^{-1/2})
\end{aligned}$$

where the last equality holds by the definition that $\hat{f}_{d,k} - f_d^0 = (N-n)^{-1}\sum_{i\in I_k^c} K_h(D_i - d) - E[K_h(D-d)] + O(h^2)$, the undersmoothing assumption that $\sqrt{N}h^2 \leq O(1)$, and the fact that $K^{-1}\sum_{k=1}^K(\hat{f}_{d,k} - E[K_h(D-d)]) = \frac{1}{N}\sum_{i=1}^N(K_h(D_i-d) - E[K_h(D-d)])$.

Step 3: "Neyman Term"

Now we consider (19), which we can rewrite as

$$\begin{aligned}
&\sqrt{N}\frac{1}{K}\sum_{k=1}^K E_{n,k}[\psi_J(Z,\theta_{0J},f_d^0,\hat{\eta}_k)] \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^N \psi_J(Z_i,\theta_{0J},f_d^0,\eta_0) \\
&+ \sqrt{N}\frac{1}{K}\sum_{k=1}^K \underbrace{(E_{n,k}[\psi_J(Z,\theta_{0J},f_d^0,\hat{\eta}_k)] - E_{n,k}[\psi_J(Z_i,\theta_{0J},f_d^0,\eta_0)])}_{R_{nk}}
\end{aligned}$$

Since K is fixed, $n = O(N)$, it suffices to show that $R_{nk} = o_p(N^{-1/2}M_J)$, so it vanishes when scaled by the (square root of) asymptotic variance. Note that by triangle inequality, we have the following decomposition

$$|R_{n,k}| \leq \frac{R_{1k} + R_{2k}}{\sqrt{n}}$$

where

$$R_{1k} \equiv |G_{nk}[\psi_J(Z,\theta_{0J},f_d^0,\hat{\eta}_k)] - G_{nk}[\psi_J(Z,\theta_{0J},f_d^0,\eta_0)]|$$

with $G_{nk}(f) = \sqrt{n}(P_n - P)(f)$ denote the empirical process, and with some abuse of notation, it will also be used to denote conditional version of the empirical process conditioning on the auxiliary sample I_k^c. Moreover,

$$R_{2k} \equiv \sqrt{n}|E[\psi_J(Z,\theta_{0J},f_d^0,\hat{\eta}_k)|I_k^c] - E[\psi_J(Z,\theta_{0J},f_d^0,\eta_0)]|.$$

For simplicity, let's suppress other arguments in ψ and denote $\psi_\eta^i \equiv \psi_J(Z_i,\theta_{0J},f_d^0,\eta)$.

First, we consider R_{1k}, in which

$$G_{nk}\psi_{\hat{\eta}_k} - G_{nk}\psi_{\eta_0} = \sqrt{n}\frac{1}{n}\sum_{i=1}^n \underbrace{\psi_{\hat{\eta}_k}^i - \psi_{\eta_0}^i - E[\psi_{\hat{\eta}_k}^i|I_k^c] - E[\psi_{\eta_0}^i]}_{\equiv\Delta_{ik}}$$

In particular, it can be shown that $E[\Delta_{ik}\Delta_{jk}] = 0$ for all $i \neq j$ using the i.i.d. assumption of the data and that the nuisance parameter $\hat{\eta}_k$ is estimated using the auxiliary sample.

Then, we have

$$\begin{aligned} E[R_{1k}^2|I_k^c] &\leq E[\Delta_{ik}^2|I_k^c] \\ &\leq E[(\psi_{\hat{\eta}_k}^i - \psi_{\eta_0}^i)^2|I_k^c] \\ &\leq \sup_{\eta \in T_N} E[(\psi_{\eta}^i - \psi_{\eta_0}^i)^2|I_k^c] \\ &\leq \sup_{\eta \in T_N} E[(\psi_{\eta}^i - \psi_{\eta_0}^i)^2] \\ &\lesssim M_J^2 \varepsilon_N^2 \qquad (c) \end{aligned}$$

and using the conditional Markov's inequality, we conclude that $R_{1k} = o_p(M_J)$. Now we bound R_{2k}. Note that by definition of the score, $E[\psi_J(Z, \theta_{0J}, f_d^0, \eta_0)] = 0$, so it suffices to bound $E[\psi_J(Z, \theta_{0J}, f_d^0, \hat{\eta}_k)|I_k^c]$. Suppressing other arguments in the score, define

$$h_k(r) \equiv E[\psi_J(\eta_0 + r(\hat{\eta}_k - \eta_0))|I_k^c]$$

where by definition $h_k(0) = E[\psi_J(\eta_0)|I_k^c] = 0$ and $h_k(1) = E[\psi_J(\hat{\eta}_k)|I_k^c]$. Use Taylor's theorem, expand $h_k(1)$ around 0, we have

$$h_k(1) = h_k(0) + h_k'(0) + \frac{1}{2} h_k''(\bar{r}), \quad \bar{r} \in (0,1).$$

Note that, by Neyman orthogonality,

$$h_k'(0) = \partial_\eta E[\psi_J(\eta_0)][\hat{\eta}_k - \eta_0] = 0$$

and use that fact that $h_k(0) = 0$, we have

$$\begin{aligned} R_{2k} &= \sqrt{n}|h_k(1)| = \sqrt{n}|h_k''(\bar{r})| \\ &\leq \sup_{r \in (0,1), \eta \in T_N} \sqrt{n}|\partial_r^2 E[\psi_J(\eta_0 + r(\hat{\eta}_k - \eta_0))]| \\ &\lesssim \sqrt{n} M_J \varepsilon_N^2 \qquad (d) \end{aligned}$$

Combining above results, we conclude that

$$\sqrt{N} R_{n,k} \lesssim M_J \varepsilon_N + \sqrt{N} M_J \varepsilon_N^2.$$

and for $\varepsilon_N = o(N^{-1/4})$, we have $\sqrt{N} R_{n,k} = o_p(M_J)$.

Step 4: Auxiliary Results

In this section, we show the auxiliary results (a)-(d) used in the previous steps. We first

show (c) as it will also be used to bound other results.

Recall that

$$(c): \qquad \sup_{\eta \in T_N} E[(\psi_\eta - \psi_{\eta_0})^2] \lesssim M_J^2 \varepsilon_N^2.$$

By definition,

$$\begin{aligned}
&\psi_\eta - \psi_{\eta_0} \\
&= \Delta Y \mathbf{1}\{D=0\} \frac{f_J(X)}{f_d^0 \cdot g(X)} + \frac{m_J(D) g(X) - \mathbf{1}\{D=0\} f_J(X)}{f_d^0 \cdot g^2(X)} \mathcal{E}_{\Delta Y}(X) \\
&- \Delta Y \mathbf{1}\{D=0\} \frac{f_J^0(X)}{f_d^0 \cdot g_0(X)} - \frac{m_J(D) g_0(X) - \mathbf{1}\{D=0\} f_J^0(X)}{f_d^0 \cdot g_0^2(X)} \mathcal{E}_{\Delta Y}^0(X) \\
&= \frac{\Delta Y \mathbf{1}\{D=0\}}{f_d^0} \left(\frac{f_J(X)}{g(X)} - \frac{f_J^0(X)}{g_0(X)}\right) + \frac{m_J(D)}{f_d^0} \left(\frac{\mathcal{E}_{\Delta Y}(X)}{g(X)} - \frac{\mathcal{E}_{\Delta Y}^0(X)}{g_0(X)}\right) \\
&- \frac{\mathbf{1}\{D=0\}}{f_d^0} \left(\frac{f_J(X) \mathcal{E}_{\Delta Y}(X)}{g^2(X)} - \frac{f_J^0(X) \mathcal{E}_{\Delta Y}^0(X)}{g_0^2(X)}\right) \\
&\lesssim C_1 (f_J(X) - f_J^0(X)) + C_2 M_J (g(X) - g_0(X)) + C_3 M_J (\mathcal{E}_{\Delta Y}(X) - \mathcal{E}_{\Delta Y}^0(X))
\end{aligned}$$

where the last line can be shown using the usual plus-minus trick with C_1, C_2, C_3 being some constants and $M_J = \|m_J\|_\infty$. Then by the definition of T_N and the assumptions on the rate of convergence of the nuisance parameters,

$$\begin{aligned}
\sup_{\eta \in T_N} E[(\psi_\eta - \psi_{\eta_0})^2] &\lesssim \|f_J - f_J^0\|_{P,2}^2 + M_J^2 \|g - g_0\|_{P,2}^2 + M_J^2 \|\mathcal{E}_{\Delta Y} - \mathcal{E}_{\Delta Y}^0\|_{P,2}^2 \\
&+ M_J \|f_J - f_J^0\|_{P,2} \|g - g_0\|_{P,2} + M_J \|f_J - f_J^0\|_{P,2} \|\mathcal{E}_{\Delta Y} - \mathcal{E}_{\Delta Y}^0\|_{P,2} \\
&+ M_J^2 \|g - g_0\|_{P,2} \|\mathcal{E}_{\Delta Y} - \mathcal{E}_{\Delta Y}^0\|_{P,2} \\
&\lesssim M_J^2 \varepsilon_N^2
\end{aligned}$$

This shows (c) with $\varepsilon_N = o(N^{-1/4})$.

Next, we consider (a). We want to show

$$(a): \qquad \sup_{f \in F_N, \eta \in T_N} E[|\partial_f^2 \psi_J(Z, \theta_{0J}, f, \eta) - \partial_f^2 \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)|^2] \lesssim \varepsilon_N^2$$

By definition,

$$\begin{aligned}
\partial_f^2 \psi_J(Z, \theta_{0J}, f, \eta) &= \frac{2}{f^2} (\psi_J(Z, \theta_{0J}, f, \eta) + \theta_{0J}) \\
\partial_f^3 \psi_J(Z, \theta_{0J}, f, \eta) &= -\frac{6}{f^3} (\psi_J(Z, \theta_{0J}, f, \eta) + \theta_{0J}).
\end{aligned}$$

Then using Taylor's theorem expand around f_d^0, we

$$\begin{aligned}
&\partial_f^2\psi_J(Z,\theta_{0J},f,\eta) - \partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0) \\
&= \partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta) - \partial_f^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0) + \partial_f^3\psi_J(Z,\theta_{0J},\bar{f},\eta)(f-f_d^0) \\
&= \frac{2}{(f_d^0)^2}(\psi_J(Z,\theta_{0J},f_d^0,\eta) - \psi_J(Z,\theta_{0J},f_d^0,\eta_0)) \quad (\star) \\
&- \frac{6}{\bar{f}^3}(\psi_J(Z,\theta_{0J},\bar{f},\eta) + \theta_{0J})(f-f_d^0) \quad (\star\star)
\end{aligned}$$

By the assumption, on F_N, $\bar{f}$ and f_d^0 are bounded away from zero, so that $(\star)$ is the leading term that can be bounded with (c). Moreover, for $\varepsilon_N = o(N^{-1/4})$, $(\star\star)$ is of smaller order and can be ignored. Therefore we conclude that

$$\sup_{f\in F_N,\eta\in T_N} E[|\partial_f^2\psi_J(Z,\theta_{0J},f,\eta) - \partial_p^2\psi_J(Z,\theta_{0J},f_d^0,\eta_0)|^2] \lesssim M_J^2\varepsilon_N^2.$$

Similarly, by definition,

$$\begin{aligned}
&\partial_f\psi_J(Z,\theta_{0J},f_d^0,\eta) - \partial_f\psi_J(Z,\theta_{0J},f_d^0,\eta_0) \\
&= -\frac{1}{f_d^0}(\psi_J(Z,\theta_{0J},f_d^0,\eta) - \psi_J(Z,\theta_{0J},f_d^0,\eta_0))
\end{aligned}$$

and using the same arguments as before, (b) follows from (a) and (c).

Last, we show (d). It suffices to show

$$\sup_{r\in(0,1),\eta\in T_N} |\partial_r^2 E[\psi_J(\eta_0 + r(\hat{\eta}_k - \eta_0))]| \lesssim M_J\varepsilon_N^2.$$

By definition,

$$\begin{aligned}
&\psi_J(\eta_0 + r(\hat{\eta}_k - \eta_0)) \\
&= \frac{\Delta Y \mathbf{1}\{D=0\}(f_J^0 + r(f_J - f_J^0))}{f_d^0\cdot(g_0 + r(g-g_0))} - \theta_{0J} \\
&+ \frac{1}{f_d^0}\left(\frac{m_J}{g_0 + r(g-g_0)} - \frac{\mathbf{1}\{D=0\}(f_J^0 + r(f_J - f_J^0))}{(g_0 + r(g-g_0))^2}\right)(\mathcal{E}_{\Delta_Y}^0 + r(\mathcal{E}_{\Delta_Y} - \mathcal{E}_{\Delta_Y}^0))
\end{aligned}$$

and we take the second order partial derivatives wrt r term by term. For simplicity, we omit the derivations, and we have

$$\begin{aligned}
&\partial_r^2\psi_J(\eta_0 + r(\hat{\eta}_k - \eta_0)) \\
&\asymp C_1\Delta_f\Delta_g + C_2(\Delta_g)^2 + C_3M_J\Delta_{\mathcal{E}}\Delta_g + C_4M_J(\Delta_g)^2 + C_5\Delta_f\Delta_{\mathcal{E}} + C_6\Delta_{\mathcal{E}}\Delta_g
\end{aligned}$$

where $\Delta_f \equiv f_J - f_J^0$, $\Delta_g \equiv g - g_0$, and $\Delta_{\mathcal{E}} \equiv \mathcal{E}_{\Delta Y} - \mathcal{E}^0_{\Delta Y}$. Then by triangle inequality, Cauchy-Schwarz, and the assumption on the space of nuisance parameters T_N, we conclude

$$\begin{aligned}\partial_r^2 E[\psi_J(\eta_0 + r(\hat{\eta}_k - \eta_0))] &\lesssim \|f_J - f_J^0\|_{P,2}\|g - g_0\|_{P,2} + \|f_J - f_J^0\|_{P,2}\|\mathcal{E}_{\Delta Y} - \mathcal{E}^0_{\Delta Y}\|_{P,2} \\ &\quad + M_J\|g - g_0\|_{P,2}\|\mathcal{E}_{\Delta Y} - \mathcal{E}^0_{\Delta Y}\|_{P,2} + M_J\|g - g_0\|^2_{P,2} \\ &\lesssim M_J\varepsilon_N^2.\end{aligned}$$

Part III: Conclusion

Combining the results in Part I and Part II, we have

$$\begin{aligned}&\widehat{ATT(d)} - ATT(d) \\ &= \frac{1}{N}\sum_{i=1}^N \frac{K_h(D_i - d)\Delta Y_i - E[(K_h(D-d)\Delta Y]}{f_d^0} \qquad ① \\ &- \frac{E[\Delta Y|D=d]}{f_d^0}\frac{1}{N}\sum_{i=1}^N (K_h(D_i - d) - E[K_h(D-d)]) \qquad ② \\ &- \frac{1}{N}\sum_{i=1}^N \psi_J(Z_i, \theta_{0J}, f_d^0, \eta_0) \qquad ③ \\ &- \frac{1}{N}\sum_{i=1}^N S_f^0(K_h(D_i - d) - E[K_h(D_i - d)]) \qquad ④ \\ &+ o_p((Nh)^{-1/2}) + o_p(N^{-1/2}M_J) \qquad ⑤ \\ &+ \theta_0 - \theta_{0J} \qquad ⑥\end{aligned}$$

where each of ①-④ is an average of i.i.d zero-mean terms with the variance growing either with kernel bandwidth h or the series term J.

Since J and h grows with N, we need a triangular array CLT to establish the asymptotic results. The Lyapunov conditions are easy to verify for the kernel terms ①,②,④. Note that we assumed that $m_J^d(D) = \sum_{j=1}^J \phi_j(D)\phi_j(d)\mathbf{1}\{D > 0\}$ satsifies $\|m_J^d(D)\|_\infty \leq M_J$. If $E[\psi_J^2] \asymp E[(m_J^d(D))^2] \asymp M_J^2$ and $E[\psi_J^3] \asymp M_J^3$, then the Lyapunov condition is also satisfied for ③. Then by CLT, together with assumption 4.7, we have

$$\frac{\widehat{ATT(d)} - ATT(d)}{\sigma_N/\sqrt{N}} \quad \to^d \quad N(0,1)$$

with σ_N defined by

$$\sigma_N^2 \equiv E[\big(\frac{1}{f_d^0}(K_h(D-d)\Delta Y - E[K_h(D-d)\Delta Y]) \\ - \psi_J + (\frac{\theta_J}{f_d^0} - \frac{\mathcal{E}_{\Delta Y}^d}{f_d^0})(K_h(D-d) - E[K_h(D-d)])\big)^2]$$

where we have used the fact that $S_f^0 = -\theta_J/f_d^0$. ■

B Supplementary Material

First, we extend our results to the repeated cross sections setting.

Algorithm B.1 (CDID Estimator)**.** *Let $\{I_k\}_{k=1}^K$ denote a random partition of a randome sample $\{Z_i\}_{i=1}^N$, each with equal size $n = N/K$, and for each $k \in \{1, \cdots, K\}$, let $I_k^c \equiv N \setminus I_k$ denote the complement.*

- *(Repeated Cross Sections) For each k, construct*

$$\widehat{ATT(d)}_k \equiv \frac{1}{n}\sum_{i \in I_k} \hat{\mathcal{E}}_{\lambda Y,k}^d - \frac{T_i - \hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)} Y_i \mathbf{1}\{D_i = 0\} \frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k} \cdot \hat{g}_k(X_i)} \\ - \frac{m_J^d(D_i)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k} \cdot \hat{g}_k^2(X_i)} \hat{\mathcal{E}}_{\lambda Y,k}(X_i)$$

  where $\hat{\lambda}_k, \hat{f}_{d,k}, \hat{\mathcal{E}}_{\lambda Y,k}^d, \hat{f}_{J,k}, \hat{g}_k, \hat{\mathcal{E}}_{\lambda Y,k}$ are the estimators of λ, f_d, $E[\frac{T-\lambda}{\lambda(1-\lambda)}Y|D=d]$, $f_J(d|X)$, $g(X)$ and $\mathcal{E}_{\lambda Y}(X)$ respectively using the rest of the sample I_k^c.

- *Average through k to obtain the final estimator*

$$\widehat{ATT(d)} \equiv \frac{1}{K}\sum_{k=1}^K \widehat{ATT(d)}_k.$$

Analogous to the repeated outcomes setting, we make the following assumptions.

Assumption B.1 (Bounds)**.**

(i) $m_J^d(D) = \sum_{j=1}^J \phi_j(D)\phi_j(d)\mathbf{1}\{D > 0\}$ are bounded by some constant M_J that grows with J such that $M_J/\sqrt{N} = o(1)$;

(ii) for some constants $0 < c < 1$ and $0 < C < \infty$, $f_d > c$, $c < \lambda < 1 - c$, $|E[\frac{T-\lambda}{\lambda(1-\lambda)}Y|D = d]| < C$, and $|\mathcal{E}_{\lambda Y}(X)| < C$ almost surely;

(iii) *for some constants* $0 < \kappa < \frac{1}{2}$ *and for all* $J \geq 1$, $\kappa < f_J(d|X), g(X) < 1 - \kappa$ *almost surely;*

(iv) f_d *and* $E[\frac{\lambda - T}{\lambda(1-\lambda)} Y | D = d]$ *are twice continuously differentiable at* $D = d \in (d_L, d_H)$ *and have bounded second derivative.*

Assumption B.2 (Rates)**.**

(i) *kernel bandwith satisfies* $Nh \to \infty$ *and* $\sqrt{Nh^5} = o(1)$ *and*

$$\frac{\sqrt{N}}{\max\{M_J, h^{-\frac{1}{2}}\}} E[|\sum_{j=J+1}^{\infty} E[\phi_j(D)\mathbf{1}\{D > 0\}|X]\phi_j(d)|] = o(1);$$

(ii) *with probability tending to* 1, $\|\hat{f}_J - f_J(d|X)\|_{P,2} \leq M_J \varepsilon_N$, $\|\hat{g}(X) - g(X)\|_{P,2} \leq \varepsilon_N$, $\|\hat{\mathcal{E}}_{\lambda Y}(X) - \mathcal{E}_{\lambda Y}(X)\|_{P,2} \leq \varepsilon_N$*;*

(iii) *with probability tending to* 1, $\|\hat{\mathcal{E}}_{\lambda Y}(X)\|_{P,\infty} < C$, $\kappa < \|\hat{f}_J(X)\|_{P,\infty} < 1 - \kappa$, *and* $\kappa < \|\hat{g}(X)\|_{P,\infty} < 1 - \kappa$.

Theorem B.1 (Repeated Cross Sections)**.** *Suppose assumptions 4.2, 4.3, 4.4, 4.6, B.1, and B.2 hold. Then for* $\varepsilon_N = o(N^{-1/4})$,

$$\frac{\widehat{ATT}(d) - ATT(d)}{\sigma_N / \sqrt{N}} \quad \to^d \quad N(0,1)$$

where

$$\sigma_N^2 \equiv E[\Big(\frac{1}{f_d}(K_h(D-d)Y^\lambda - E[K_h(D-d)Y^\lambda]) \\ - \psi_J + (\frac{\theta_J}{f_d} - \frac{\mathcal{E}_{\lambda Y}^d}{f_d})(K_h(D-d) - E[K_h(D-d)])\Big)^2].$$

and ψ_J *is defined as in (16) and* $Y^\lambda \equiv \frac{T - \lambda}{\lambda(1-\lambda)} Y$.

Similarly as before, we construct

$$\hat{\sigma}_N^2 := \frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\Big(\frac{1}{\hat{f}_{d,k}}(K_h(D-d)Y^{\hat{\lambda}_k} - E_{n^c,k}[K_h(D-d)Y^{\hat{\lambda}_k}]) \\ - \psi_J(Z, \hat{\theta}_J, \hat{\lambda}_k, \hat{f}_{d,k}, \hat{\eta}_k) \\ + (\frac{\hat{\theta}_J}{\hat{f}_{d,k}} - \frac{\hat{\mathcal{E}}_{\lambda Y,k}^d}{\hat{f}_{d,k}})(K_h(D-d) - E_{n^c,k}[K_h(D-d)])\Big)^2]$$

where

$$\hat{\theta}_J \equiv \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\frac{T_i-\hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)}Y_i\mathbf{1}\{D_i=0\}\frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k(X_i)}$$
$$+\frac{m_J^d(D_i)\hat{g}_k(X_i)-\mathbf{1}\{D_i=0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k^2(X_i)}\hat{\mathcal{E}}_{\lambda Y,k}(X_i),$$

$Y^{\hat{\lambda}_k}\equiv\frac{T-\hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)}Y$, and $E_{n^c,k}$ denotes the empirical average using the auxiliary sample I_k^c. Then, the $1-\alpha$ confidence interval can be constructed as $[\widehat{ATT}(d)-z_{1-\alpha/2}\hat{\sigma}_N/\sqrt{N},\widehat{ATT}(d)+z_{1-\alpha/2}\hat{\sigma}_N/\sqrt{N}]$ where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal random variable.

Alternatively, one can use a multiplier bootstrap type of procedure to construct the confidence interval for our estimator. Specifically, let $\{\xi_i\}_{i=1}^N$ be an i.i.d. sequence of sub-exponential random variables independent of $\{Y_i,T_i,D_i,X_i\}_{i=1}^N$ such that $E[\xi_i]=E[\xi_i^2]=1$. Then for each $b=1,\cdots,B$, we draw such a sequence $\{\xi_i\}_{i=1}^N$ and construct

$$\widehat{ATT(d)}_b^* \equiv \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\xi_i\Big(\hat{\mathcal{E}}_{\lambda Y,k}^d-\frac{T_i-\hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)}Y_i\mathbf{1}\{D_i=0\}\frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k(X_i)}$$
$$-\frac{m_J^d(D_i)\hat{g}_k(X_i)-\mathbf{1}\{D_i=0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k^2(X_i)}\hat{\mathcal{E}}_{\lambda Y,k}(X_i)\Big)$$

Let $\hat{c}_\alpha$ be the α's quantile of $\{\widehat{ATT(d)}_b^*-\widehat{ATT(d)}\}_{b=1}^B$, and we construct the confidence interval as $[\widehat{ATT(d)}-\hat{c}_{1-\alpha/2},\widehat{ATT(d)}-\hat{c}_{\alpha/2}]$.

B.1 Proof of Theorem B.1 (Repeated Cross Sections)

The proof for the repeated cross sections case follows very closely to that of the repeated outcomes case, with only minor modifications due to the presence of a new parameter $\lambda=P(T=1)$, which can be estimated at parametric rate.

Let T_N be the set of square integrable $\eta:=(f_J,g(X),\mathcal{E}_{\lambda Y}(X))$ such that assumption B.2 holds. Let F_N,E_N be the set of $f>0$ and $\mathcal{E}_{\lambda Y}^d$ such that $|f-f_d^0|\leq(Nh)^{-1/2}$ and $|\mathcal{E}_{\lambda Y}^d-\mathcal{E}_{\lambda Y,0}^d|\leq(Nh)^{-1/2}$. Then assumption B.2 implies that, with probability tending to 1, $\hat{\eta}_k\in T_N$, $\hat{f}_{d,k}\in F_N$, $\hat{\lambda}_k\in P_N$, and $\hat{\mathcal{E}}_{\lambda Y}^d\in E_N$.

First, recall that for $1 \leq k \leq K$,

$$\widehat{ATT(d)}_k \equiv \frac{1}{n}\sum_{i\in I_k} \underbrace{\hat{\mathcal{E}}^d_{\lambda Y,k}}_{(1)} - \underbrace{\frac{T_i - \hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)} Y_i \mathbf{1}\{D_i = 0\} \frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot \hat{g}_k(X_i)}}_{(2)}$$
$$\underbrace{- \frac{m^d_J(D_i)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot \hat{g}^2_k(X_i)} \hat{\mathcal{E}}_{\lambda Y,k}(X_i)}_{(3)}$$

We first focus on (1), and then on (2) and (3).

Part I: Kernel Regression Results

We first consider the (1), $\hat{\mathcal{E}}^d_{\lambda Y,k} \equiv \hat{E}[\frac{T-\lambda}{\lambda(1-\lambda)}Y|D=d]$, which is estimated using kernel (and the density f_d is estimated using the same bandwidth h):

$$\hat{\mathcal{E}}^d_{\lambda Y,k} = \frac{\frac{1}{n}\sum_{i\in I_k} K_h(D_i - d)\frac{T_i - \hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)} Y_i}{\hat{f}_{d,k}}$$

where

$$\hat{f}_{d,k} = \frac{1}{n}\sum_{i\in I_k} K_h(D_i - d); \quad \hat{\lambda}_k = \frac{1}{n}\sum_{i\in I_k} T_i.$$

For notation simplicity, denote $Y^\lambda \equiv \frac{T-\lambda}{\lambda(1-\lambda)}Y$. Then using the similar arguments as before, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{i=1}^K \hat{\mathcal{E}}^d_{\lambda Y,k} - \mathcal{E}^d_{\lambda Y} \\
&= \frac{1}{N}\sum_{i=1}^N \frac{K_h(D_i - d)Y^\lambda_i - E[(K_h(D-d)Y^\lambda])}{f_d} \\
&- \frac{E[Y^\lambda|D=d]}{f_d}\frac{1}{N}\sum_{i=1}^N K_h(D_i - d) - E[K_h(D-d)] \\
&+ o_p((Nh)^{-1/2}).
\end{aligned}$$

Part II: Orthogonal Scores

Let $\hat{\theta}_J$ be defined as

$$\hat{\theta}_J \equiv \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i\in I_k}\frac{T_i-\hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)}Y_i\mathbf{1}\{D_i=1\}\frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k(X_i)} + \frac{m_J^d(D_i)\hat{g}_k(X_i)-\mathbf{1}\{D_i=0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot\hat{g}_k^2(X_i)}\hat{\mathcal{E}}_{\lambda Y,k}(X_i).$$

Then by definition,

$$\sqrt{N}(\hat{\theta}_J-\theta_{0J}) = \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\psi_J(Z_i,\theta_{0,J},\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k)] \tag{22}$$

where ψ_J is defined as in (16), and $E_{n,k}(f)=\frac{1}{n}\sum_{i\in I_k}f(Z_i)$ denotes the empirical average. Then by multivariate version of Taylor's theorem,

$$\sqrt{N}(\hat{\theta}_J-\theta_{0J}) = \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)] \tag{23}$$

$$+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_\lambda\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)](\hat{\lambda}_k-\lambda_0) \tag{24}$$

$$+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_f\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)](\hat{f}_{d,k}-f_d^0) \tag{25}$$

$$+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_\lambda^2\psi_J(Z,\theta_{0J},\bar{\lambda}_k,\bar{f}_k,\hat{\eta}_k)](\hat{\lambda}_k-\lambda_0)^2 \tag{26}$$

$$+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},\bar{\lambda}_k,\bar{f}_k,\hat{\eta}_k)](\hat{f}_{d,k}-f_d^0)^2 \tag{27}$$

$$+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_\lambda\partial_f\psi_J(Z,\theta_{0J},\bar{\lambda}_k,\bar{f}_k,\hat{\eta}_k)](\hat{f}_{d,k}-f_d^0)(\hat{\lambda}_k-\lambda_0) \tag{28}$$

where $\bar{\lambda}_k\in(\lambda_0,\hat{\lambda}_k)$ and $\bar{f}_k\in(f_d^0,\hat{f}_{d,k})$. All the second order terms (26)-(28) can be shown to be $o_p(1)$. The first order term (25) can be analyzed in the same way as the repeat outcomes case. Moreover, since $\hat{\lambda}_k=E_{n,k}T_i$ converges at parametric rate while the kernel estimator $\hat{f}_{d,k}$ converges at slower rate, the influence of (24) on the asymptotic variance is negligible. The main term (23) can be analyzed in the same way as in the repeated outcomes case.

Step 1: Second Order Terms

First, we consider (26). By triangle inequality, we have

$$\begin{aligned}
&|E_{n,k}[\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)] - E[\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]| \\
\leq &\underbrace{|E_{n,k}[\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)] - E_{n,k}[\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|}_{J_{1k}} \\
+ &\underbrace{|E_{n,k}[\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)] - E[\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|}_{J_{2k}}
\end{aligned}$$

To bound J_{2k}, since $0 < c < \lambda_0 < 1 - c$ and the score ψ is bounded by M_J, we have

$$\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0) \lesssim M_J$$

and hence

$$E[J_{2k}^2] \leq \frac{1}{N} E[(\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0))^2] \lesssim M_J^2/N.$$

Then by Markov's inequality, we have $J_{2k} \leq O_p(M_J/\sqrt{N})$. For J_{1k}, note that

$$\begin{aligned}
E[J_{1k}^2 | I_k^c] &= E[|E_{n,k}[\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)] - E_{n,k}[\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|^2 | I_k^c] \\
&\leq \sup_{\substack{\lambda \in P_N, \ f \in F_N \\ \eta \in T_N}} E[|\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda, f, \eta) - \partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2 | I_k^c] \\
&\leq \sup_{\substack{\lambda \in P_N, \ f \in F_N \\ \eta \in T_N}} E[|\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda, f, \eta) - \partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim M_J^2 \varepsilon_N^2 \qquad \text{(a)}
\end{aligned}$$

where the first equation holds by definition, the second line holds by Cauchy-Schwarz and the third line holds by the construction that all the parameters are estimated using auxiliary sample I_k^c. Then by conditional Markov's inequality, $(\hat{\lambda}_k - \lambda)^2 \leq O_p(N^{-1})$, and assumption B.1, we conclude that (26) $= o_p(1)$. We will show (a) at the end of this section.

Term (27) is bounded in the same way as the repeated outcomes case. By triangle inequality, we have

$$\begin{aligned}
&|E_{n,k}[\partial_f^2 \psi_J(Z, \theta_{0J}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)] - E[\partial_f^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]| \\
\leq &\underbrace{|E_{n,k}[\partial_f^2 \psi_J(Z, \theta_{0J}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)] - E_{n,k}[\partial_f^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|}_{J_{3k}}
\end{aligned}$$

$$+\underbrace{|E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)]-E[\partial_f^2\psi_J(Z,\theta_{0J},\lambda_0 f_d^0,\eta_0)]|}_{J_{4k}}.$$

To bound J_{4k}, note that since f_d^0 is bounded away from zero and the score ψ is bounded by M_J,

$$\partial_f^2\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)=\frac{2}{(f_d^0)^2}(\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)+\theta_{0J})\lesssim M_J$$

which implies that

$$E[J_{4k}^2]\leq\frac{1}{N}E[(\partial_f^2\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0))^2]\lesssim M_J^2/N$$

and by Markov's inequality, we have $J_{4k}\leq O_p(M_J/\sqrt{N})$. For J_{3k}, we have

$$\begin{aligned}
E[J_{3k}^2|I_k^c]&=E[|E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},\bar{\lambda}_k,\bar{f}_k,\hat{\eta}_k)]-E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)]|^2|I_k^c]\\
&\leq\sup_{\substack{\lambda\in P_N,\ f\in F_N\\ \eta\in T_N}}E[|\partial_f^2\psi_J(Z,\theta_{0J},\lambda,f,\eta)-\partial_f^2\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)|^2|I_k^c]\\
&\leq\sup_{\substack{\lambda\in P_N,\ f\in F_N\\ \eta\in T_N}}E[|\partial_f^2\psi_J(Z,\theta_{0J},\lambda,f,\eta)-\partial_f^2\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)|^2]\\
&\lesssim M_J^2\varepsilon_N^2\qquad\text{(b)}
\end{aligned}$$

Then by conditional Markov's inequality, $(\hat{f}_{d,k}-f_d^0)^2\leq O_p((Nh)^{-1})$, and assumption B.1, we conclude that (27) $=o_p(1)$. We postpone the proof of (b) to the end of this section.

Finally, we can bound (28) using similar arguments as those for (26) and (27). To avoid repetitiveness, we only highlight the difference. In particular, we need

$$\sup_{\substack{\lambda\in P_N,\ f\in F_N\\ \eta\in T_N}}E[|\partial_\lambda\partial_f\psi_J(Z,\theta_{0J},\bar{\lambda}_k,\bar{f}_k,\hat{\eta}_k)-\partial_\lambda\partial_f\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)|^2]\lesssim M_J^2\varepsilon_N^2\qquad\text{(c)}$$

and using conditional Markov's inequality, $(\hat{f}_{d,k}-f_d)(\hat{\lambda}_k-\lambda_0)\leq O_p(N^{-1}h^{-1/2})$, and assumption B.1, we conclude that (28) $=o_p(1)$. Claim (c) will be shown later. This shows that all the second order terms are negligible in the asymptotic distribution.

Step 2: First Order Terms

We first consider (24). By triangle inequality, we have

$$\begin{aligned}
&|E_{n,k}[\partial_\lambda\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)]-E[\partial_\lambda\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)]|\\
\leq&\underbrace{|E_{n,k}[\partial_\lambda\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)]-E_{n,k}[\partial_\lambda\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)]|}_{J_{5k}}
\end{aligned}$$

$$+\underbrace{|E_{n,k}[\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)] - E[\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)]|}_{J_{6k}}.$$

To bound J_{6k}, note that since λ_0 is bounded away from zero and the score ψ is bounded by M_J,

$$\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0) \lesssim M_J$$

which implies that

$$E[J_{6k}^2] \leq \frac{1}{N} E[(\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0))^2] \lesssim M_J^2/N$$

and by Markov's inequality, we have $J_{6k} \leq O_p(M_J/\sqrt{N})$. With the assumption that $M_J/\sqrt{N} = o(1)$, we have $J_{6k} = o_p(1)$.

On the other hand, for J_{5k}, note that

$$\begin{aligned}
E[J_{5k}^2|I_k^c] &= E[|E_{n,k}[\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)] - E_{n,k}[\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)]|^2|I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta) - \partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)|^2|I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta) - \partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)|^2] \\
&\lesssim M_J^2 \varepsilon_N^2 \qquad \text{(d)}
\end{aligned}$$

where the first equation holds by definition, the second line holds by Cauchy-Schwarz and the third line holds by the construction that all the parameters are estimated using auxiliary sample I_k^c. Then we conclude with conditional Markov's inequality that $J_{5k} = o_p(1)$. Therefore,

$$E_{n,k}[\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)] \to^p E[\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)] \equiv S_\lambda^0$$

Note that $(\hat{\lambda}_k - \lambda_0) = O_p(N^{-1/2})$, we can rewrite (24) as

$$\begin{aligned}
(24) &= \sqrt{N}\frac{1}{K}\sum_{k=1}^K E_{n,k}[\partial_\lambda \psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)](\hat{\lambda}_k - \lambda_0) \\
&= \sqrt{N}\frac{1}{K}\sum_{k=1}^K S_\lambda^0(\hat{\lambda}_k - \lambda_0) + o_p(1) \\
&= \sqrt{N}\frac{1}{N}\sum_{i=1}^N S_\lambda^0(T_i - \lambda_0) + o_p(1)
\end{aligned}$$

where the last equality holds by the definition that $\hat{\lambda}_k - \lambda_0 = (N-n)^{-1}\sum_{i \in I_k^c} T_i - \lambda_0$ and

the fact that $K^{-1}\sum_{k=1}^{K}(\hat{\lambda}_k - \lambda_0) = \frac{1}{N}\sum_{i=1}^{N}(T_i - \lambda_0)$. We remark that, since $S_\lambda^0 = E[\partial_\lambda \psi_J^0]$ is bounded by a constant and $\hat{\lambda}$ converges at parametric rate, (24) vanishes when scaled by the square-root of the asymptotic variance.

Term (25) will be bounded using the same argument as in the repeated outcomes setting. First, by triangle inequality

$$\begin{aligned}
&|E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)] - E[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]| \\
\leq &\underbrace{|E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)] - E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|}_{J_{7k}} \\
+ &\underbrace{|E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)] - E[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|}_{J_{8k}}.
\end{aligned}$$

We first bound J_{8k}. Note that since f_d^0 is bounded away from zero and the score ψ is bounded by M_J, we have

$$\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0) = -\frac{1}{f_d^0}(\psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0) + \theta_{0J}) \lesssim M_J$$

which implies that

$$E[J_{8k}^2] \leq \frac{1}{N} E[(\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0))^2] \lesssim M_J^2/N$$

and by Markov's inequality, we have $J_{8k} \leq O_p(M_J/\sqrt{N})$. With the assumption that $M_J/\sqrt{N} = o(1)$, we have $J_{8k} = o_p(1)$.

Second, to bound J_{7k}, note that

$$\begin{aligned}
E[J_{7k}^2 | I_k^c] &= E[|E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)] - E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|^2 | I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta) - \partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2 | I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta) - \partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim M_J^2 \varepsilon_N^2 \qquad \text{(b)}
\end{aligned}$$

where the first equation holds by definition, the second line holds by Cauchy-Schwarz and the third line holds by the construction that all the parameters are estimated using auxiliary sample I_k^c. Then we conclude with the conditional Markov's inequality that $J_{7k} = o_p(1)$. Therefore,

$$E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)] \to^p E[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)] \equiv S_f^0$$

Note that under the assumption, $(\hat{f}_{d,k} - f_d^0) = O_p((Nh)^{-1/2})$, we can rewrite (25) as

$$\begin{aligned}
(25) &= \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)](\hat{f}_{d,k} - f_d^0) \\
&= \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} S_f^0(\hat{f}_{d,k} - f_d^0) + o_p(h^{-1/2}) \\
&= \sqrt{N}\frac{1}{N}\sum_{i=1}^{N} S_f^0(K_h(D_i - d) - E[K_h(D - d)]) + o_p(h^{-1/2})
\end{aligned}$$

where the last equality holds by the definition that $\hat{f}_{d,k} - f_d^0 = (N-n)^{-1}\sum_{i \in I_k^c} K_h(D_i - d) - E[K_h(D-d)] + O(h^2)$, the undersmoothing assumption that $\sqrt{N}h^2 \leq O(1)$, and the fact that $K^{-1}\sum_{k=1}^{K}(\hat{f}_{d,k} - E[K_h(D-d)]) = \frac{1}{N}\sum_{i=1}^{N}(K_h(D_i - d) - E[K_h(D-d)])$. This term will contribute to the asymptotic variance.

■

Acknowledgements

The author would like to express his gratitude to Andres Santos, Rosa Matzkin, and Denis Chetverikov for their generous time and extremely helpful discussions that have led to substantial improvements to this project. The author would also like to thank Oscar H. Madrid-Padilla, Mingli Chen, Rodrigo Pinto, and participants at UCLA econometrics proseminars for their helpful suggestions.

References

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies* **72(1)**, 1–19.

Altonji, J. G. and Matzkin, R. L. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* **73(4)**, 1053–1102.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys* **4**, 40–79.

Ashraf, N., Bau, N., Nunn, N., and Voena, A. (2020). Bride price and female education. *Journal of Political Economy* **128(2)**, 591–641.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica* **74(2)**, 431-497.

Athey, S. and Imbens, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics* **226(1)**, 62–79.

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

Blundell, R. W., Kristensen, D., and Matzkin, R. L. (2020). Individual counterfactuals with multidimensional unobserved heterogeneity. Working Paper.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics* **225(2)**, 200–230.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. (2021). Difference-in-differences with a continuous treatment. *arXiv preprint arXiv:2107.02637*.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *The American Economic Review* **84(4)**, 772.

Cattaneo, M. D. and Jansson, M. (2021). Average density estimators: Efficiency and bootstrap consistency. *Econometric Theory*, 1–35.

Colangelo, K. and Lee, Y. Y. (2022). Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*.

Chang, N. C. (2020). Double/debiased machine learning for difference-in-differences models. *Econometrics Journal* **23(2)**, 177–191.

Chen, M., Jiang, H., Liao, W., and Zhao, T. (2019). Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *arXiv preprint arXiv:1908.01842v5*.

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

De Chaisemartin, C. and D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review* **110(9)**, 2964–96.

de Chaisemartin, C., D'Haultfoeuille, X., Pasquier, F., and Vazquez-Bare, G. (2022). Difference-in-Differences Estimators for Treatments Continuously Distributed at Every Period. *arXiv preprint arXiv:2201.06898*.

D'Haultfoeuille, X., Hoderlein, S., and Sasaki, Y. (2021). Nonparametric difference-in-differences in repeated cross-sections with continuous treatments. *arXiv preprint arXiv:2104.14458*.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica* **64(5)**, 1001–1044.

Duflo, E. (2001). Schooling and labor market consequences of school construction in Indonesia: evidence from an unusual policy experiment. *American Economic Review* **91(4)**, 795–813.

Efromovich, S. (2010). Dimension reduction and adaptation in conditional density estimation. *Journal of the American Statistical Association* **105(490)**, 761–774.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83(1)**, 189–206.

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika* **91(4)**, 819-834.

Fan, Q., Hsu, Y. C., Lieli, R. P., and Zhang, Y. (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics* **40(1)**, 313–327.

Fortin, N., Lemieux, T., and Firpo, S. (2011). Decomposition methods in economics. *Handbook of Labor Economics* **Vol.4**, 1–102. Elsevier.

Guerre, E., Perrigne, I., and Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica* **68**, 525—574.

Gajek, L. (1986). On improving density estimators which are not bona fide functions. *Annals of Statistics* **14**, 1612–1618.

Haile, P., Hong, H., and Shum, M. (2006). Nonparametric tests for common value in first-price auctions. Working Paper, Yale University, New Haven, CT.

Hall, P., Wolff, R. C., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* **94(445)**, 154–163.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* **99(468)**, 1015–1026.

Hayakawa, S. and Suzuki, T. (2020). On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks* **123**, 343-361.

Heckman, J. (1990). Varieties of selection bias. *The American Economic Review* **80(2)**, 313–318.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 226164, 73–84.

Izbicki, R. and Lee, A. B. (2016). Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics* **25(4)**, 1297–1316.

Izbicki, R. and Lee, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics* **11(2)**, 2800–2831.

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

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

Kronmal, R. and Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *Journal of the American Statistical Association* **63(323)**, 925–952.

Lecué, G. and Mitchell, C. (2012). Oracle inequalities for cross-validation type procedures. *Electronic Journal of Statistics* **6**, 1803–1837.

Lorentz, G. G. (1966). Metric entropy and approximation. *Bulletin of the American Mathematical Society* **72**, 903–937.

Matzkin, R. L. (2007). Nonparametric identification. *Handbook of Econometrics* **6**, 5307–5368.

Matzkin, R. L. (2013). Nonparametric identification in structural economic models. *Annual Review of Economics* **5(1)**, 457-486.

Matzkin, R. L. (2015). Estimation of nonparametric models with simultaneity. *Econometrica* **83(1)**, 1-66.

Perrigne, I. and Vuong, Q. (2019). Econometrics of auctions and nonlinear pricing. *Annual Review of Economics* **11**, 27–54.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. *Multivariate Analysis II* **25**, 31.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66(5)**, 688–701.

Sant'Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics* **219(1)**, 101–122.

Semenova, V. and Chernozhukov, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* **24(2)**, 264–289.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 1040–1053.

Su, L., Ura, T., and Zhang, Y. (2019). Non-separable models with high-dimensional data. *Journal of Econometrics* **212(2)**, 646–677.

Suzuki, T. (2018, September). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. *International Conference on Learning Representations*.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Walter, G. and Blum, J. (1979). Probability density estimation using delta sequences. *Annals of Statistics* 328–340.

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* **27**, 1564–1599.