High-Dimensional Conditional Density Estimation and Continuous Difference-in-Differences

Lucas Z. Zhang
Department of Economics
University of California, Los Angeles
lucaszz@g.ucla.edu
[Click Here for the Latest Version]

September 26, 2023

Abstract

Conditional density enjoys a series representation, with each term being a known function multiplied by its conditional expectation. This structure is especially beneficial in high-dimensional settings, where these conditional expectations can be flexibly estimated using various machine learning methods. However, choosing the right series terms is challenging. We introduce a data-driven estimator using a cross-validation procedure and demonstrate its optimality through an oracle inequality that bounds the estimation error. Beyond our theory-backed estimation strategy, we underscore the extensive role of conditional density in economics, especially as the generalized propensity score in causal inference with continuous treatment. Furthering this discourse, we extend the widely-used difference-in-differences models to accommodate continuous treatment. Specifically, we establish identification, estimation, and inference results for the causal parameter of interest under the double/debiased machine learning framework. To illustrate the practicality of our methods, we revisit two notable empirical studies: Acemoglu and Finkelstein (2008) on technology adoption in U.S. healthcare industries, and Duflo (2001) on the impact of a large-scale development policy in Indonesia.

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

JEL Classification Codes: C13; C14; C31; C33

1 Introduction

Researchers are often interested in how the distribution of an outcome Y depends on covariates X. The conditional density of Y given X, denoted as $f_{Y|X}$, is a fundamental statistical object that summarizes such a relationship. Its role in economics is especially pronounced with a wide range of applications. For instance, when studying the identification of structural economic models, conditional densities are used to establish the connection between what can be observed from the data and the structural parameters (e.g. Matzkin (2007, 2013)). Take the first-price auction as an example: the conditional density of the bids can be used to recover the private values of the bidders (e.g. Guerre et al. (2000); Perrigne and Vuong (2019)). Other notable examples[1] 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)). Given the crucial role of conditional density in economics, researchers might be inclined to avoid making potentially restrictive parametric assumptions and, instead, prefer its nonparametric estimation. This can be especially challenging in the high-dimensional setting where the number of covariates X is large.

The literature on nonparametric conditional density estimation is vast. The most well-known nonparametric method is perhaps the kernel method proposed in Rosenblatt (1969) and a subsequent literature 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 the local polynomial regression studied in Fan et al. (1996) and Fan and Yim (2004), and more recently the methods 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 estimators have many attractive theoretical properties, they converges slowly as the dimension of the conditioning variable becomes large.[2] 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.[3] Moreover, even the data-driven estimators

[1]We will examine the role of the conditional density in these examples in detail in Section 2.

[2]See also Ma and Zhu (2013) for a review of various dimension reduction techniques, which often require very strong assumptions.

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

from Hall et al. (2004), Fan and Yim (2004) and Efromovich (2010) have drawbacks: Hall et al. (2004) require cross-validation searching over each covariate, which becomes computationally intractable as number of covariates 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 not yet been studied.[4]

To improve upon previous literature, we propose a data-driven nonparametric conditional density estimator that is feasible in the high-dimensional setting. First, for a suitable sequence of known functions $\{\phi_j\}_{j=1}^{\infty}$ of Y, we 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. That is, the conditional density can be expressed as an infinite sum of known functions multiplied by their conditional expectations. In particular, for a high-dimensional conditioning variable X, instead of estimating the conditional density directly, this representation allows researchers to estimate the conditional expectation $E[\phi_j(Y)|X]$ in each series term using any state-of-the-art machine learners, such as deep neural networks. This motivates an estimator of the form

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

and notably Izbicki and Lee (2017) have studied the properties of such an estimator for a *deterministic* series cutoff J. Nevertheless, choosing the optimal series cutoff deterministically requires researchers to make potentially unrealistic assumptions that are difficult to verify in practice. Therefore, it is preferred to choose J in a data-driven way with theoretical guarantees. To this end, we resort to a cross-validation procedure, in which the series cutoff $\hat{J}$ is chosen by minimizing an empirical risk. Our final estimator takes the form of an average of sub-sample estimators with this cutoff using every training sample. Following the general strategy proposed by Lecué and Mitchell (2012), we establish an oracle inequality that shows our estimator is asymptotically optimal. To the best of our knowledge, this is the first such result of a nonparametric conditional density estimator that is both data-driven and feasible in the high-dimensional setting. We recognize that there is an extensive literature on cross-validation, and due to space limitations, we refer the readers to Arlot

[4]There is also a large literature on parametric or semiparametric density/conditional density estimation. For example, Rothfuss et al. (2019) use neural networks to estimate conditional densities with flexible parametric mixture models (see also the references therein for a review on the related literature).

and Celisse (2010) for a comprehensive survey.

To add to the growing literature in economics where the conditional density plays a key role, we study in detail a new 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 recent theoretical advancements aim to accommodate the complexities in empirical research. Notably, these developments include extensions to semiparametric settings (Abadie (2005)), nonlinear settings (Athey and Imbens (2006)), and multiple periods and staggered treatment timing settings (e.g., Callaway and Sant'Anna (2021); de Chaisemartin and D'Haultfoeuille (2020); Athey and Imbens (2022)). While most of the existing literature has primarily focused on DiD models with binary or discrete treatments, recent studies (e.g., Callaway et al. (2021) and D'Haultfoeuille et al. (2021)) have explored DiD settings with continuous treatment, opening up new avenues for investigation.

In this study, we expand upon this new line of research by considering a 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 intensity under the *conditional parallel trends* assumption. In this setting, the conditional density of the continuous treatment plays the role of a *generalized propensity score*. The fully nonparametric estimator of the ATT in this setting involves averaging over estimated infinite-dimensional nuisance parameters, which can result in substantial biases. To address this issue, in context of DiD settings with binary or discrete treatments, Sant'Anna and Zhao (2020) propose doubly robust estimators, while Chang (2020) studies doubly/debiased machine learning estimators that allow for high-dimensional controls. Building on the insights from Chang (2020), we adopt and extend the double/debiased machine learning (DML) framework from CCDDHNR (2018) to accommodate continuous treatments.

To illustrate the usefulness of our methods, we revisit two influential empirical studies. First, we apply our method to Acemoglu and Finkelstein (2008), which examines the impact of the 1983 Medicare Prospective Payment System (PPS) reform on the adoption of new technologies in the heavily regulated healthcare industry. The PPS reform notably altered the reimbursement structure for Medicare inpatient expenses, thus affecting hospitals with varying proportions of Medicare patients differently. Given this context, the share of Medicare patients can be interpreted as a continuous treatment variable, making Acemoglu and Finkelstein (2008) an examplary case for applying our methodology. We thus nonparametrically estimate the ATT of the PPS reform in a continuous DiD context, providing a more detailed understanding of the effects of this policy reform.

The second empirical study we consider is Duflo (2001), which studies the effect of a large-scale policy intervention in Indonesia (INPRES) on educational outcomes. One analysis in Duflo (2001) relies on a binary DiD comparing high vs. low treatment intensity regions. In contrast, we model this policy intervention as a continuous treatment, thus allowing the treatment group to have varied treatment intensities. While Duflo (2001) finds a small positive effect of this policy, our more granular analysis finds important and statistically significant heterogeneity in the ATTs across different treatment intensities.

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 involving conditional densities. In section 3, we first show the validity of the series representations of the conditional densities, then discuss the construction of our cross-validated estimator in detail, and finally, establish the theoretical properties of our estimator. In section 4, we formally set up the DiD models with continuous treatment, show identification, estimation, and inference under the DML framework, and illustrate the usefulness of our results with empirical applications. Finally, we conclude in Section 5. All the proofs will be given in the appendix.

2 Applications

In this section, we discuss several empirical examples in which the estimation of a conditional density plays a crucial role.

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 $V_i \in [v_L, v_H] \subset \mathbf{R}$. In an auction with characteristics X, each bidder bids $B_i = s(V_i, X)$ that maximizes the expected utility. If the equilibrium bid function s is monotonic in V, then using the first-order condition, the unobserved private value V_i can be written as

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

where $G(\cdot|I,X)$ and $g(\cdot|I,X)$ denote the observed equilibrium bid distribution and density conditional on the number of bidders I and the covariates X. This is the main identification equation that enables the researcher to recover the model primitives $(V, f_{V|X,I})$. Using this identification result, Guerre et al. (2000) study the nonparametric estimation of these primitives using kernel methods. For potentially high-dimensional covariates X, Haile et al. (2006) and Perrigne and Vuong (2019) propose single index restrictions on the relationship between the private value V and covariates X to reduce the dimension. While the estimators based on such single index restrictions are easy to implement, they can suffer from significant misspecification errors if the single index assumptions do not hold. In contrast, our method will allow researchers to nonparametrically estimate the conditional bid distribution $f_{V|I,X}$

for high-dimensional X using machine learning methods in a data-driven way without having to rely on such single index restrictions. ■

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 a 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 $X \perp (\epsilon_1, \cdots, \epsilon_K)|Z$.[5] Specifically, the authors consider the identification of 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 $\beta(x)$ is identified as

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

A nonparametric estimator can be constructed based on this expression, which requires the estimation of the conditional density $f_{Z|X}$. For another example, in nonparametric nonseparable simultaneous equation models, Matzkin (2015) shows that the structural derivatives can be constructively identified as the functionals of conditional densities of observed variables. As before, the nonparametric estimation based on such identification results rely on the nonparametric estimation of the conditional densities. The literature typically employs kernel estimators due to their well-established theoretical properties; however, such estimators typically require the researchers to specify the kernel bandwidth, and even with covariates of moderate dimensions, the rate of convergence of such estimators can be slow. Therefore, our data-driven estimator can be used as an alternative that potentially achieves a faster rate of convergence even with high-dimensional covariates. ■

Example 2.3 (**Continuous Treatment**). Hirano and Imbens (2004) introduce 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 $Y_i := Y_i(T_i)$ 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) refer to the conditional density $f_{T|X}$ as the generalized propensity score. Under the weak unconfoundedness assumption

[5] 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.

that $Y(t) \perp T|X$ for all $t \in [t_0, t_1]$, the average potential outcome at $T = t$ is identified as

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

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, which can fail to capture the complexities of the true conditional densities.

In a related study, Kennedy et al. (2017) propose an alternative identification result of $E[Y(t)]$ using a doubly robust signal $Y(\eta)$, where $\eta = (E[Y|T, X], f_{T|X})$ denotes the infinite-dimensional nuisance parameters, such that

$$E[Y(t)] = E[Y(\eta)|T = t]. \tag{2}$$

To estimate $E[Y(t)]$ using this expression, researchers first need to estimate the conditional density $f_{T|X}$.[6] Kennedy et al. (2017) estimate such conditional density by first assuming a model $T = \mu(X) + \sigma(X)\epsilon$, then using a suite of ML methods to estimate $\mu(X) = E[T|X]$ and $\sigma(X) = Var(T|X)$, and in the final step, estimating $f_{T|X}$, now effectively a univariate density estimation problem, using the standard kernel method. One concern is that this approach only captures the relationship between treatment T and covariates X up to a second moment. In contrast to Hirano and Imbens (2004) and Kennedy et al. (2017), our nonparametric conditional density estimator does not require additional modeling assumptions while still being computationally tractable. ■

Example 2.4 (Conditional Average Partial Derivative). Let $T \in \mathbf{R}$ be a continuous treatment variable, $Y = Y(T)$ the observed potential outcome, Z a vector of controls, and X be a subvector of Z. Semenova and Chernozhukov (2021) define the conditional average partial derivative $\partial_t E[Y(t)|X = x]$ as the parameter of interest. Under the conditional independence assumption $\{Y(t), t \in \mathbf{R}\} \perp T|Z$, Semenova and Chernozhukov (2021) show that $\partial_t E[Y(t)|X = x]$ is identified as

$$\partial_t E[Y(t)|X = x] = E[Y(\eta)|X = x] \tag{3}$$

where $Y(\eta)$ is a signal that depends on the nuisance parameter $\eta := (E[Y|T, Z], f_{T|Z})$. The estimation of $\partial_t E[Y(t)|X = x]$ based on (3) requires first estimating the nuisance parameters $\hat{\eta}$, particularly the conditional density $f_{T|Z}$. Semenova and Chernozhukov (2021) first assume a model $T = \mu(Z) + \epsilon$ with $\epsilon \perp Z$, then estimate $\mu(Z)$ using LASSO, and finally,

[6] In recent works, Kallus and Zhou (2018), Su et al. (2019), and Colangelo and Lee (2022) also consider the estimation (and inference in the latter two studies) of $E[Y(t)]$ using an alternative score. Nevertheless, the conditional densities still have to be estimated as a first step.

estimate the conditional density as a univariate density. Nevertheless, the independence assumption $\epsilon \perp Z$ can be difficulty to verify in practice, and the conditional density estimator based on such a model can only capture the relationship between T and Z up to the first moment. In contrast, our nonparametric estimator can be employed here without the additional modeling assumption that $\epsilon \perp X$, and can capture the rich complexity in $f_{T|X}$ beyond the first moment. ■

Example 2.5 (**Counterfactual Distributions**). Counterfactual distributions have been employed extensively in the studies of wage inequality. 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). Under an assumption of invariance of counterfactual distributions (see Fortin el al. (2011)), the counterfactual density of group A can be identified as

$$f^c_{Y_A}(y) = \int f_{Y_A|X_A}(y|x) \frac{dF_{X_B}(x)}{dF_{X_A}(x)} dF_{X_A}(x) \tag{4}$$

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 (4), which requires estimation of the conditional density $f_{Y_A|X_A}$, and our estimator can be employed directly here. Alternatively, an orthogonal score for (4) can be constructed for high-dimensional covariates,[7] and our data-driven conditional density estimator that utilizes machine learning methods can be particularly useful in this setting. ■

3 Conditional Density Estimation

In this section, we show that conditional density can be represented as a series, discuss the construction of our cross-validated estimator based on such representation in detail, and establish theoretical results on estimation error for such an estimator.

[7] Currently we are studying this as a work in progress in a separate project.

3.1 Series Representation

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

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

Assumption 3.1 is a set of mild regularity conditions generally satisfied in most cases in economics. For example, economic variables Y and X typically take values in well-behaved subsets $\mathbf{Y} \times \mathbf{X} \subseteq \mathbf{R} \times \mathbf{R}^d$, which, together with assumption 3.1 (iii), ensure that[8] $L^2(\nu_Y)$ is separable and countable orthonormal bases exist. Such orthonormal bases will provide the functions used in the series representation of the conditional densities. Moreover, assumption 3.1 (iii) does impose restrictions on the support of Y and X and rules out random variables with degenerate distributions; nevertheless, both continuous and discrete X's are allowed. Under this assumption, the Radon-Nikodym derivative of P w.r.t. ν exists, i.e., there is a density $f_{Y,X}$ s.t.

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

The conditional density can then be defined as:

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

Note that since $f_X(x) = 0$ implies $f_{Y,X}(\cdot, x) = 0$ ν_Y-a.e., defining $f_{Y|X}(y|x) := 0$ for $f_X(x) = 0$ has little impact in a measure-theoretic sense. However, such a definition ensures $f_{Y|X}(y|x) f_X(x) = f_{Y,X}(y,x)$ for all $(y,x) \in \mathbf{Y} \times \mathbf{X}$, which will help us simplify the formal arguments when showing the series representation is valid. Finally, let P_X be the projection of P onto $\mathbf{X}$, that is, for any $B \in \mathcal{B}_X$, $P_X(B) = P(\mathbf{Y} \times B)$. 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;*

[8] $L^2(\nu_Y)$ is defined as the set of square-integrable functions of Y w.r.t. the measure ν_Y.

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

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

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

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

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 $f_{Y|X}$ is square integrable w.r.t the product measure $\nu_Y \otimes P_X$, the series expansion holds P_X-a.e. (in the sense that for a.e. x, the series converges in $L^2(\nu_Y)$) as well as in $L^2(\nu_Y \otimes P_X)$. From now on, we will use the following representation whenever the convergence holds:

$$f_{Y|X}(y|x) = \sum_{j=1}^{\infty} E[\phi_j(Y)|X = x]\phi_j(y). \tag{5}$$

In particular, $L^2(\nu_Y)$ being separable guarantees the existence of a countable orthonormal basis (due to Zorn's lemma and Gram-Schmidt process). Since ν_Y is known, in practice, there are many well-known orthonormal bases for the researchers to choose from. Therefore, each term in the series expansion (5) is the multiplication of a known function and its conditional expectation, which motivates a series estimator for the conditional density. In the next section, we will discuss the construction of our estimator based on such series expansions in detail.

3.2 Cross-Validated Estimator

Suppose we have an i.i.d. random sample $\{(Y_i, X_i)\}_{i=1}^{n} \sim (Y, X)$ that satisfies assumption 3.1 and an orthonormal basis $\{\phi_j\}_{j=1}^{\infty}$ on $L^2(\nu_Y)$. Building on the series expansion established in the previous section, an estimator can be constructed by first picking a cutoff J and estimating the conditional expectations $h_j(X) := E[\phi_j(Y)|X]$ for $j = 1, \cdots, J$, then forming

$$\hat{f}_J(y|x) = \sum_{j=1}^{J} \hat{h}_j(x)\phi_j(y). \tag{6}$$

For potentially high-dimensional covariates X, researchers can estimate the conditional expectations $\{h_j\}_{j=1}^J$ using any of their preferred machine learning methods.

In order to assess the quality of such an estimator, we need a metric to quantify how "close" this estimator is from the true conditional density $f_{Y|X}$. Since the series expansion holds for $f_{Y|X} \in L^2(\nu_Y \otimes P_X)$, it is natural to consider the L^2 norm w.r.t. the product measure $\nu_Y \otimes P_X$. For notational simplicity, for any function g of (y,x) in $L^2(\nu_Y \otimes P_X)$, we denote this norm as

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

where the second equality holds by definition since P_X is the probability measure.

Suppose we want to find an esitmator $\hat{f}$ that minimizes the L^2 norm:

$$\|\hat{f} - f_{Y|X}\|_H^2 = \int \left(\hat{f}(y|x) - f_{Y|X}(y|x)\right)^2 d\nu_Y(y) dP_X(x) \tag{7}$$

which is the same as minimizing the following [9]:

$$\|\hat{f} - f_{Y|X}\|_H^2 - \|f_{Y|X}\|_H^2 = \int \hat{f}^2(y|x) - 2\hat{f}(y|x) f_{Y|X}(y|x) d\nu_Y(y) dP_X(x). \tag{8}$$

This expression is impractical to work with since it requires knowledge of the true conditional density $f_{Y|X}$. However, the following lemma shows that this objective is equivalent to a *risk* function that can be estimated from data.

Lemma 3.1. *Define a loss function*

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

and the associated risk of an estimator $\hat{f}$ *as*

$$R(\hat{f}) := E[Q((Y,X),\hat{f})] = E[\int \hat{f}^2(y|X) d\nu_Y(y) - 2\hat{f}(Y|X)]. \tag{10}$$

Then risk $R(\hat{f})$ *satisfies*

$$R(\hat{f}) = \|\hat{f} - f_{Y|X}\|_H^2 - \|f_{Y|X}\|_H^2. \tag{11}$$

This lemma can be shown using the Law of Iterated Expectations and the fact that $f_{Y|X}$ is the conditional density of Y given X. The proof is given in the appendix. The

[9] This holds because $\|f_{Y|X}\|_H^2 = E_X[\int f_{Y|X}^2(y,X) d\nu_Y(y)]$ is a constant.

lemma suggests that our problem is essentially a risk minimization problem and the risk is minimized at the true conditional density $f_{Y|X}$. In particular, given data $\{(Y_i, X_i)\}_{i=1}^n$ and $\hat{f}$, we can define the *empirical risk* of $\hat{f}$ as

$$R_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} Q((Y_i, X_i), \hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\int \hat{f}^2(y|X_i)d\nu_Y(y) - 2\hat{f}(Y_i|X_i). \tag{12}$$

We now have all the neceassary ingredients to describe our cross-validation procedure, adapting the general framework laid out in Lecué and Mitchell (2012) to our setting. This cross-validation procedure is formally summarized in Algorithm 1.

Algorithm 1 Average Cross-Validated Conditional Density Estimator

Input: Data $D^{(n)} = \{(Y_i, X_i)\}_{i=1}^n$, orthonormal basis $\{\phi_j\}_{j=1}^\infty$ of Y, a maximum cutoff p, a method for estimating conditional expectations, and an integer $K \geq 2$.
Output: Estimator $\bar{f}^{(n)}(y|x)$.

1: Split $D^{(n)}$ into K disjoint subsets $D_1^{(n_V)}, \ldots, D_K^{(n_V)}$ as validation sets and their complements $\{D_k^{(n_T)} = D^{(n)} \setminus D_k^{(n_V)}\}_{k=1}^K$ as training sets.
2: **for all** $1 \leq k \leq K$ **do**
3: **for all** $1 \leq j \leq p$ **do**
4: Estimate $h_l = E[\phi_l(Y)|X]$ for $l = 1, \ldots, j$ using training set $D_k^{(n_T)}$.
5: Construct $\hat{f}_j^{(n_T)}(D_k^{(n_T)})(y|x) = \sum_{l=1}^j \hat{h}_l(x)\phi_l(y)$.
6: **end for**
7: **end for**
8: **for all** $1 \leq j \leq p$ **do**
9: Calculate K-fold empirical risk $R_{n,K}$ according to (13) using $\{\hat{f}_j^{(n_T)}\}_{k=1}^K$.
10: **end for**
11: Solve $\hat{j}^* = \arg\min_{1\leq j\leq p} R_{n,K}$.
12: **return** $\bar{f}^{(n)}(y|x) = \sum_{l=1}^{\hat{j}^*} \tilde{h}_l(x)\phi_l(y)$, where $\tilde{h}_l(x) = K^{-1}\sum_{k=1}^K \hat{h}_l(D_k^{(n_T)})$.

The first step is to split a sample into training and validating subsamples. Formally, let n denote the sample size and without loss of generality suppose n is divisible by some fixed integer K. Then we split the sample[10] $D^{(n)} := \{(Y_i, X_i\}_{i=1}^n$ into K disjoint validating sets $D^{(n_V)}$ of equal size $n_V := n/K$. These validating sets will be used to compute the empirical risks of candidate estimators. In addition, 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.

In the second step, we use the training sets to train a large dictionary of candidate

[10] Although we assume an i.i.d. random sample, in practice, the data researchers received might have been sorted by certain criteria independent of the data-generating process beforehand. In this case, the researchers can use an external randomization device independent of the data-generating process to reshuffle the data before the sample splitting.

estimators. To be more precise: first, we pick a large p, which denotes the cardinality of the dictionary, and consider a set of statistics[11] $\{\hat{f}_1, \cdots, \hat{f}_p\}$ such that its j-th element is $\hat{f}_j(y|x) = \sum_{l=1}^{j} \hat{h}_l(y)\phi_l(x)$ (recall $\hat{h}_l$'s are the preferred machine learners of $h_l = E[\phi_l(Y)|X]$'s); second, on each of the $k = 1, \cdots, K$ training sets $D_k^{(n_T)}$ of size n_T, we train the machine learners of conditional expectations $\{h_l\}_{j=1}^{p}$ and then construct $\hat{f}_j^{(n_T)}(D_k^{(n_T)})$ for $j = 1, \cdots, p$ using the trained $\{\hat{h}_l(D_k^{(n_T)})\}_{l=1}^{p}$.

In the third step, we use these trained estimators to evaluate a empirical version of the risk on the validating sets. Specifically, we define the *K-fold empirical risk* of $\hat{f} \in \{\hat{f}_j\}_{j=1}^{p}$ as

$$R_{n,K}(\hat{f}) := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_V} \sum_{i \in D_k^{(n_V)}} Q((Y_i, X_i), \hat{f}^{(n_T)}(D_k^{(n_T)})) \tag{13}$$

where recall $D_k^{(n_T)}$ is the k-th training set and $D_k^{(n_V)} = D^{(n)} \backslash D_k^{(n_T)}$ is the k-th validating set. That is, for each $\hat{f}^{(n_T)}(D_k^{(n_T)})$ trained using $D_k^{(n_T)}$, we evaluate its empirical risk on the validating set $D_k^{(n_V)}$. Then we average over the K validating sets to obtain the K-fold empirical risk. One potential concern is that the empirical risk $R_{n,K}$ takes the form of an empirical average of loss Q, which involves integral calculations. However, note that the estimators we consider take the form $\hat{f}_j = \sum_{l=1}^{j} \hat{h}_l \phi_l$ with ϕ_l's being elements in an orthonormal basis. Then by orthonormality, the loss can be rewritten as $Q((y, x), \hat{f}_j) = \sum_{l=1}^{j} \hat{h}_l^2(x) - 2\hat{f}_j(y, x)$, which only requires simple summations when computing the empirical risk.

In the final step, we construct our estimator by first finding the index $\hat{j}^*$ that corresponds to the smallest K-fold empirical risk, and then average over $\hat{f}_{\hat{j}^*}$ trained on each training sets. Formally, we define our estimator as

$$\bar{f}^{(n)} := \frac{1}{K} \sum_{k=1}^{K} \hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)}) \quad \text{with} \quad \hat{j}^* = \arg \min_{1 \le j \le p} R_{n,K}(\hat{f}_j). \tag{14}$$

Although $\bar{f}$ aggregates sub-sample estimators, it can still be expressed 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. We note that this estimator differs from the typical K-fold CV estimator $\hat{f}_{CV} := \hat{f}_{\hat{j}^*}^{(n)}$ that is trained by using the full sample $D^{(n)}$ after finding the $\hat{j}^*$ above. While we do not compare[12] the quality of $\bar{f}^{(n)}$ to $\hat{f}_{CV}$, we emphasize that $\bar{f}^{(n)}$ is also constructed using the full sample and does not require re-training after

[11] We follow Lecué and Mitchell (2012) and define a statistic $\hat{f} = (\hat{f}^{(m)})_{m \in \mathbf{N}}$ as a sequence such that each $\hat{f}^{(m)}$ is associated with $\hat{f}^{(m)}(D^{(m)})$ trained using data $D^{(m)}$.

[12] As commented in Lecué and Mitchell (2012), with additional regularity conditions, the estimation error of $\hat{f}_{CV}$ can be bounded by that of the sub-sample estimator.

selecting $\hat{j}^*$.

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 one 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 latter, we consider the following set

$$C := \{c \in \ell^2 : \sum_{j=2}^{\infty} c_j\phi_j(y) \geq -\phi_1\}. \tag{15}$$

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 each of the sub-sample estimator $\hat{f}_{\hat{j}^*}$. In particular, since for each x, $f_{Y|X}(\cdot|x)$ is a density in $L^2(\nu_Y)$, one can consider the orthogonal projection algorithms (e.g., the *p-algorithm* in Gajek (1986)), which can be shown to weakly reduce the estimation error (see Theorem 1 in Gajek (1986) for example). Therefore, our main results will be established for the pre-processed estimators, and in practice researchers can decide what post-processing methods to use if they find the estimator is negative.

3.3 Theoretical Results

We first establish an oracle inequality[13] for our estimator, that is, an inequality that relates our estimator to an "ideal" estimator that, in our case, minimizes the estimation error. 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. *Let $\{(Y_i, X_i)\}_{i=1}^n$ be an i.i.d random sample distributed according to (Y, X) such that assumption 3.1 is satisfied. Assume $f_{Y|X} \in L^2(\nu_Y \otimes P_X)$ and let $\{\phi_j\}_{j=1}^{\infty}$ be an orthonormal basis on $L^2(\nu_Y)$. Moreover, assume $f_{Y|X}$ and the statistics $\{\hat{f}_j\}_{j=1}^p$ defined as in (6) are bounded by some constant M. Let $\bar{f}$ be the estimator defined in (14). Then for any constant $a > 0$, there exists a constant C that only depends on a such that*

$$E[\|\bar{f}^{(n)} - f_{Y|X}\|_H^2] \leq (1+a) \min_{1\leq j\leq p} E[\|\hat{f}_j^{(n_T)} - f_{Y|X}\|_H^2] + C\frac{\log p}{n_V}. \tag{16}$$

[13] See, for example, section 4 in Candes (2006) for an introduction.

This oracle inequality essentially states that the estimation error[14] of our estimator $\bar{f}$ is bounded above (up to a constant) by the smallest achievable estimation error for a given dictionary of estimators $\{\hat{f}_j\}_{j=1}^p$. In particular, the theorem accommodates any machine learning estimators of the conditional expectations $\hat{h}_l$'s in each $\hat{f}_j(y|x) = \sum_{l=1}^j \hat{h}_l(x)\phi_l(y)$. Note that the oracle inequality (16) is established 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 estimators $\{\hat{f}_j\}_{j=1}^p$ are uniformly bounded above by some constant. We can even modify the theorem to allow for this bound to grow with p.[15] Moreover, the convexity of the loss Q and the associated risk R defined in section 3.2 plays a major role in the proof. Specifically, the convexity of the risk allows us to bound the expected difference of $R(\bar{f}^{(n)}) - R(f_{Y|X})$ by two terms, one 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 estimator, which gives rise to the $\log(p)/n_V$ term in (16).

On the other hand, to obtain a concrete estimation error that is more familiar to practitioners, additional assumptions on our estimator and on the true conditional density $f_{Y|X}$ are needed. Recall that the estimation error of $\hat{f}_J$ satisfies the bias-variance decomposition

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

which suggests that this estimation error should be minimized at some cutoff J under suitable regularity conditions. Moreover, as long as K (as in K-fold cross-validation) is fixed, the sample sizes of the training set (n_T) and validating set (n_V) are in the same order as the sample size n. Hence, for sufficiently large p, the minimum is achieved in the oracle in equation (16), which establishes an upper bound on the estimation error of our cross-validated estimator $\bar{f}$. In the next theorem, we show such a result 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 $0 < \delta \le 1$, $E[(\hat{h}_j(X) - h_j(X))^2] \asymp n^{-\delta}$ *for all* $j \ge 1$;

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

[14] The expectation is taken w.r.t. the estimator.

[15] 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.

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

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

Condition (i) in Theorem 3.2 makes an assumption on the quality of the conditional expectation estimators $\hat{h}_j(X) = \hat{E}[\phi_j(Y)|X]$. 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 investigating the estimation error of various state-of-the-art machine learning estimators. For example, Chen et al. (2022) establish the estimation error in the form of condition (i) (up to a log term) 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 section 4 in Izbicki and Lee (2017) for several other examples that satisfy (i). In particular, machine learning estimators such as deep neural networks are particularly useful in the setting with high-dimensional covariates X: such ML estimators can often adapt to the intrinsically low-dimensional structures typically exhibited in high-dimensional data, which translates to a much faster rate of convergence (see, e.g., Chen et al. (2022)).

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 appealing features of our results. First, our conditional density estimator accommodates any estimators for conditional expectations in the series. In particular, 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[16] for fast convergence of

[16] For example, such conditions include but are not limited to the sparsity or approximate sparsity as-

ML estimators $\hat{h}_j$ in the high-dimensional setting are satisfied, our estimator achieves a fast rate of convergence. Last but not least, our estimator is data-driven with theoretical guarantees. In particular, the optimal cutoff J is 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, researchers may be interested in the conditional density at a point, i.e., $f_{Y|X}(y|X)$ at a specific y. For example, such a result can be useful in our continuous difference-in-differences framework, which will be discussed in the next section. Therefore, we conclude this section with our next theorem that shows the rate in Theorem 3.2 can also be achieved in this point-wise case under the proposed conditions.

Theorem 3.3. *Suppose conditions in Theorem 3.1 and 3.2 are satisfied. Moreover, assume*

(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 $\Sigma_J := E[B_J(X)B_J(X)']$ with $B_J(X)$ being the column vector $B_J(X) := (h_j(X) - \hat{h}_j(X))_{j=1}^J$;

(iii) there exist a measurable function $c(\cdot)$ that satisfies $E[c^2(X)] < \infty$ and a constant $\gamma > 0$ such that for all $J \geq 0$, $|\sum_{j=J+1}^{\infty} h_j(x)\phi_j(y)| \lesssim c(x)J^{-\gamma/2}$.

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

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

In the theorem, condition (i) ensures that the magnitude of each basis term does not affect the bounds on variance and bias. Examples of bounded bases include trigonometric bases on intervals in $\mathbf{R}$ and Hermite basis on the whole $\mathbf{R}$. This condition can be relaxed to allow for unbounded bases, potentially at the cost of a slower rate of convergence. Moreover, condition (ii) is a high-level assumption, which is determined by the quality of the estimators $\hat{h}_j$'s. In particular, the diagonal entries of the matrix Σ_J measure the variances of each conditional mean estimator in the series, while the off-diagonal entries measure the cross-term correlations. In contrast, when establishing MISE in Theorem 3.2, there is no such correlation due to the orthonormality of ϕ_j's. Additionally, we assume (iii) to control the point-wise bias, which is motivated by the analogous conditions in the (unconditional) orthogonal series density estimations. For the unconditional case, such conditions can be

sumptions typically assumed in the literature.

satisfied under certain smoothness assumptions for specific orthonormal bases; see discussions in Wahba (1975) for the cosine basis and Liebscher (1990) for the Hermite basis. In our case, however, we require such conditions on the tail-sum to hold uniformly on the support of the conditioning variable X (up to a square-integrable function $c(\cdot)$).

Remark 3.1. *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 the decomposition:*

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

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 extend the difference-in-differences models to the case of continuous treatment. In this setting, the conditional density of the continuous treatment plays a crucial role in identifying the parameter of interest, and 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 work. While the more common DiD settings focus on binary or discrete multi-valued treatments, there has been an increasing amount of interest in DiD with continuous treatments. The main idea of continuous DiD is simple: the treatment group rarely receives the treatment at the same level, and the treatment effect can vary with the “dose/intensity” of the treatment. Therefore, instead of comparing the outcomes of the treated and the controls before and after the treatment at the *group* level, one can futher examine the treated group and compare the outcomes at *different treatment intensity*.

In fact, continuous treatment is prevalent in many empirical settings. For instance, each affected individual can have varied exposure to policy interventions, marketing campaigns, or environmental pollutants, all of which can be modeled as continuous treatments. In particular, several recent studies in various fields have employed DiD with continuous treatments. These include the study by Zeng et al. (2022) on the impact of online advertising sites shutdowns, Cook et al. (2023)'s work on racial discrimination in public accommodations, and Ananat et al. (2022)'s study on the effects of the expanded child tax credit.

Nevertheless, while continuous DiD finds its popularity among empirical studies, its theoretical foundation is still limited, and a few recent studies have just started to fill this gap, notably Callaway et al. (2021); D'Haultfoeuille et al. (2021); de Chaisemartin et al. (2022). For instance, Callaway et al. (2021) examine continuous DiD in the context of the commonly used two-way fixed effect (TWFE) regression setting. Concurrently, D'Haultfoeuille et al. (2021) generalize the change-in-changes model studied in Athey and Imbens (2006) to continuous treatment. In contrast to the aforementioned literature, our results build upon the semiparametric framework proposed in Abadie (2005), broadening its applicability to settings involving continuous treatments.

The main advantage of our approach is that it explicitly accounts for the presence of covariates and focuses directly on causal parameter: the average treatment effect on the treated (ATT) at any given treatment intensity. As noted in Abadie (2005), the (unconditional) parallel trends assumption[17] can be restrictive if there are covariates that affect outcome dynamics and their distributions differ between control and treatment groups. Therefore, we follow the same motivation and incorporate covariates into our identification and estimation strategy. However, one major difference sets our results apart from Abadie (2005) is the presence of the continuous treatment, particularly its conditional density, which is commonly referred to as the "generalized propensity score" (see Hirano and Imbens (2004)). In this context, the causal parameter of interest, the ATT, becomes a functional of the infinite-dimensional conditional density. This motivates us to consider the estimation and inference of the causal parameters under the double/debiased machine learning (DML) framework studied in CCDDHNR (2018).

In particular, the estimation of the causal parameter requires first estimating nuisance parameters, including the conditional density of the continuous treatment. For potentially high-dimensional controls, researchers have to resort to machine learning methods to estimate these nuisance parameters. However, the use of machine learning methods can often result in substantial bias in the estimation of the causal parameter, see CCDDHNR (2018) and the references therein for further examples. Moreover, if one estimates the nuisance parameters and the causal parameter using the same sample, another source of bias due

[17] That is, on average, in the absence treatment, the time trends in the outcomes between the controls and the treated are the same.

to overfitting can also arise. To address these concerns, DML employs both an orthogonalization procedure and a cross-fitting procedure to reduce the influence of the nuisance parameters.

Due to these attractive properties of DML, drawing parallels with Chang (2020)—which provides insights into the DiD with discrete treatments under the DML framework—we extend the DML to our continuous DiD setting. Specifically, we derive orthogonal scores in both repeated outcomes (panel data) and repeated cross-sections settings. Using these scores, we construct DML estimators of the ATTs and study their asymptotic properties. In particular, we show that the DML estimators are asymptotically normal and derive their asymptotic variance. In addition, in each of these settings, we provide a detailed empirical example to illustrate the usefulness of our method. Specifically, in the panel data setting, we revisit Acemoglu and Finkelstein (2008) which studies the impact of the 1983 Medicare payment system reform on the heavily regulated health care industry. Moreover, in the repeated cross-sections setting, we re-examine the effect of a large-scale policy intervention in Indonesia on education outcomes first studied by Duflo (2001).

4.1 Setup and Identification

In this section, we formally set up the difference-in-differences with continuous treatment following Abadie (2005). First, using the potential outcome notation (e.g. Rubin (1974)), 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.

The treatment variable D is modeled as a random variable with a mixture distribution[18]: a probability mass at 0 and a continuous distribution on an interval $[d_L, d_H]$ excluding 0. Specifically, the control group consists of individuals who receive treatment $D = 0$, and we need a relatively large number of individuals in the control group so that the the comparison between the treated and the control group is meaningful. On the other hand, the treated individuals can receive varied treatments, each with a potentially different treatment dose/intensity $D = d \in [d_L, d_H]$ according to some continuous distribution. Moreover, we will assume throughout that assumption 3.1 holds for (D, X) so that the conditional probability $P(D = 0|X)$ and density $f_{D|X}(d|X)$ for $d > 0$ are well defined.

Remark 4.1. *To formalize the mixture distribution of the treatment variable, consider a*

[18] We are going to implicitly assume that the treatment status and treatment intensity are independently determined.

measure[19] $\nu = \delta_0 + \lambda$, *with* λ *being the Lebesgue measure and* δ_0 *being the Dirac delta at* 0. *Suppose* F_D *is the distribution of* D. *Then the density of* D *w.r.t.* ν *is given by* $dF_D/d\nu := \mathbf{1}\{D = 0\}P(D = 0) + \mathbf{1}\{D > 0\}f_D$ *with* f_D *being the probability density of* D *on* $[d_L, d_H]$. *In particular,* $F_D(0) = \int \mathbf{1}\{D = 0\}\frac{dF}{d\nu}d\nu = P(D = 0)$ *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 restrict our attention to the two-period $(t-1, t)$ models and, as in the usual DiD setting, suppose that no subject receives treatment at period 0, so we may suppress the time notation in treatment D_i. Let X_i denote the set of individual level covariates. 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* $Y_i = (1 - T_i)Y_{i,t-1} + T_iY_{i,t}$;

(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 D *satisfies* $supp(D) = \{0\} \sqcup [d_L, d_H]$ *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*

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

Assumptions 4.1 and 4.2 are standard in the DiD literature. In particular, Assumption 4.1 does not allow the covariates to vary over time, while Assumption 4.2(ii) requires that

[19]This ν is essentially the dominant measure ν_Y we discussed in section 3.1. We drop the subscript here to avoid the confusion in notation, as here the Y is not the variable that we are interested in establishing the density.

the sample is not stratified by the outcome, treatment, or covariates.[20] Moreover, Assumption 4.3 describes the requirements on the support of the treatment. Specifically, in the continuous DiD setting, the control group ($D = 0$) must have a positive measure, and the treated group must have a positive likelihood of being treated at any intensity $d \in (d_L, d_H)$.

We want to emphasize the importance of Assumption 4.4, the conditional parallel trends condition that generalizes the discrete case of Abadie (2005), as the main identifying assumption that enables us to identify the causal parameter of interest. This assumption essentially states that, conditional on covariates, the unobserved counterfactual trend of the treated *at each given treatment intensity* is the same as the observed trend of the control group. In other words, the conditional parallel trends assumption 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. Importantly, our extension of this assumption to the continuous treatment setting allows us to consider the heterogeneity in another dimension: the treatment intensity.

As commented in Abadie (2005), the covariates in DiD can serve two purposes, which also apply to our continuous treatment setting. First, covariates can be used to account for compositional differences between control and treatment groups that affect outcome dynamics. Moreover, covariates allows researchers to capture the heterogeneous treatment effects across different groups/individuals characterized by the covariates. In particular, the conditional parallel trends assumption allows us to explicitly incorporate the covariates in DiD nonparametrically, in contrast to commonly used parametric approaches in the literature, such as a linear model, which can potentially introduce misspecification biases.

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

$$ATT(d) := E[Y_t(d) - Y_t(0)|D = d]. \tag{17}$$

The interpretation of this parameter is analogous to the cases with discrete treatment: the expected effect of a treatment with intensity d for those who actually received treatment with intensity d. Note that ATT is a local measure, and in the absence of stronger assumptions, the average treatment effect $ATE(d) := E[Y_t(d) - Y_t(0)]$, which is the expected effect of treatment with intensity d across the entire population, is not identified[21].

The following theorem presents the main results of this section, in which we establish the

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

[21] We note that ATE in this setting can be identified under a stronger form of parallel trends assumption and can be shown to be numerically equivalent to ATT, see Callaway et al. (2021) Section 3.3 for details.

identifications of $ATT(d)$ for both repeated outcomes and repeated cross-sections settings.

Theorem 4.1 (Identification of ATT)**.**

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

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

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

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

  where $\lambda := P(T = 1)$.

Here we use the repeated outcomes case to illustrate the main idea. The proof for the repeated cross-sections case is similar and is deferred to the appendix. We beginning by writing the ATT as

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

First, by the modeling assumptions that $Y_t = Y_t(D)$ and $Y_{t-1} = Y_{t-1}(0)$ since no one receives treatment in the pre-treatment period, we have

$$E[Y_t(d) - Y_{t-1}(0)|D = d] = E[Y_t - Y_{t-1}|D = d]. \tag{18}$$

Second, by the law of iterated expectation, Bayes' rule, and conditional parallel trends assumption, we can express the counterfactual quantity as follows:

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

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

$$= E\left[(Y_t - Y_{t-1})\mathbf{1}\{D = 0\}\frac{f_{D|X}(d|X)}{f_D(d)P(D = 0|X)}\right] \tag{19}$$

Subtracting (19) from (18), we obtain the desired result. In particular, in the third equality in (19), we substitute the unobserved counterfactual trend $E[(Y_t(0)-Y_{t-1}(0))|X = x, D = d]$ by the observed trend $E[(Y_t(0) - Y_{t-1}(0))|X = x, D = 0]$ of the control group, which is allowed by the conditional parallel trends assumption.

With Theorem 4.1, one can build estimators for $ATT(d)$ using the estimated sample analogues. For potentially high-dimensional covariates, machine learning methods can be employed to estimate the nuisance parameters, including the conditional density $f_{D|X}(d|X)$ and the conditional probability $P(D = 0|X)$. However, the use of machine learning methods can often result in non-trivial first-order biases in the estimation of the causal parameter[22], which makes such "plug-in" estimators less desirable. One way to alleviate such biases is to consider alternative estimating equations that reduce the "sensitivity" of the causal parameters to the nuisance parameters. We formalize this idea in detail in the next section.

4.2 Orthogonal Scores

First, recall that in the repeated outcomes case,

$$\begin{aligned} ATT(d) =& E[\Delta Y|D = d] - E\Big[\underbrace{\Delta Y\mathbf{1}\{D = 0\}\frac{f_{D|X}(d)}{f_D(d)P(D = 0|X)}}_{:=\varphi}\Big] \\ :=& E[\Delta Y|D = d] - \theta_0 \end{aligned}$$

where $\Delta Y := Y_t - Y_{t-1}$ and $\theta_0 := E[\varphi]$. Since the potentially high-dimensional covariates X only affects $ATT(d)$ through infinite-dimensional nuisance parameters $f_{D|X}(d|X)$ and $P(D = 0|X)$ via a function φ, we focus on $\theta_0 = E[\varphi]$. In particular, we need to adjust φ such that the first-order biases from estimating the infinite-dimensional nuisance parameters are negligible. To make this statement more precise, we introduce the notion of *Neyman orthogonality*. We use the repeated outcomes case as our main example for illustration as the analogous discussion on repeated cross-sections only requires minor modifications.

For simplicity, consider the following notations: let $\theta_0 \in \Theta \subset \mathbf{R}$ be the low-dimensional parameter of interest; let $\rho_0 \in \mathcal{H}$ denote the true low-dimensional nuisance parameters, e.g.,

[22] See CCDDHNR (2018) and references therein for a detailed discussion

in the repeated outcomes case, $\rho_0 = f_D(d)$ for a given d; let $\eta_0 \in \mathcal{T}$ denote the true infinite-dimensional nuisance parameters, which in our case include[23] $f_{D|X}(d|X)$ and $P(D = 0|X)$; let $\mathcal{T}_n \subset \mathcal{T}$ be a nuisance realization set in which the estimated $\hat{\eta}$ takes values with high probability; let Z be the observable random vector, e.g. $Z = (Y_{t-1}, Y_t, D, X)$ in the repeated outcomes setting; let $\psi : (Z, \theta, \rho, \eta) \mapsto \mathbf{R}$ denote a score[24].

With these notations, following CCDDHNR (2018) and Chang (2020), we formally define the Neyman orthogonality with respect to the infinite-dimensional nuisance parameters:

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) θ_0 satisfies the moment condition $E_P[\psi(Z, \theta_0, \rho_0, \eta_0)] = 0$;

(ii) for $r \in [0, 1)$ and $\eta \in \mathcal{T}_n$, the Gateaux (directional) derivative satisfies

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

In the above definition, (ii) ensures that the first-order bias from estimating the *infinite-dimensional* nuisance parameters is zero. We will construct scores that satisfy this orthogonality condition with some modifications that we will clarify shortly. Recall that in our case, $\varphi - \theta_0$ can be considered as a score since

$$E[\varphi - \theta_0] = E\Big[\Delta Y \mathbf{1}\{D = 0\} \frac{f_{D|X}(d|X)}{f_D(d)P(D = 0|X)} - \theta_0\Big] = 0.$$

This expression has two features that are worth noting. First, if $f_D(d)$ is estimated non-parametrically, e.g. using a kernel density estimator, we can no longer achieve root-N rate when estimating θ_0. The slower than root-N rate 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 $\varphi - \theta_0$ does not satisfy Neyman orthogonality, and an adjustment term has to be added.

In general, the adjustment term is straightforward to construct if the nuisance parameters can be written as conditional expectations. However, in our case, while $P(D = 0|X)$ can be expressed as a conditional expectation $E[\mathbf{1}\{D = 0\}|X]$, $f_{D|X}(d|X)$ being the conditional density presents additional challenges. To address this issue, we use a modified

[23] New infinite-dimensional nuisance parameters can arise when constructing the orthogonal scores.

[24] We say ψ is a score function if at the true nuisance parameters (ρ_0, η_0) and the true θ_0, the moment condition $E[\psi(Z, \theta_0, \rho_0, \eta_0)] = 0$ holds.

series representation of the conditional density introduced in Section 3, which allows us to approximate the conditional density using a finite series of conditional expectations.

Specifically, let $\{\phi_j\}_{j=1}^{\infty}$ be an orthonormal basis of D, and for a strictly positive $d \in (d_L, d_H)$, we can represent $f_{D|X}(d|X)$ as

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

Then, under suitable regularity conditions,[25]

$$\underbrace{E[\Delta Y \mathbf{1}\{D = 0\}\frac{f_{D|X}(d|X)}{f_D(d)P(D = 0|X)}]}_{:=\ \theta_0} = \lim_{J\to\infty} E[\underbrace{\Delta Y \mathbf{1}\{D = 0\}\frac{f_J(d|X)}{f_D(d)P(D = 0|X)}}_{:=\ \varphi_J}]$$

where $f_J(d|X) := \sum_{j=1}^{J} E[\phi_j(D)\mathbf{1}\{D > 0\}|X]\phi_j(d)$. 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: $m_J^d(D) := \sum_{j=1}^{J} \phi_j(D)\phi_j(d)\mathbf{1}\{D > 0\}$; $g(X) := P(D = 0|X)$; $\mathcal{E}_{\Delta Y}(X) := E[\Delta Y \mathbf{1}\{D = 0\}|X]$; $\mathcal{E}_{\lambda Y}(X) := E\left[\frac{T-\lambda}{\lambda(1-\lambda)} Y \mathbf{1}\{D = 0\}|X\right]$ with $\lambda = P(T = 1)$; $f_d := f_D(d)$. The following lemma introduces scores that satisfy Neyman orthogonality.

Lemma 4.1. *Suppose there exists $M_J^{(1)} \in L^1(P_{Y_{t-1},Y_t,D,X})$ and $M_J^{(2)} \in L^1(P_{Y,T,D,X})$ such that $|\psi_J^{(1)}| \leq M_J^{(1)}$ and $|\psi_J^{(2)}| \leq M_J^{(2)}$ almost surely. Then the scores $\psi_J^{(1)}$ and $\psi_J^{(2)}$ satisfy Neyman orthogonality defined in (4.1), where*

(i) for the repeated outcomes setting,

$$\begin{aligned}\psi_J^{(1)} := \Delta Y \mathbf{1}\{D = 0\}\frac{f_J(d|X)}{f_d \cdot g(X)} - \theta_{0,J} \\ + \frac{m_J^d(D)g(X) - \mathbf{1}\{D = 0\}f_J(d|X)}{f_d \cdot g^2(X)}\mathcal{E}_{\Delta Y}(X);\end{aligned} \tag{20}$$

[25] 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.

(ii) for the repeated cross-sections setting,

$$\begin{aligned}\psi_J^{(2)} := \frac{T-\lambda}{\lambda(1-\lambda)} Y \mathbf{1}\{D=0\} \frac{f_J(d|X)}{f_d \cdot g(X)} - \theta_{0,J} \\ + \frac{m_J^d(D) g(X) - \mathbf{1}\{D=0\} f_J(d|X)}{f_d \cdot g^2(X)} \mathcal{E}_{\lambda Y}(X).\end{aligned} \tag{21}$$

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. The assumption on the existence of integrable functions $M_J^{(1)}$ and $M_J^{(2)}$ is a mild regularity condition that allows us to interchange expectation and derivative. This assumption can be readily checked under the boundedness of the nuisance parameters in the scores, which will be made precise in the next section. For notational simplicity, we drop the superscripts on $\psi_J^{(1)}$ and $\psi_J^{(2)}$ whenever the context is clear.

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. In particular, the estimating moments for $\theta_{0,J}$'s based on these orthogonal scores are not sensitive to potentially biased estimates of these nuisance parameters. In the next section, we will construct DML estimators of $\theta_{0,J}$'s using these scores and establish their asymptotic properties.

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.

Since (20) is a score, we have $E[\psi_J] = 0$ for any J, from which we obtain a moment condition for the target parameter $\theta_{0,J}$:

$$\theta_{0,J} = E\Big[\Delta Y \mathbf{1}\{D=0\} \frac{f_J(d|X)}{f_d \cdot g(X)} + \frac{m_J^d(D) g(X) - \mathbf{1}\{D=0\} f_J(d|X)}{f_d \cdot g^2(X)} \mathcal{E}_{\Delta Y}(X)\Big]. \tag{22}$$

Then, the $ATT(d)$ is identified as $ATT(d) = E[\Delta Y|D=d] - \lim_{J\to\infty} \theta_{0,J}$, which suggests we can approximate $ATT(d)$ as

$$\begin{aligned}ATT(d) \approx E\Big[E[\Delta Y|D=d] - \Delta Y \mathbf{1}\{D=0\} \frac{f_J(d|X)}{f_d \cdot g(X)} + \\ \frac{m_J^d(D) g(X) - \mathbf{1}\{D=0\} f_J(d|X)}{f_d \cdot g^2(X)} \mathcal{E}_{\Delta Y}(X)\Big]\end{aligned} \tag{23}$$

for some "large" J. Therefore, we can construct an estimator for $ATT(d)$ based on the sample analogue of (23). In particular, since ψ_J is an orthogonal score, for high-dimensional X, we can use machine learning methods to estimate the nuisance parameters without having to worry about the first order biases from doing so.

Moreover, another key aspect of the DML estimator is the use of cross-fitting techniques to reduce the overfitting bias from estimating the nuisance parameters using machine learners. We follow the precedure studied in CCDDHNR (2018) and the main idea is as follows. First, we partition the random sample into $K \geq 2$ disjoint subsets $\{I_k\}_{k=1}^{K}$ of equal size $n = N/K$. Then, for each $k \in \{1, \cdots, K\}$, we use the sample $I_k^c := N \setminus I_k$ to estimate the nuisance parameters with the preferred machine learning methods. Next, we compute sample averages according to (23) using the estimated nuisance parameters evaulated at the sample I_k to obtain the k-th estimate $\widehat{ATT(d)}_k$ for $ATT(d)$. Finally, we average through the K estimates to obtain the final estimator. The following algorithm summarizes the procedure.

Algorithm 4.1 (CDID Estimator, Repeated Outcomes). *Let $\{I_k\}_{k=1}^{K}$ denote a random partition of a random 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 $I_k^c := N \setminus I_k$ denote the complement.*

- *Step 1: for each k, construct*

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

  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. In particular, $\hat{f}_{d,k}, \hat{\mathcal{E}}^d_{\Delta Y,k}$ are kernel estimators, $\hat{g}_k, \hat{\mathcal{E}}_{\Delta Y,k}$ are estimated using ML methods (e.g. deep neural networks), and each term in $\hat{f}_{J,k}$ is estimated estimated using ML for a large J.

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

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

Remark 4.2. *It is important to note that at each $k = 1, \cdots, K$, the nuisance parameters and the ATT are estimated using disjoint subsamples. While doing so helps reduce the overfitting bias, the sample splitting also significantly simplifies the asymptotic analysis, which* *itself has a long history in the literature (see CCDDHNR (2018) and references therein). Moreover, the cross-fitting ensures that the final estimator uses the full sample, and hence the choice of K does not affect the asymptotic analysis of our estimator. In practice, we recommend using $K = 5$ as a rule of thumb.*

As we will establish shortly, to achieve valid inference results, we need an undermoothing J for the conditional density estimator $\hat{f}_J(d|X)$. This is the main reason why we opt to use a large J instead of the cross-validated J in section 3 that is shown to balance the variance and bias and hence may fail to be under-smoothing. Alternatively, we can also consider using the cross-validated J multiplied by a term that grows with the sample size. 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 (Kernel). *The kernel K satisfies:*

(i) *K is bounded and differentiable;*

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

and define $K_h(u) := h^{-1}K(u/h)$.

Assumption 4.6 (Orthonormal Basis). $\{\phi_j\}_{j=1}^{\infty}$ *is an orthonormal basis on the support of D such that*

(i) $m_J^d(D) = \sum_{j=1}^{J} \phi_j(D)\phi_j(d)\mathbf{1}\{D > 0\}$ *satisfies* $\|m_J^d(D)\|_\infty \leq M_J$ *for some constant M_J that grows with J;*

(ii) $E[(m_J^d(D))^2] \asymp \tilde{M}_J^2$ *and* $E[|m_J^d(D)|^3] \asymp \tilde{M}_J^3$ *for some constant $\tilde{M}_J$ that grows with J.*

Assumption 4.7 (Bounds).

(i) *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;*

(ii) *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;*

(iii) *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 4.8 (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) $M_J/\sqrt{N} = o(1)$*;*

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

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

We use kernel to estimate the low-dimensional parameters $f_D(d)$ and $E[\Delta Y|D = d]$ given its well-established theoretical properties and these kernel estimators will play a role in the asymptotic distributions of our estimator for ATT. We assume the standard regularity conditions for kernel estimators in assumption 4.5, which are sufficient for a triangular array CLT to hold. Assumption 4.6 is a set of regularity conditions on the orthonormal basis: (i) is stated in very general terms and can usually be verified by the choice of the orthonormal basis; similar to the assumptions on the kernel, (ii) is sufficient for Lyapunov conditions[26] to hold so that a triangular array CLT can apply as J growing with n, which can be checked with additional assumptions on the orthonormal basis (e.g. trigonometric basis or Hermite basis). Assumption 4.8 concerns the quality of the nonparametric estimators: (i) requires under-smoothing tuning parameters so that the bias vanishes asymptotically (otherwise asymptotic normality still holds but not centered at θ_0); (iii) is the standard assumptions on the nuisance estimators in the DML 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 $M_J N^{-1/4}$. 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 4.2 (Repeated Outcomes)**.** *Suppose assumptions 4.1, 4.3, 4.4, 4.5, 4.6, 4.7, and 4.8 hold. If* $\varepsilon_N = o(N^{-1/4})$, *then*

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

where

$$\sigma_N^2 := E\Big[\Big(\frac{1}{f_d}(K_h(D-d)\Delta Y - E[K_h(D-d)\Delta Y])$$

[26] Alternatively, we can make a set of alternative assumptions and check the weaker Lindeberg's conditions.

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

and ψ_J is defined as in (20)

The proof follows the general framework for DML estimators 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, which holds with only minor modifications.

With a consistent estimator $\hat{\sigma}^2_N$ 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

$$\begin{aligned}
\hat{\sigma}^2_N :=& \frac{1}{K}\sum_{k=1}^{K} E_{n,k}\Big[\Big(\frac{1}{\hat{f}_{d,k}}(K_h(D-d)\Delta Y - E_{n^c,k}[K_h(D-d)\Delta Y]) \\
&- \psi_J(Z, \hat{\theta}_J, \hat{f}_{d,k}, \hat{\eta}_k) \\
&+ (\frac{\hat{\theta}_J}{\hat{f}_{d,k}} - \frac{\hat{\mathcal{E}}^d_{\Delta Y,k}}{\hat{f}_{d,k}})(K_h(D-d) - E_{n^c,k}[K_h(D-d)])\Big)^2\Big]
\end{aligned} \tag{24}$$

where

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

and $E_{n^c,k}$ denotes the empirical average using the auxiliary sample I^c_k. 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, we can consider a multiplier bootstrap procedure to construct the confidence interval for our estimator, which has been discussed extensively 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 $E[\xi_i] = Var(\xi_i) = 1$.

Then for each $b = 1, \cdots, B$, we draw such a sequence $\{\xi_i\}_{i=1}^N$ and construct

$$\widehat{ATT(d)}_b^* := \frac{1}{N} \sum_{k=1}^K \sum_{i \in I_k} \xi_i \Big(\hat{\mathcal{E}}_{\Delta Y,k}^d - \Delta Y_i \mathbf{1}\{D_i = 0\} \frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k} \cdot \hat{g}_k(X_i)} - \frac{m_J^d(D_i)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k} \cdot \hat{g}_k^2(X_i)} \hat{\mathcal{E}}_{\Delta Y,k}(X_i) \Big). \tag{25}$$

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 $[\widehat{ATT(d)} - \hat{c}_{1-\alpha/2}, \widehat{ATT(d)} - \hat{c}_{\alpha/2}]$. We defer the theoretical discussions of this procedure to future work.

Next, we apply our methods to two empirical applications where the research designs can be reframed as continuous DiD.

4.4 Empirical Application 1: Acemoglu and Finkelstein (2008)

The Medicare Prospective Payment System (PPS) reform, introduced in 1983, changed the way Medicare reimburses hospitals for inpatient care. Instead of a full-cost reimbursement model based on actual expenses, hospitals began receiving a predetermined amount per patient based on the diagnosis. Notably, during the first three years[27] post-reform, reimbursements for capital costs still reflected actual expenses. This meant that hospitals treating Medicare patients experienced a relative increase in labor costs compared to capital costs. Acemoglu and Finkelstein (2008) highlighted this unique aspect of the PPS reform. Their research revealed that the PPS reform not only significantly raised the capital-labor ratio in hospitals but also promoted the adoption of new technologies.

Specifically, one of the main theoretical predictions in Acemoglu and Finkelstein (2008) posits that the PPS reform would result in a higher capital-labor ratio in hospitals. Furthermore, if the elasticity of substitution between capital and labor is sufficiently large, PPS reform should lead to an increase in demand for capital/technology. It is important to note that, since only hospitals with Medicare patients are affected by this reform, these effects should be bigger for hospitals with higher shares of Medicare patients. To test these predictions empirically, Acemoglu and Finkelstein (2008) uses data from the Annual American Hospital Association (AHA) survey of hospitals from 1980 to 1986, which contains information on hospital expenditure, employment, and other characteristics related to the technologies at the hospital level.

[27] In fact, as noted in Acemoglu and Finkelstein (2008), there was no change to the Medicare's reimbursement for capital costs until 1991 due to various delays.

The baseline specification in Acemoglu and Finkelstein (2008) takes the following form of a linear regression:

$$Y_{i,t} = \alpha_i + \gamma_t + X_{it}'\eta + \beta \cdot (D_i \cdot \text{Post}_t) + \varepsilon_{i,t}, \tag{26}$$

where $Y_{i,t}$ denotes either the capital-labor ratio or the total number of medical facilities[28] of hospital i in year t, D_i denotes the share of Medicare inpatient days in hospital i *prior* to the PPS reform, $\text{Post}_t = \mathbf{1}\{t \in \text{post-PPS years}\}$ denotes the treatment-timing indicator, X_{it} denotes a vector of covariates, and α_i and γ_t denote hospital and year fixed effects respectively. Acemoglu and Finkelstein (2008) argue that the coefficient β captures the causal effect of the PPS reform on the capital-labor ratio or the technological adoption. The main identifying assumption is that, in the absence of the PPS reform, hospitals with different shares D_i should have experienced similar changes in outcome variables over time, i.e., a parallel trends assumption.

Notably, regression in (26) closely resembles the commonly used Two-Way Fixed Effects (TWFE) design, with an important distinction that the treatment variable D_i here is continuous. In fact, as pointed out in Callaway et al. (2021), with continuous treatment, the coefficient β in (26) can be expressed as a weighted average of the $ATT(d)$ over all the treatment intenties with potentially *negative* weights[29], which makes β difficult to interpret. This is where our continuous DiD framework can be useful. In particular, we can reframe the research design in Acemoglu and Finkelstein (2008) as a continuous DiD design with the following setup:

- Prior to the PPS reform, no hospital was treated.
- Since the PPS reform only affected hospitals with Medicare patients, hospitals with Medicare share $D_i = 0$ can serve as the control group.
- The treatment group consists of hospitals with positive Medicare shares $D_i > 0$. Since the Medicare shares differ widely across hospitals, we can model the positive shares as continuous treatment intensities.
- We consider the same outcome variables as the ones in (26): Y_{it} can be either the capital-labor ratio or some measures of technological adoption.
- We assume a conditional parallel trends assumption:

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

[28]The total number of facilities can be used as a measure of technological adoption.

[29]See Proposition 10 in Callaway et al. (2021).

That is, on average, in the absence of the PPS reform, the outcome variables of hospitals with share $D_i = d$ should have experienced similar changes over time as hospitals with no Medicare patients (shares $D_i = 0$), *conditional on a set of hospital-specific covariates X determined prior to the PPS reform.* We note that this assumption strengthens the parallel trends assumption in Acemoglu and Finkelstein (2008) by allowing covariates X to enter the identification nonparametrically.

- We also include a rich set of covariates X that are determined prior to the PPS reform: number of beds, number of doctors/residents, whether in a metro area, and a full set of states (or regions) dummies[30]. In addition, when the outcome variable is the capital-labor ratio, we will include a set of binary variables that indicate whether the hospital has a particular type of capital equipment (e.g., CT, MRI, etc.).

The causal effect of the PPS reform can be identified as the average treatment effect on the treated (ATT) at each intensity d:

$$ATT(d) = E[Y_t(d) - Y_t(0)|D = d].$$

Importantly, in contrast to the constant β in (26), the causal parameter $ATT(d)$ can be directly employed to validate the main theoretical predictions of Acemoglu and Finkelstein (2008) at a much more granular level. For example, the prediction that the PPS reform should lead to an increase in capital-labor ratio can be validated if $ATT(d) > 0$ for all $d > 0$. Moreover, the prediction that hospitals with higher shares of Medicare inpatients should experience a greater increase in capital-labor ratio would hold if $ATT(d)$ is increasing in d.

In fact, there are two potential methods to estimate $ATT(d)$. First, the dataset in Acemoglu and Finkelstein (2008) possesses a panel structure, allowing us to utilize our estimator for the repeated outcomes case. As an illustration, the year 1983 to be designated as the pre-treatment year $(t-1)$, while any subsequent years can be considered as the post-treatment year (t). On the other hand, given that the treatment intensity D represents the Medicare share – information available for all years both prior to and following the PPS reform – we can also employ our estimator for the repeated cross-sections scenario. Therefore, we demonstrate our methods in both cases, specifically:

- In the repeated outcomes setting, we set $t - 1 = 1983$ and, for each $t \geq 1984$, estimate the ATT at various treatment intensities. The outcome variables under consideration

[30] There are several other covariates that were mentioned in Acemoglu and Finkelstein (2008), including whether the hospital is a general hospital, a short-term hospital, or a federal hospital. We opt to not include these covariates since they can be used to determine a hospital's exemption status from the PPS reform and hence can violate the conditional parallel trends assumption.

are capital-labor ratio and a measure of technological adoption (number of medical facilities)[31].

- In the repeated cross-sections setting, we also set $t-1=1983$ and estimate ATT at various treatment intensities for each $t \geq 1984$. To provide a clearer illustration of this concept, we center our analysis on the capital-labor ratio.

To begin with our analysis, let's first examine the distribution of the treatment variable, defined as the Medicare inpatient share for each hospital in 1983 prior to the PPS reform. Figure 1 depicts the histogram of D for 1983, and the distribution of D is suitable for our

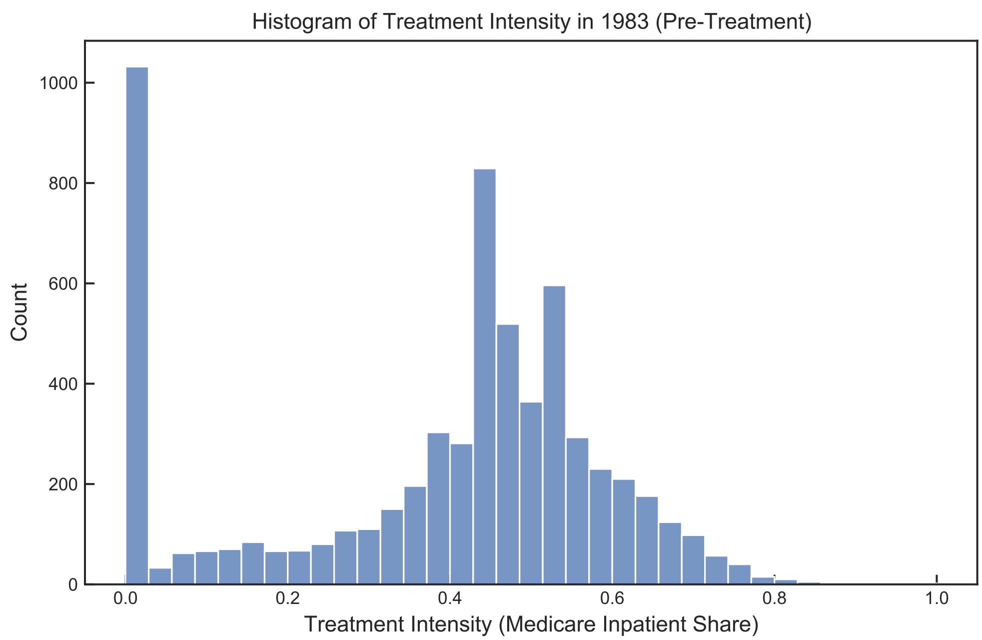

Figure 1: Histogram of Treatment Intensity (Medicare Share in 1983)

continuous DiD framework. Specifically, a significant number of hospitals register at $D=0$, enabling us to consider these hospitals as the control group. Moreover, the positive Medicare shares ($D>0$) varies widely across hospitals and appear to follow a continuous distribution, which allows us to view these positive shares as continuous treatment intensities.

We now turn to the results for the repeated outcomes (panel) setting, where the outcome variable is the capital-labor ratio. In particular, using $t-1=1983$ as the pre-treatment

[31] When the outcome variable is the technological adoption, we do not consider year 1986 due to data availability.

year, we estimate the causal parameter ATT(d) at various intensity d ranging from 0.1 to 0.8 for each $t = 1984, 1985, 1986$. The results are shown in Table 1 and Figure 2. In the table, we provide standard errors in parentheses as well as bootstrap confidence intervals. In Figure 2, we only plot the estimated ATTs for various intensities d and opt to omit the confidence intervals for clearer visualization. It is crucial to note that all the estimates are statistically significant.

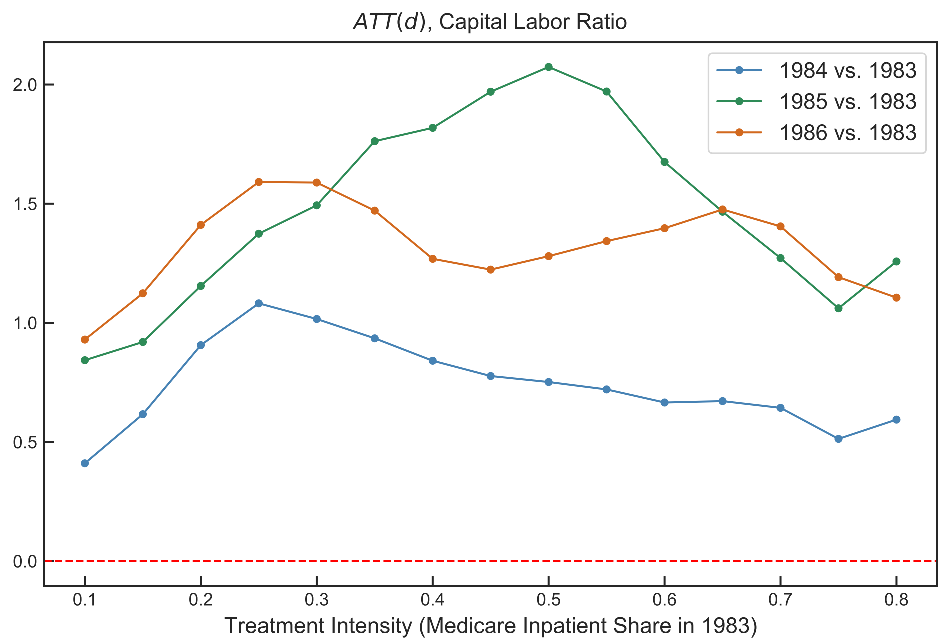

Figure 2: Estimated ATT(d) for Capital-Labor Ratio (Panel Data)

Specifically, we observe that all the estimated ATTs for the capital-labor ratio are positive, which corroborates the empirical findings in Acemoglu and Finkelstein (2008) and provides further evidence that the PPS reform led to an increase in the capital-labor ratio. Moreover, compared to the results from $t = 1984$, the estimates for $t = 1985$ and $t = 1986$ are much larger in magnitude, which implies that the hospitals respond to the PPS reform gradually. For comparison, the estimated β in Acemoglu and Finkelstein (2008) is 1.13 for the capital-labor ratio, which is larger than our estimates for $t = 1984$ but much smaller for many of our estimates for $t = 1985$ and $t = 1986$. Interesting, such differentials by year are consistent with the alternative research specifications in Acemoglu and Finkelstein (2008) (see Table 2 column (3)), which also found that the impact of the PPS reform is incremental over time.

Finally, for all three years, the estimated ATTs vary widely across treatment intenties and don't display increasing trends, which is inconsis tent with the theoretical prediction that hospitals with higher Medicare shares should experience a more substantial increase in the capital-labor ratio. One possible explanation is that our estimates are not precise enough to detect such pattern. Notebly, even though all our estimates are statistically significantly different from zero, the associated confidence intervals are relatively wide, which is an inherent feature given the relatively small sample size for using nonparametric methods.

Similarly, we present evidence of increased technological adoption following the PPS reform. The outcome variable here is the total number of various medical facilities in each hospital, which can be used as a measure of technological adoption. As with our prior analysis, we designate $t - 1 = 1983$ as the pre-treatment year. However, due to data availability, we restrict our analysis of post treatment years to 1984 and 1985. We then estimate the causal parameter ATT(d) at varying intensities d ranging from 0.1 to 0.8 for both $t = 1984$ and $t = 1985$. The findings are shown in Table 2 and Figure 3.

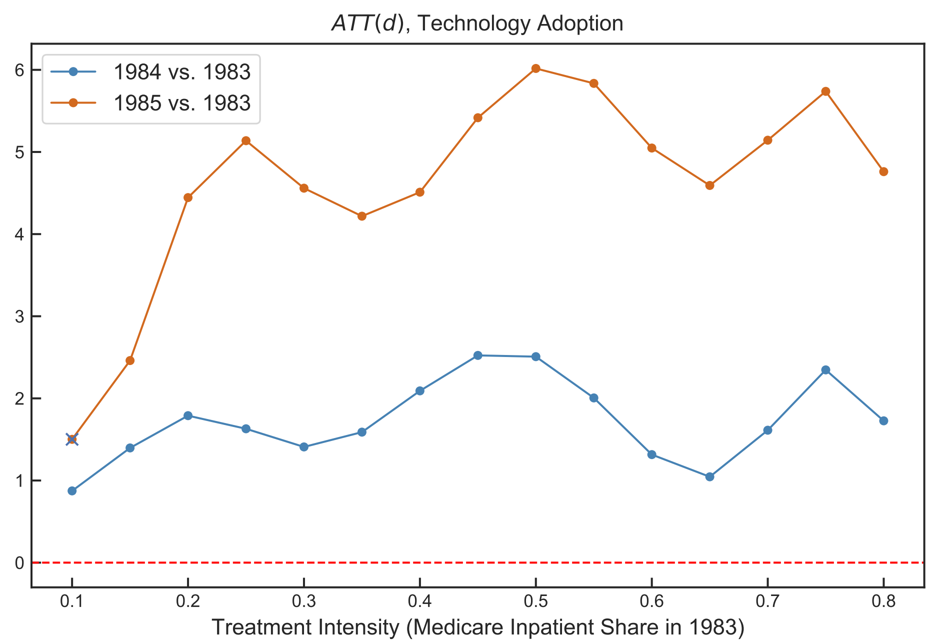

Figure 3: Estimated ATT(d) for Technological Adoption (Panel Data)

Figure 3 further reveals that the estimated ATTs for technological adoption are positive at all the treatment intensities we considered. This validates the theoretical prediction in Acemoglu and Finkelstein (2008) that the PPS reform should lead to an increase in the technological adoption. Moreover, similar to the findings for the capital-labor ratio, the 1985 estimates are much more substantial in magnitude comparing to their 1984 counterparts, further suggesting that the impact of the PPS reform is incremental over time. Finally, for both years, the estimates are increasing for lower treatment intensities and leveling off for higher treatment intensities. This trend partially validates the theoretical prediction that hospitals with higher Medicare inpatient shares should experience a bigger increase in the technological adoption following the PPS reform [32].

In our analysis thus far, we have adhered to the research design of Acemoglu and Finkelstein (2008), utilizing the Medicare share from 1983 as our quasi-experimental variation for causal analysis. However, it is crucial to acknoledge the potential changes in Medicare share as a result of the PPS reform. Specifically, the PPS reform could lead to a reduction in the Medicare share for hospitals with positive shares initially. Indeed, a comparison of the histograms of Medicare share between 1983 and 1985, as displayed in Figure 4, reveals a leftward shift in the distribution:

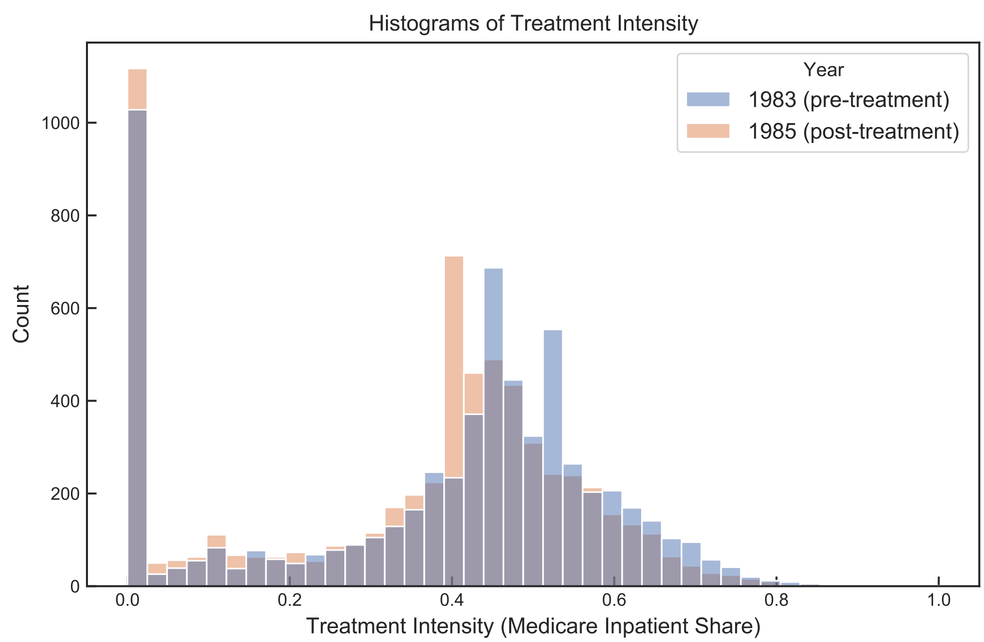

Figure 4: Histograms of Treatment Intensity (1983 vs. 1985)

[32]We need to be cautious when comparing the estimates for different treatment intensities since the confidence intervals are relatively wide, even though all but one estimates are statistically significant from zero.

Therefore, to account for the changes in the treatment intensity (Medicare share), we treat the data as repeated cross-sections and apply our estimator accordingly. Specifically, we focus on the capital-labor ratio as our main outcome variable, and we estimate ATTs across a wide range of treatment intensities for $t-1 = 1983$ and $t = 1983, 1984, 1985$. The results from our repeated cross-sections methods, as shown in Table 3, differ considerably from those in the panel setting. As a highlight, we plot the results for 1985 in Figure 5.

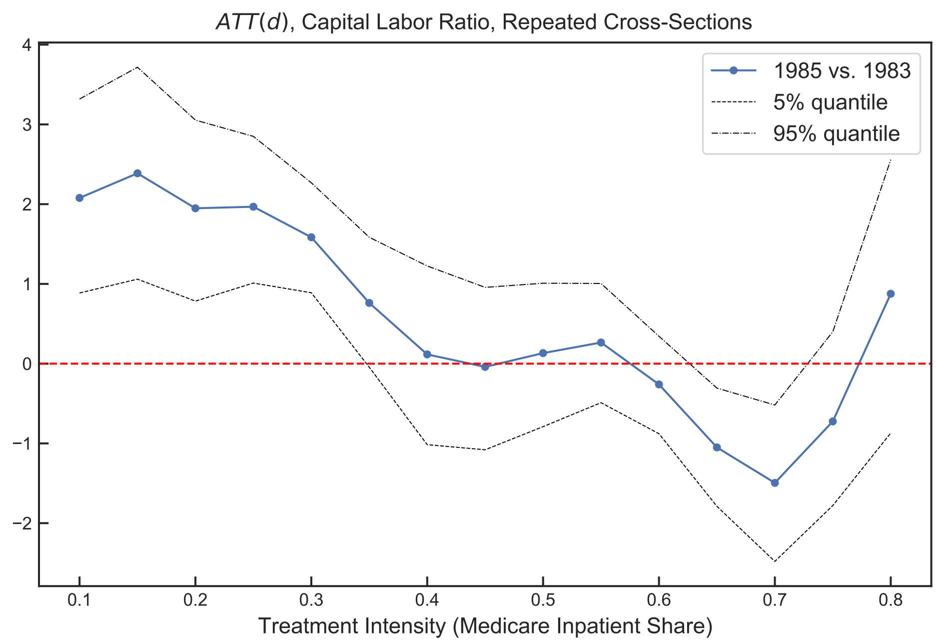

Figure 5: Estimated ATT(d) for Capital-Labor Ratio (Repeated Cross-Sections)

Notably, most of the estimates for year 1984 are not significantly different from zero. On the other hand, for 1985, as shown in Figure 5, the estimated ATTs are positive and large for low treatment intensities. However, as the treatment intensity increases, these estimates decrease in magnitude and can even become negative. A similar trend is evident for 1986. This pattern markedly differs from what we see in the panel setting, where the estimated ATTs are consistently positive across all treatment intensities. These findings suggest that the PPS reform could lead to a decrease in the capital-labor ratio for hospitals with high Medicare inpatient shares, which is in contradiction to the theoretical predictions of Acemoglu and Finkelstein (2008). One possible explanation is that hospitals with high volume of Medicare inpatients might have developed administrative and clinical systems to effectively manage these patients, making it easier to adapt to PPS changes. Nevertheless, further investigations and formal theoretical analysis are needed to understand the underlying mechanism behind this phenomenon.

Remark 4.3. *The estimators for the causal parameters are constructed based on the results from the previous section. Here are the details of our implementation:*

- *A 5-fold cross-fitting is employed with data randomly shuffled before the sample splitting step*[33]*.*
- *The second-order Gaussian kernel with bandwidth $h = O(N^{-1/4})$ is used to estimate the density $f_D(d)$ and the conditional mean $E[\Delta Y|D = d]$.*
- *The infinite-dimensional nuisance parameters are estimated using the Random Forest (RF). The main advantage of using RF is that it can handle both continuous and discrete covariates, which is crucial for our analysis since our covariates X include both continuous variables and a large number of states dummies. However, we note that other ML methods, such as deep neural networks, can also be used to estimate the nuisance parameters.*
- *The orthonormal cosine basis on $[0, 1]$ is utilized to estimate the conditional density $f_J(d|X)$, where the series cutoff $J = O(N^{1/4})$ is chosen so that the under-smoothing assumption is more plausible; each conditional expectation in $f_J(d|X)$ is estimated using the ML method mentioned above.*
- *The standard errors are calculated using the cross-fitted estimator defined in (24) and (44). In addition, we also prenset 90-percent bootstrap confidence intervals constructed using the multiplier bootstrap procedure defined in (25) and (45): Gaussian multipliers $\{\xi_i\}_{i=1}^N$ are drawn from a normal distribution with $E[\xi_i] = Var[\xi_i] = 1$ for $B = 1000$ repetitions.*

[33]To avoid having clusters of data being over-represented in the subsamples.

Table 1: Estimated ATT(d) for Capital-Labor Ratio (Panel)

	($t = 1984$) ATT(D=d)	Bootstrap CI	($t = 1985$) ATT(D=d)	Bootstrap CI	($t = 1986$) ATT(D=d)	Bootstrap CI
d = 0.1	0.4105 (0.2269)	[0.0283, 0.7880]	0.8432 (0.2492)	[0.4181, 1.2435]	0.9297 (0.2423)	[0.5151, 1.3111]
d = 0.15	0.6165 (0.2683)	[0.1306, 1.0352]	0.9196 (0.2826)	[0.4656, 1.3646]	1.1239 (0.3008)	[0.6018, 1.6008]
d = 0.2	0.9064 (0.2999)	[0.3945, 1.3840]	1.1548 (0.5333)	[0.2944, 1.9581]	1.4110 (0.3303)	[0.8718, 1.9154]
d = 0.25	1.0818 (0.2616)	[0.6433, 1.5033]	1.3741 (0.5482)	[0.4625, 2.2294]	1.5910 (0.2821)	[1.1369, 1.9916]
d = 0.3	1.0161 (0.1780)	[0.7364, 1.2873]	1.4931 (0.5869)	[0.5107, 2.4309]	1.5887 (0.2507)	[1.1839, 1.9737]
d = 0.35	0.9350 (0.2034)	[0.6164, 1.2750]	1.7622 (0.7452)	[0.4935, 2.9886]	1.4709 (0.3260)	[0.9493, 1.9832]
d = 0.4	0.8410 (0.2262)	[0.4821, 1.2066]	1.8183 (0.7911)	[0.4812, 3.0997]	1.2686 (0.3424)	[0.7295, 1.8312]
d = 0.45	0.7768 (0.2648)	[0.3425, 1.2068]	1.9698 (0.8052)	[0.5605, 3.2337]	1.2234 (0.3453)	[0.6507, 1.7854]
d = 0.5	0.7517 (0.3316)	[0.1745, 1.2620]	2.0736 (0.8763)	[0.5411, 3.3901]	1.2798 (0.3951)	[0.5861, 1.8911]
d = 0.55	0.7206 (0.3114)	[0.1896, 1.1964]	1.9706 (0.8500)	[0.5072, 3.2700]	1.3430 (0.3893)	[0.6656, 1.9392]
d = 0.6	0.6657 (0.2144)	[0.3188, 0.9999]	1.6750 (0.7133)	[0.4863, 2.7614]	1.3971 (0.3303)	[0.8462, 1.9081]
d = 0.65	0.6716 (0.1978)	[0.3339, 0.9708]	1.4662 (0.6191)	[0.4409, 2.4900]	1.4757 (0.3406)	[0.9230, 2.0442]
d = 0.7	0.6433 (0.2064)	[0.2934, 0.9761]	1.2718 (0.4566)	[0.5338, 2.0447]	1.4047 (0.3523)	[0.8620, 1.9712]
d = 0.75	0.5131 (0.2533)	[0.1213, 0.9389]	1.0610 (0.4584)	[0.3335, 1.8266]	1.1917 (0.3670)	[0.6169, 1.7734]
d = 0.8	0.5939 (0.3197)	[0.0823, 1.1203]	1.2576 (0.4973)	[0.4764, 2.0920]	1.1058 (0.3752)	[0.5136, 1.6793]

Notes: (i) d indicates the treatment intensity; (ii) standard errors calculated using cross-fitted formula are shown in parentheses; (iii) 90%-CI using multiplier bootstrap shown in separate columns; (iv) for all post-treatment period $t = 1984, 1985, 1986$, the baseline pre-treatment year is $t = 1983$; (v) all the nuisance parameters are estimated nonparametrically using random forests.

Table 2: Estimated ATT(d) for Technological Adoption (Panel)

	($t = 1984$) ATT(D=d)	Bootstrap CI	($t = 1985$) ATT(D=d)	Bootstrap CI
d = 0.1	0.8735 (0.4965)	[0.0915, 1.7011]	1.5012 (1.0257)	[-0.1882, 3.2009]
d = 0.15	1.3966 (0.5939)	[0.4536, 2.3642]	2.4604 (1.3896)	[0.2910, 4.7959]
d = 0.2	1.7896 (0.5797)	[0.8658, 2.6735]	4.4447 (1.2381)	[2.4560, 6.4869]
d = 0.25	1.6298 (0.6205)	[0.6312, 2.6376]	5.1374 (0.8375)	[3.7879, 6.4412]
d = 0.3	1.4089 (0.4495)	[0.7211, 2.1478]	4.5592 (0.5956)	[3.6112, 5.5063]
d = 0.35	1.5890 (0.7052)	[0.5026, 2.6889]	4.2181 (0.8108)	[2.9679, 5.5625]
d = 0.4	2.0915 (1.2231)	[0.1672, 4.0489]	4.5103 (0.9589)	[3.0573, 6.0712]
d = 0.45	2.5231 (1.4911)	[0.1427, 4.8928]	5.4169 (1.0446)	[3.8158, 6.9531]
d = 0.5	2.5080 (1.3630)	[0.3598, 4.6075]	6.0168 (1.0569)	[4.3189, 7.5972]
d = 0.55	2.0055 (0.9108)	[0.5179, 3.3895]	5.8351 (0.9049)	[4.3630, 7.1993]
d = 0.6	1.3154 (0.5214)	[0.5011, 2.1110]	5.0480 (0.9178)	[3.5340, 6.5263]
d = 0.65	1.0451 (0.4889)	[0.2648, 1.8434]	4.5923 (1.3030)	[2.4368, 6.7603]
d = 0.7	1.6124 (0.6122)	[0.6022, 2.5588]	5.1433 (1.5093)	[2.6244, 7.6275]
d = 0.75	2.3469 (1.2027)	[0.4561, 4.2525]	5.7380 (1.3071)	[3.5262, 7.8050]
d = 0.8	1.7266 (0.9883)	[0.0718, 3.2958]	4.7605 (0.8721)	[3.3513, 6.1138]

Notes: (i) d indicates the treatment intensity; (ii) standard errors calculated using cross-fitted formula are shown in parentheses; (iii) 90%-CI using multiplier bootstrap shown in separate columns; (iv) for all post-treatment period $t = 1984, 1985, 1986$, the baseline pre-treatment year is $t = 1983$; (v) all the nuisance parameters are estimated nonparametrically using random forests.

Table 3: Estimated ATT(d) for Capital-Labor Ratio (Repeated Cross-Sections)

	($t = 1984$) ATT(D=d)	Bootstrap CI	($t = 1985$) ATT(D=d)	Bootstrap CI	($t = 1986$) ATT(D=d)	Bootstrap CI
d = 0.1	0.2353 (0.7875)	[-1.2893, 1.7216]	2.0781 (0.7167)	[0.8862, 3.3181]	2.9498 (0.7068)	[1.6976, 4.1190]
d = 0.15	0.0883 (1.3344)	[-2.7366, 2.6403]	2.3869 (0.8000)	[1.0594, 3.7171]	2.5135 (0.9556)	[0.8613, 4.0502]
d = 0.2	0.6917 (1.6479)	[-2.6378, 4.0499]	1.9482 (0.6948)	[0.7840, 3.0539]	2.4511 (1.0943)	[0.6327, 4.2523]
d = 0.25	1.3698 (1.0835)	[-0.7596, 3.4644]	1.9678 (0.5569)	[1.0108, 2.8500]	3.0380 (0.8298)	[1.6193, 4.3150]
d = 0.3	1.5204 (0.5881)	[0.4651, 2.6079]	1.5850 (0.4476)	[0.8892, 2.2682]	3.9194 (0.9160)	[2.3577, 5.3676]
d = 0.35	1.1936 (0.5139)	[0.2130, 2.2490]	0.7617 (0.4964)	[-0.0446, 1.5849]	4.0648 (0.9844)	[2.4131, 5.7084]
d = 0.4	0.8755 (0.6608)	[-0.3741, 2.1867]	0.1159 (0.6696)	[-1.0158, 1.2250]	3.4234 (1.0771)	[1.6636, 5.1449]
d = 0.45	0.6473 (0.6623)	[-0.6370, 1.8988]	-0.0407 (0.6219)	[-1.0800, 0.9564]	2.2998 (0.8918)	[0.8718, 3.8182]
d = 0.5	0.3921 (0.6865)	[-0.8886, 1.6837]	0.1321 (0.5542)	[-0.7900, 1.0085]	1.4233 (0.7015)	[0.2360, 2.6172]
d = 0.55	0.1512 (0.6113)	[-1.0443, 1.2511]	0.2651 (0.4626)	[-0.4893, 1.0051]	1.1309 (0.5948)	[0.1526, 2.1676]
d = 0.6	-0.0071 (0.4919)	[-0.8876, 0.9202]	-0.2605 (0.3994)	[-0.8790, 0.3434]	0.8571 (0.6139)	[-0.1388, 1.8580]
d = 0.65	-0.0453 (0.5674)	[-1.1205, 1.0748]	-1.0503 (0.4693)	[-1.7891, -0.3067]	0.0161 (0.7651)	[-1.3120, 1.2156]
d = 0.7	0.0866 (0.8271)	[-1.5172, 1.6274]	-1.4948 (0.6167)	[-2.4798, -0.5188]	-0.7250 (0.9520)	[-2.3715, 0.9577]
d = 0.75	0.2596 (1.2530)	[-2.0182, 2.5673]	-0.7242 (0.6966)	[-1.7790, 0.3998]	0.6415 (0.8792)	[-0.8437, 2.0133]
d = 0.8	0.4381 (2.2005)	[-3.8990, 4.6216]	0.8767 (1.1105)	[-0.8703, 2.5578]	3.5775 (1.5692)	[1.1035, 5.8811]

Notes: (i) d indicates the treatment intensity; (ii) standard errors calculated using cross-fitted formula are shown in parentheses; (iii) 90%-CI using multiplier bootstrap shown in separate columns; (iv) for all post-treatment period $t = 1984, 1985, 1986$, the baseline pre-treatment year is $t = 1983$; (v) all the nuisance parameters are estimated nonparametrically using random forests.

4.5 Empirical Application 2: 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 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)).[34]

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 Y_i denote the educational level of individual i, which is our outcome variable;
- 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);

[34] 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.

- 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 $d_L > 0$;
- we include the following covariates X_i: gender, religion, land ownership (as a proxy for family wealth), community size, urban/rural residency;
- finally, for our sample, we consider all individuals who 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 4.4. *Duflo (2001) divides the sample into low-intensity and high-intensity groups based on the treatment intensity. Although we are not able to locate the exact criteria used in Duflo (2001), we found that using the* 40/60 *cutoff roughly matches the mean difference in treatment intensities between "low" and "high" groups in our setting to that in Duflo (2001).*

Remark 4.5. *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 4. 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 ATT with 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. Specifically, for this DML estimator, we estimate the nuisance parameters using deep neural networks, and we report the estimated ATT in the second row in Table 4. We note that the nonparametric estimator from Chang (2020) with covariates shows a much larger treatment effect and has statistical significance.

Table 4: 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 17 different treatment intensities ranging from 10-percentile to 90-percentile of the empirical distribution of the treatment intensities in the treatment group. Here are the implementation details:

- we consider a 5-fold cross-fitting; in particular, we first randomly shuffle the data[35] before splitting the sample.
- the density $f_D(d)$ and conditional expectation $E[\frac{T-\lambda}{\lambda(1-\lambda)} Y | D = d]$ are estimated using a Gaussian kernel with bandwidth $h = N^{-1/4}$;
- nuisance parameters $\mathcal{E}_{\lambda Y}(X)$ and $g(X)$ are estimated using either the Random Forest (RF) or deep neural networks (DNN)[36]: for RF, we set the number of trees to 100 and the max-depth to 50; DNNs are implemented using multi-layer perceptron (MLP) with ReLU activation and optimized using the popular *Adam* algorithm (Kingma and Ba (2017));
- estimation of $f_J(d|X)$: we use the cosine basis on $[0, 2\pi]$, which roughly corresponds to the support of the treatment; we consider $J = N^{1/4}$ so that the under-smoothing

[35] The data we initially had was sorted by region. Without reshuffling, the sample splitting would have resulted in observations with certain treatment intensities being contained in only one subsample.

[36] Both are readily available in the scikit-learn library in Python.

Table 5: Diff-in-Diff with Continuous Treatment

	(Random Forest) ATT($D=d$)	Bootstrap CI	(Neural Network) ATT($D=d$)	Bootstrap CI
$\alpha = 0.1$	1.0065	[0.7743, 1.2474]	1.0193	[0.7586, 1.2675]
	(0.2206)		(0.2237)	
$\alpha = 0.2$	0.4428	[0.1843, 0.7116]	0.4519	[0.1690, 0.7284]
	(0.2215)		(0.2242)	
$\alpha = 0.3$	0.058	[-0.3434, 0.4730]	0.0706	[-0.4182, 0.5326]
	(0.2967)		(0.3113)	
$\alpha = 0.4$	1.5853	[1.2388, 1.9886]	1.5417	[1.1663, 1.8898]
	(0.2872)		(0.2857)	
$\alpha = 0.5$	0.8010	[0.4473, 1.1696]	0.8496	[0.4120, 1.2599]
	(0.2890)		(0.2977)	
$\alpha = 0.6$	0.7720	[0.4244, 1.0869]	0.7700	[0.4499, 1.0935]
	(0.2803)		(0.2767)	
$\alpha = 0.7$	0.8861	[0.6224, 1.1342]	0.8837	[0.6226, 1.1412]
	(0.2632)		(0.2646)	
$\alpha = 0.8$	-0.3350	[-0.7421, 0.0700]	-0.4176	[-0.8074, -0.0273]
	(0.3834)		(0.3816)	
$\alpha = 0.9$	1.0055	[0.7257, 1.3195]	1.0152	[0.7227, 1.3018]
	(0.4281)		(0.4268)	

Notes: (i) α indicates the treatment intensity d being the corresponding percentile values, with standard errors calculated using cross-fitted formula in parentheses; (ii) 95%-CI using multiplier bootstrap; (iii) in column 2, all the nuisance parameters are estimated using the Random Forest (RF) methods; (iv) in column 3, all the nuisance parameters are estimated using the deep neural network of multi-layer perceptron (MLP) class with ReLU activation

assumption is more plausible; each conditional expectation in $f_J(d|X)$ is estimated using the ML methods mentioned above;

- standard errors are calculated using the cross-fitted estimator defined in (44); we also construct 95-percent bootstrap confidence intervals using the multiplier bootstrap procedure defined in (45): we use Gaussian multipliers $\{\xi_i\}_{i=1}^N$ with $E[\xi_i] = Var[\xi_i] = 1$ for $B = 1000$ repetitions.

We present a few selected results for our estimator in Table 5, with visualizations in Figure 6 (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 binary treatment results, our results suggest that the ATTs vary 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 (e.g., 40 and 50 percentile values) but small and even negative values at other intensities. One potential explanation is that, since the treatment is determined at the district level, the variations may reflect other district-specific characteristics. Indeed, as commented in Duflo

(2001), during the same period as the school constructions, there was also a large scale of water and sanitation programs being implemented, which can be a potential confounding factor. Unfortunately, as we mentioned previously, we are unable to include district-specific covariates as they do not have enough variations, in which case ML estimators can use such variables to predict the treatment status perfectly. 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.

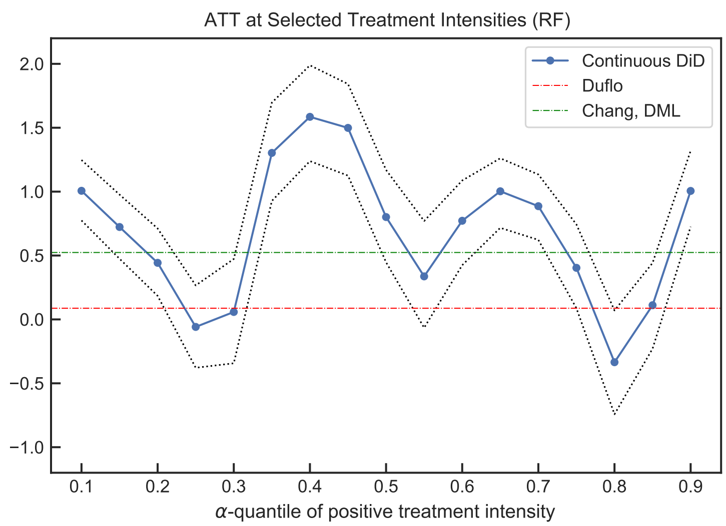

Figure 6: Diff-in-Diff with Continuous Treatment

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-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 a better understanding of the dependence relationships between the economic variables albeit the richer data sources and the increasing complexity of the economic models. Moreover, adding to the growing list of economics applications where conditional densities play a crucial role, we study the nonparametric difference-in-differences models with continuous treatments in detail. Such models have important implications in empirical research, and we hope our methods can provide new tools for researchers to analyze the effects of continuous treatment variables in the future.

Acknowledgements

The author would like to express his gratitude to Andres Santos, Denis Chetverikov, and Rosa Matzkin for their generous time and extremely helpful discussions that have led to substantial improvements to this paper. 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. Last but not least, the author would like to thank Kathleen McGarry, Amy Finkelstein, and NBER for generously providing the data used in the empirical application.

A Proofs

A.1 Proof of Proposition 3.1

For the first claim, note that $\mathbf{Y}$ is assumed to be a Polish space, and in particular, any compact subset of a Polish space is also Polish. Given that ν_Y is a Radon measure[37], by 7.14.13 in Bogachev (2007b), ν_Y on $\mathcal{B}_Y$ is therefore separable. Then by 4.7.63 in Bogachev (2007a), we conclude that $L^2(\nu_Y)$ is separable.[38]

To show the second claim, let $f_{Y|X} \in L^2(\nu_Y \otimes P_X)$ and let $\{\phi_j\}_{j=1}^{\infty}$ be an orthonormal basis on $L^2(\nu_Y)$. By Fubini's theorem,

$$\int f_{Y|X}^2(y|x) d\nu_Y dP_X < \infty \implies P_X(x \in \mathbf{X} : \int f_{Y|X}^2(y|x) d\nu_Y < \infty) = 1. \tag{27}$$

That is, $f_{Y|X}(\cdot|x) \in L^2(\nu_Y)$ for almost every $x \in \mathbf{X}$. Since $\{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis on $L^2(\nu_Y)$, by Parseval's identity (e.g. Theorem 5.27 in Folland (1999)), for $f_{Y|X}(\cdot|x) \in L^2(\nu_Y)$, there exists $\{h_j(x)\}_{j=1}^{\infty} \in \ell^2$ such that

$$\lim_{J\to\infty} \sum_{j=J+1}^{\infty} h_j^2(x) = \lim_{J\to\infty} \int (f_{Y|X}(y|x) - \sum_{j=1}^{J} h_j(x)\phi_j(y))^2 d\nu_Y = 0 \tag{28}$$

where the first equality holds by orthonormality. In particular, for every j,

$$h_j(x) := \int \phi_j(y) f_{Y|X}(y|x) d\nu_Y. \tag{29}$$

Since (29) holds for a.e. $x \in \mathbf{X}$, by the definition of conditional expectation (formally, see Proposition 10.4.18 in Bogachev (2007b)), we have

$$P(h_j(X) = E[\phi_j(Y)|X]) = 1. \tag{30}$$

Then the claim follows from (28) and (30).

To show the final claim, again we assume $f_{Y|X} \in L^2(\nu_Y \otimes P_X)$ and let $\{\phi_j\}_{j=1}^{\infty}$ be an orthonormal basis on $L^2(\nu_Y)$. First, for one direction, assume

$$\lim_{J\to\infty} \sum_{j=1}^{J} E[(E[\phi_j(Y|X)])^2] < \infty. \tag{31}$$

[37] We assume ν_Y to be Radon to rule out pathological cases involving counting measures.

[38] Separable measure allows us to construct a countable dense subset of simple functions, and since simple functions are dense in $L^2(\nu_Y)$, then the result follows.

Then by Fatou's Lemma,

$$E[\lim_{J\to\infty}\sum_{j=1}^{J}(E[\phi_j(Y|X)])^2] \le \lim_{J\to\infty}\sum_{j=1}^{J}E[(E[\phi_j(Y|X)])^2] < \infty \tag{32}$$

which also implies that

$$P(\lim_{J\to\infty}\sum_{j=1}^{J}(E[\phi_j(Y|X)])^2] < \infty) = 1. \tag{33}$$

By orthonormality,

$$\begin{aligned}
&\int (f_{Y|X}(y|X) - \sum_{j=1}^{J}E[\phi_j(Y)|X]\phi_j(y))^2 d\nu_Y \\
\le& 2\int f_{Y|X}^2(y|X)d\nu_Y + \lim_{J\to\infty} 2\sum_{j=1}^{J}(E[\phi_j(Y)|X])^2 \equiv M(X)
\end{aligned}$$

By $f_{Y|X} \in L^2(\nu_Y)$ and (32), $M(X) \in L^1(P_X)$. Therefore, by the second second claim in the theorem, applying dominated convergence theorem, we have

$$\lim_{J\to\infty} E[\int (f_{Y|X}(y|X) - \sum_{j=1}^{J}E[\phi_j(Y)|X]\phi_j(y))^2 d\nu_Y] = 0. \tag{34}$$

To show the other direction, assume (34) holds. Note that by orthonormality,

$$\sum_{j=1}^{J}E[(E[\phi_j(Y|X)])^2] = E[\sum_{j=1}^{J}\int (E[\phi_j(Y|X)]\phi_j(y))^2 d\nu_Y].$$

Then by $f_{Y|X} \in L^2(\nu_Y \otimes P_X)$ and (34),

$$\begin{aligned}
&\lim_{J\to\infty}\sum_{j=1}^{J}E[(E[\phi_j(Y|X)])^2] \\
=& \lim_{J\to\infty} E[\sum_{j=1}^{J}\int (E[\phi_j(Y|X)]\phi_j(y))^2 d\nu_Y] \\
\le& 2E[\int f_{Y|X}^2(y|X)d\nu_Y] + 2\lim_{J\to\infty} E[\int (f_{Y|X}(y|X) - \sum_{j=1}^{J}E[\phi_j(Y)|X]\phi_j(y))^2 d\nu_Y] < \infty.
\end{aligned}$$

This concludes the proof. ■

A.2 Proof of Lemma 3.1

To prove the claim of the lemma, consider $\hat{f}(Y, X)$ as a function of two random variables (Y, X), and let $f_{Y|X}$ denote the true conditional density. Then by definition, we have

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

In particular, note that the first two terms give us the results, and we only need to show that the last two terms add up to zero. To show this, we use the fact that $f_{Y|X}$ is the conditional density, and by the law of iterated expectations, we have

$$
E[\hat{f}(Y, X)] = E[E[\hat{f}(Y, X)|X]] = E[\int \hat{f}(y, X)f_{Y|X}(y|X)d\nu_Y(y)].
$$

■

A.3 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 $f_1, f_2 \in L^2(\nu_Y \otimes P_X)$, 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 $R_{n,K}(\hat{f}_j)$, we define $R^*_{n,K}$ as the minimized empirical risk, that is,

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

Then, the difference in the risk of our estimator and the risk at the true conditional density satisfies

$$\begin{aligned}&R(\bar{f}^{(n)}) - R(f_{Y|X}) \\ =&(1+a)(R^*_{n,K} - R_{n,K}(f_{Y|X})) + (R(\bar{f}^{(n)} - R(f_{Y|X})) - (1+a)(R^*_{n,K} - R_{n,K}(f_{Y|X})) \\ \leq&(1+a)(R_{n,K}(\hat{f}_j) - R_{n,K}(f_{Y|X})) + (R(\bar{f}^{(n)}) - R(f_{Y|X})) - (1+a)(R^*_{n,K} - R_{n,K}(f_{Y|X}))\end{aligned} \tag{35}$$

for all $a > 0$ and $1 \leq j \leq p$. The inequality holds since $R^*_{n,K}$ is the minimized risk using the dictionary and therefore $R^*_{n,K} \leq R_{n,K}(\hat{f}_j)$ for all $1 \leq j \leq p$.

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

$$\begin{aligned}&E[R_{n,K}(\hat{f}_j) - R_{n,K}(f_{Y|X})] \\ =&E[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_V}\sum_{i \in D_k^{(n_V)}} Q((Y_i,X_i), \hat{f}_j^{(n_T)}(D_k^{(n_T)})) - Q((Y_i,X_i), f_{Y|X})] \\ =&\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_V}\sum_{i \in D_k^{(n_V)}} E[Q((Y_i,X_i), \hat{f}_j^{(n_T)}(D_k^{(n_T)}))] - E[Q((Y_i,X_i), f_{Y|X})] \\ =&E_{D^{(n_T)}}[R(\hat{f}_j^{(n_T)}(D^{(n_T)}))] - R(f_{Y|X})\end{aligned} \tag{36}$$

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

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

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

$$
\begin{aligned}
& E[(R(\bar{f}^{(n)}) - R(f_{Y|X})) - (1+a)(R_{n,K}^* - R_{n,K}(f_{Y|X}))] \\
\leq & E[\frac{1}{K}\sum_{k=1}^{K} E_P[Q((Y,X), \hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)}))] - E_P[Q((Y,X), f_{Y|X})] \\
& - (1+a)(\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_V}\sum_{i\in D_k^{(n_V)}} Q((Y_i,X_i), \hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)})) - Q((Y_i,X_i), f_{Y|X})))] \\
= & \frac{1}{K}\sum_{k=1}^{K} E[E_P[(Q((Y,X), \hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)})) - Q((Y,X), f_{Y|X}))] \\
& - \frac{1+a}{n_V}\sum_{i\in D_k^{(n_V)}} Q((Y_i,X_i), \hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)})) - Q((Y_i,X_i), f_{Y|X})))] \\
\leq & E[\max_{1\leq j\leq p}(P - (1+a)P_{n_V})(Q((Y,X), \hat{f}_j^{(n_T)}(D^{(n_T)})) - Q((Y,X), f_{Y|X})))].
\end{aligned}
\tag{37}
$$

In the above derivation, the first inequality holds by convexity and definition of $R, R_{n,K}$, and the second equality holds by the i.i.d. sampling assumption and that the validating sets $D_k^{(n_V)}$ are of equal size n_V and are disjoint from each other. In the last line, we use P to denote the expectation E_P and P_{n_V} to denote the empirical average using validating set $D^{(n_V)}$, and the inequality holds since $\hat{j}^* \in \{1, \cdots, p\}$.

Then combining (35), (36), and (37), we have

$$
\begin{aligned}
&E[\|\bar{f}^{(n)} - f_{Y|X}\|_H^2] \\
=&E[R(\bar{f}^{(n)}) - R(f_{Y|X})] \\
\leq &\min_{1\leq j\leq p}(1+a)E_{D^{(n_T)}}[R(\hat{f}_j^{(n_T)}(D^{(n_T)}))] - R(f_{Y|X}) \\
&+ E[\max_{1\leq j\leq p}(P-(1+a)P_{n_V})(Q((Y,X),\hat{f}_j^{(n_T)}(D^{(n_T)})) - Q((Y,X),f_{Y|X})))] \\
\leq &\min_{1\leq j\leq p}(1+a)E[\|\hat{f}_j^{(n_T)} - f_{Y|X}\|_H^2] \\
&+ E[\max_{1\leq j\leq p}(P-(1+a)P_{n_V})(Q((Y,X),\hat{f}_j^{(n_T)}(D^{(n_T)})) - Q((Y,X),f_{Y|X})))]
\end{aligned}
\tag{38}
$$

where the first equality and last inequality hold by definition that $R(f) = \|f - f_{Y|X}\|_H^2 - \|f_{Y|X}\|_H^2$ and $R(f_{Y|X}) = -\|f_{Y|X}\|_H^2$ for $f = \bar{f}^{(n)}$ and $f = \hat{f}_j^{(n_T)}$, and the second inequality holds by boundedness assumption and monotonicity of expectations. In the next section, we bound the the maximum of the shifted empirical process term in (38) 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 $\mathbf{Z}$ and $\{Z_i\}_{i=1}^n \sim Z$ a sequence of i.i.d. random variables with $Z \in \mathbf{Z}$ distributed according to a probability measure P_Z on Borel σ-algebra $\mathcal{B}_Z$. Moreover, we assume that, for all $1 \leq j \leq p$, (i) $E[G_j(Z)] \geq 0$; (ii) $\|G_j\|_\infty \leq \tilde{M}$ for some constant $\tilde{M}$; (iii) $(E[G_j^2(Z)])^{1/2} \leq C(E[G_j(Z)])^{1/2}$ for some constant $C > 0$.

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 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,

$$
P\left[E[G_j(Z)] - \frac{1}{n}\sum_{i=1}^n G_j(Z_i) \geq \frac{x + aE[G_j(Z)]}{1+a}\right]
$$

$$
\begin{aligned}
&\le \exp\left(-c_1 n \frac{(\frac{x+aE[G_j(Z)]}{1+a})^2}{E[G_j^2(Z)] + \tilde{M}\frac{x+aE[G_j(Z)]}{1+a}}\right) \\
&\le \exp\left(-c_2 n \left[\frac{(\frac{x+aE[G_j(Z)]}{1+a})^2}{E[G_j^2(Z)]} \wedge \frac{\frac{x+aE[G_j(Z)]}{1+a}}{\tilde{M}}\right]\right) \\
&\le \exp\left(-c_3 n \left[\frac{(x+aE[G_j(Z)])^2}{E[G_j^2(Z)]} \wedge \frac{x+aE[G_j(Z)]}{\tilde{M}}\right]\right) \\
&\le \exp\left(-c_4 n \left[\left(\frac{x+aE[G_j(Z)]}{(E[G_j(Z)])^{1/2}}\right)^2 \wedge \frac{x+aE[G_j(Z)]}{\tilde{M}}\right]\right)
\end{aligned}
$$

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 $(E[G_j^2(Z)])^{1/2} \le C(E[G_j(Z)])^{1/2}$.

Note that, for $x \ge E[G_j(Z)]$, we have

$$
\left(\frac{x+aE[G_j(Z)]}{(E[G_j(Z)])^{1/2}}\right)^2 \ge \left(\frac{x+aE[G_j(Z)]}{x^{\frac{1}{2}}}\right)^2 \ge x
$$

where the second inequality holds by the assumption that $E[G_j(Z)] \ge 0$, which implies that

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

On the other hand, for $0 < x < E[G_j(Z)]$,

$$
\left(\frac{x+aE[G_j(Z)]}{(E[G_j(Z)])^{1/2}}\right)^2 > \left(\frac{aE[G_j(Z)]}{(E[G_j(Z)])^{1/2}}\right)^2 = a^2 E[G_j(Z)] > a^2 x
$$

where the first inequality holds by $x > 0$, which again implies that

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

Therefore, we have for all $x > 0$,

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

which implies that for some constant C_1,

$$P\left[\max_{1\leq j\leq p} E[G_j(Z)] - (1+a)\frac{1}{n}\sum_{i=1}^{n} G_j(Z_i) \geq x\right] \leq p\exp(-C_1 n\frac{x}{\tilde{M}}). \tag{39}$$

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

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

where the first inequality holds since $E[X] = \int_{\mathbf{R}} 1_{x\geq 0}(x) - F_X(x)dx$; the second inequality holds since the probability is bounded above by one; the third inequality holds by 39; 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_1,

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

Note that throughout the derivation, we kept the constant $\tilde{M}$ explicit to accommodate the possibility of $\tilde{M}$ potentially growing with p.[39]

Step 4: Bound on Shifted Empirical Process

Now we apply this maximal inequality in our case. We need to first verify the assumptions

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

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 $(E[G_j^2])^{1/2} \leq C(E[G_j])^{1/2}$. Plug in the definition of the loss Q, we have

$$\begin{aligned}
&\left(E[G_j^2(Z)]\right)^{\frac{1}{2}} \\
&= \left(E\left[(Q(Z, \hat{f}_j) - Q(Z, f_{Y|X}))^2\right]\right)^{\frac{1}{2}} \\
&= \left(E\left[\left(\int \hat{f}_j(y|X)^2 d\nu(y) - 2\hat{f}_j(Y|X) - \int f_{Y|X}(y)^2 d\nu_Y(y) - 2f_{Y|X}\right)^2\right]\right)^{\frac{1}{2}} \\
&= \left(E\left[\left(\int (\hat{f}_j(y|X) - f_{Y|X}(y))(\hat{f}_j(y|X) + f_{Y|X}(y)) d\nu_Y(y) - 2(\hat{f}_j(Y|X) - f_{Y|X})\right)^2\right]\right)^{\frac{1}{2}} \\
&\leq \left(E\left[\left(\int (\hat{f}_j(y|X) - f_{Y|X}(y))(\hat{f}_j(y|X) + f_{Y|X}(y)) d\nu_Y(y)\right)^2\right]\right)^{\frac{1}{2}} + 2\left(E\left[(\hat{f}_j(Y|X) - f_{Y|X})^2\right]\right)^{\frac{1}{2}}
\end{aligned}$$

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

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

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

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

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

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

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

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

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 the law of iterated expectation and monotonicity of expectation to conclude. We want to emphasize that we can allow the bound on the dictionary $\{\hat{f}_j\}_{j=1}^p$ to 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.4 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 subscripts. 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

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

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.5 Proof of Theorem 3.3

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

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

$$\leq 2E[\int \sum_{j=1}^{J}(h_j(x)-\hat{h}_j(x))^2\phi_j^2(y)dP_X(x)] + 2\int (\sum_{j=J+1}^{\infty} h_j(x)\phi_j(y))^2 dP_X(x).$$

First, we focus on the second term. By condition (iv), we have

$$\int (\sum_{j=J+1}^{\infty} h_j(x)\phi_j(y))^2 dP_X(x) \lesssim \int (c(x)J^{-\gamma/2})^2 dP_X(x) = J^{-\gamma}\int c^2(x)dP_X(x) \lesssim J^{-\gamma}.$$

Note that this is the same upper bound on the bias as the MISE case.

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

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

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

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

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

$$E[\|\hat{f}_J(y|X) - f_{Y|X}(y|X)\|_{P_X}^2] \lesssim Jn^{-\delta} + J^{-\gamma}$$

which is the same bound as in the MISE case. Then use the cross-validated $\hat{J}^*$and Theorem 3.2, we conclude that

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

■

A.6 Proof of Theorem 4.1

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

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

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

Second,

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

where the first equality holds by the law total probability, second equality holds by law of iterated expectation, the third equality holds by that $Y_t = Y_t(D)$ and $Y_{t-1} = Y_{t-1}(0)$, 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

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

Next, for repeated cross-sections, we have

$$E[\frac{T - \lambda}{\lambda(1 - \lambda)}Y|D = d]$$

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

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 follows from the repeated outcomes case. ■

A.7 Proof of Lemma 4.1

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

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

where $\Delta Y = Y_t - Y_{t-1}$, $f_d^0 := f_D(d)$, $f_J^0(d|X) := f_{D|X}(d)$, $g_0(X) := P(D=0|X)$. We will add an adjustment term to the original score so that the new score satisfies the Neyman orthogonality w.r.t. the infinite-dimensional parameters. Let $m_J^d(D) :=$

$\sum_{j=1}^{J} \phi_j(D)\phi_j(d)\mathbf{1}\{D > 0\}$.

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

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

where ∂_1 and ∂_2 denotes the partial derivatives w.r.t. the positions of $f_J^0(d|X)$ and $g_0(X)$ respectively. Then, we have

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

Now it remains to show the new score $\psi_J := \varphi_J + c_J$ 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 the moment condition $E[\psi_J] = 0$. 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 $\eta_0 := (f_J^0(d|X), g_0(X), \mathcal{E}_{\Delta Y}^0(X))$ 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 w.r.t. each nuisance parameter separately. In particular, in the following derivations, by assumption in the lemma, we can use the dominated convergence theorem to interchange the derivatives and the expectations.

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

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

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

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

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

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

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

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

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

where the first line holds by definition with $\Delta\mathcal{E}(X) = \mathcal{E}_{\Delta Y}(X) - \mathcal{E}^0_{\Delta Y}(X)$, 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 w.r.t. 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.8 Proof of Theorem 4.2 (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 $|f - f_d^0| \leq (Nh)^{-1/2}$ 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

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

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

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

where $K_h(u) := h^{-1}(u/h)$ as defined in the assumption. Then, with the standard results for kernel regression (e.g., Härdle (1990)), we have

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

Part II: Orthogonal Scores

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

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

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 ($\dagger$) will be our main focus while the bias term ($\dagger\dagger$) will be taken care of by undersmoothing requirement in assumption 4.7.

By definition,

$$\sqrt{N}(\hat{\theta}_J - \theta_{0J}) = \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\psi_J(Z_i,\theta_{0,J},\hat{f}_{d,k},\hat{\eta}_k)] \tag{40}$$

where ψ_J is defined as in (20), 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:

$$\sqrt{N}(\hat{\theta}_J - \theta_{0J}) = \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\psi_J(Z,\theta_{0J},f^0_d,\hat{\eta}_k)] \tag{41}$$

$$+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_f\psi_J(Z,\theta_{0J},f^0_d,\hat{\eta}_k)](\hat{f}_{d,k} - f^0_d) \tag{42}$$

$$+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},\bar{f}_k,\hat{\eta}_k)](\hat{f}_{d,k}-f_d^0)^2 \tag{43}$$

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 (43) vanishes rapidly and does not contribute to the asymptotic variance. In the second step, we bound first-order term (42), which potentially contributes to the asymptotic variance. In step 3, we expand (41) 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 (43). By triangle inequality, we have

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

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

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

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

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

Then by conditional Markov's inequality, $(\hat{f}_{d,k}-f_d^0)^2\le O_p((Nh)^{-1})$, and assumption 4.5, we conclude that (43) $= o_p(1)$. We will show (a) at the end of this section.

Step 2: First-Order Terms

To bound (42), we use first the triangle inequality to obtain the decomposition

$$\begin{aligned}
&|E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \hat{\eta}_k)] - E[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)]| \\
\leq &\underbrace{|E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \hat{\eta}_k)] - E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)]|}_{J_{3k}} \\
+ &\underbrace{|E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)] - E[\partial_f \psi_J(Z, \theta_{0J}, f_d^0, \eta_0)]|}_{J_{4k}}.
\end{aligned}$$

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,

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

Note that the kernel density estimator satisfies $(\hat{f}_{d,k} - f_d^0) = O_p((Nh)^{-1/2})$, so we can rewrite (42) as

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

where the last equality holds by the definition that $\hat{f}_{d,k}-f_d^0 = (N-n)^{-1}\sum_{i\in I_k^c} K_h(D_i - d) - E[K_h(D-d)] + O(h^2)$ (where $N-n$ is the sample size of each auxiliary subsample used to estimate the nuisance parameters), the under-smoothing assumption that $\sqrt{N}h^2 \leq O(1)$, and the fact that $K^{-1}\sum_{k=1}^{K}(\hat{f}_{d,k} - E[K_h(D-d)]) = \frac{1}{N}\sum_{i=1}^{N}(K_h(D_i-d)-E[K_h(D-d)])$. In particular, the kernel expression in the last line is mean-zero and it will contribute to the asymptotic variance.

Step 3: "Neyman Term"

Now we consider (41), 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

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

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,

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

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

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

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

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

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

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,

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

$$
\begin{aligned}
&+ M_J^2 \|g - g_0\|_{P,2} \|\mathcal{E}_{\Delta Y} - \mathcal{E}^0_{\Delta Y}\|_{P,2} \\
&\lesssim M_J^2 \varepsilon_N^2
\end{aligned}
$$

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,

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

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

$$\begin{aligned}
&\partial_r^2 \psi_J(\eta_0 + r(\hat{\eta}_k - \eta_0)) \\
&\asymp C_1 \Delta_f \Delta_g + C_2(\Delta_g)^2 + C_3 M_J \Delta_{\mathcal{E}} \Delta_g + C_4 M_J (\Delta_g)^2 + C_5 \Delta_f \Delta_{\mathcal{E}} + C_6 \Delta_{\mathcal{E}} \Delta_g
\end{aligned}$$

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

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

Part III: Conclusion

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

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

$$
\begin{aligned}
&- \frac{1}{N}\sum_{i=1}^{N} S_f^0 (K_h(D_i - d) - E[K_h(D_i - d)]) \qquad ④ \\
&+ o_p((Nh)^{-1/2}) + o_p(N^{-1/2}M_J) \qquad ⑤ \\
&+ \theta_0 - \theta_{0J} \qquad ⑥
\end{aligned}
$$

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. Moreover, by assumption, $E[(m_J^d(D))^2] \asymp \tilde{M}_J^2$ and $E[|m_J^d(D)|^3] \asymp \tilde{M}_J^3$, and using boundedness assumptions on the nuisance parameters, we have $E[\psi_J^2] \asymp \tilde{M}_J^2$ and $E[\psi_J^3] \asymp \tilde{M}_J^3$, then the Lyapunov condition is also satisfied for ③. Then by CLT, together with assumptions 4.7 and 4.8, we have

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

with σ_N defined by

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

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 random sample $\{Z_i\}_{i=1}^N$, each with equal size $n = N/K$, and for each $k \in \{1, \cdots, K\}$, let $I_k^c := N \backslash I_k$ denote the complement.*

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

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

where $\hat{f}_{d,k}, \hat{\mathcal{E}}^d_{\lambda Y,k}, \hat{f}_{J,k}, \hat{g}_k, \hat{\mathcal{E}}_{\lambda Y,k}$ are the estimators of $f_d, E[\lambda Y|D = d], f_J(d|X), g(X)$ and $\mathcal{E}_{\lambda Y}(X)$ respectively using the rest of the sample I^c_k. In particular, $\hat{f}_{d,k}, \hat{\mathcal{E}}^d_{\lambda Y,k}$ are kernel estimators, $\hat{g}_k, \hat{\mathcal{E}}_{\lambda Y,k}$ are estimated using ML methods (e.g. deep neural networks), and each term in $\hat{f}_{J,k}$ is estimated estimated using ML for a large J.

- *Average through the K estimators to obtain the final estimator*

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

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

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

(i) *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;*

(ii) *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;*

(iii) *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 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) *$M_J/\sqrt{N} = o(1)$;*

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

(iv) *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, 4.5, 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

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

and ψ_J is defined as in (21) and $Y^{\lambda} := \frac{T-\lambda}{\lambda(1-\lambda)}Y$.

Similarly as before, we construct

$$\begin{aligned}\hat{\sigma}_N^2 :=& \frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\Big(\frac{1}{\hat{f}_{d,k}}(K_h(D-d)Y^{\hat{\lambda}_k} - E_{n^c,k}[K_h(D-d)Y^{\hat{\lambda}_k}]) \\ &- \psi_J(Z, \hat{\theta}_J, \hat{\lambda}_k, \hat{f}_{d,k}, \hat{\eta}_k) \\ &+ (\frac{\hat{\theta}_J}{\hat{f}_{d,k}} - \frac{\hat{\mathcal{E}}^d_{\lambda Y,k}}{\hat{f}_{d,k}})(K_h(D-d) - E_{n^c,k}[K_h(D-d)])\Big)^2]\end{aligned} \tag{44}$$

where

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

$Y^{\hat{\lambda}_k} := \frac{T-\hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)}Y$, 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 $E[\xi_i] = Var[\xi_i^2] = 1$. Then for each $b = 1, \cdots, B$, we draw such a sequence $\{\xi_i\}_{i=1}^N$ and construct

$$\begin{aligned}\widehat{ATT(d)}_b^* :=& \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k} \xi_i\Big(\hat{\mathcal{E}}^d_{\lambda Y,k} - \frac{T_i - \hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)} Y_i \mathbf{1}\{D_i = 0\} \frac{\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot \hat{g}_k(X_i)} \\ &- \frac{m_J^d(D_i)\hat{g}_k(X_i) - \mathbf{1}\{D_i=0\}\hat{f}_{J,k}(d|X_i)}{\hat{f}_{d,k}\cdot \hat{g}_k^2(X_i)} \hat{\mathcal{E}}_{\lambda Y,k}(X_i)\Big)\end{aligned} \tag{45}$$

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.1 holds. Let F_N, E_N be the set of $f > 0$ and $\mathcal{E}^d_{\lambda Y}$ such that $|f - f^0_d| \le (Nh)^{-1/2}$ and $|\mathcal{E}^d_{\lambda Y} - \mathcal{E}^d_{\lambda Y,0}| \le (Nh)^{-1/2}$. 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 $\hat{\mathcal{E}}^d_{\lambda Y} \in E_N$.

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

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

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

Part I: Kernel Regression Results

We first consider the (1), $\hat{\mathcal{E}}^d_{\lambda Y,k} := \hat{E}[\frac{T-\lambda}{\lambda(1-\lambda)} Y | D = d]$, 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 $Y^{\lambda} := \frac{T-\lambda}{\lambda(1-\lambda)}Y$. Then using the similar arguments as in the repeated outcomes case, we have

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

Part II: Orthogonal Scores

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

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

Then by definition,

$$
\sqrt{N}(\hat{\theta}_J - \theta_{0J}) = \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\psi_J(Z_i,\theta_{0,J},\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k)] \tag{46}
$$

where ψ_J is defined as in (21), 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,

$$
\begin{aligned}
\sqrt{N}(\hat{\theta}_J - \theta_{0J}) &= \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)] && (47) \\
&+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_{\lambda}\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)](\hat{\lambda}_k-\lambda_0) && (48) \\
&+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_{f}\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\hat{\eta}_k)](\hat{f}_{d,k}-f_d^0) && (49) \\
&+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_{\lambda}^2\psi_J(Z,\theta_{0J},\bar{\lambda}_k,\bar{f}_k,\hat{\eta}_k)](\hat{\lambda}_k-\lambda_0)^2 && (50)
\end{aligned}
$$

$$+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_f^2\psi_J(Z,\theta_{0J},\bar{\lambda}_k,\bar{f}_k,\hat{\eta}_k)](\hat{f}_{d,k}-f_d^0)^2 \tag{51}$$

$$+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}E_{n,k}[\partial_\lambda\partial_f\psi_J(Z,\theta_{0J},\bar{\lambda}_k,\bar{f}_k,\hat{\eta}_k)](\hat{f}_{d,k}-f_d^0)(\hat{\lambda}_k-\lambda_0) \tag{52}$$

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 (50)-(52) can be shown to be $o_p(1)$. The first-order term (49) 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 (48) on the asymptotic variance is negligible. The main term (47) can be analyzed in the same way as in the repeated outcomes case.

Step 1: Second Order Terms

First, we consider (50). By triangle inequality, we have

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

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

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

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

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

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 (50) $= o_p(1)$. We will show (a) at the end of this section.

Term (51) is bounded in the same way as the repeated outcomes case. By triangle inequality, we have

$$
\begin{aligned}
&|E_{n,k}[\partial_f^2 \psi_J(Z, \theta_{0J}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)] - E[\partial_f^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]| \\
\leq & \underbrace{|E_{n,k}[\partial_f^2 \psi_J(Z, \theta_{0J}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)] - E_{n,k}[\partial_f^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|}_{J_{3k}} \\
+ & \underbrace{|E_{n,k}[\partial_f^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)] - E[\partial_f^2 \psi_J(Z, \theta_{0J}, \lambda_0 f_d^0, \eta_0)]|}_{J_{4k}}.
\end{aligned}
$$

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 (51) $= o_p(1)$. We postpone the proof of (b) to the end of this section.

Finally, we can bound (52) using similar arguments as those for (50) and (51). 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 (52) $= o_p(1)$. 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 (48). By triangle inequality, we have

$$
\begin{aligned}
&|E_{n,k}[\partial_\lambda \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)] - E[\partial_\lambda \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]| \\
\leq &\underbrace{|E_{n,k}[\partial_\lambda \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)] - E_{n,k}[\partial_\lambda \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|}_{J_{5k}} \\
+ &\underbrace{|E_{n,k}[\partial_\lambda \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)] - E[\partial_\lambda \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|}_{J_{6k}}.
\end{aligned}
$$

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,

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

Note that $(\hat{\lambda}_k - \lambda_0) = O_p(N^{-1/2})$, we can rewrite (48) as

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

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, (48) vanishes when scaled by the square-root of the asymptotic variance.

Term (49) 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

$$
E[J_{7k}^2|I_k^c] = E[|E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)] - E_{n,k}[\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|^2 | I_k^c]
$$

$$
\begin{aligned}
&\leq \sup_{\eta \in T_N} E[|\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta) - \partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2 | I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta) - \partial_f \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim M_J^2 \varepsilon_N^2 \qquad\qquad \text{(e)}
\end{aligned}
$$

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,

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

Note that under the assumption, $(\hat{f}_{d,k} - f_d^0) = O_p((Nh)^{-1/2})$, we can rewrite (49) as

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

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 under-smoothing 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.

Step 3: "Neyman Term"

Now we consider (47), which can be shown using the same argument as the repeated outcomes case.

$$
\begin{aligned}
&\sqrt{N} \frac{1}{K} \sum_{k=1}^{K} E_{n,k}[\psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)] \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_J(Z_i, \theta_{0J}, \lambda_0, f_d^0, \eta_0) \\
&+ \sqrt{N} \frac{1}{K} \sum_{k=1}^{K} \underbrace{(E_{n,k}[\psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)] - E_{n,k}[\psi_J(Z_i, \theta_{0J}, \lambda_0, f_d^0, \eta_0)])}_{R_{nk}}
\end{aligned}
$$

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

$$R_{1k} := |G_{nk}[\psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)] - G_{nk}[\psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|$$

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,

$$R_{2k} := \sqrt{n}|E[\psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)|I_k^c] - E[\psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)]|.$$

For simplicity, let's suppress other arguments in ψ and denote $\psi_\eta^i := \psi_J(Z_i, \theta_{0J}, \lambda_0, f_d^0, \eta)$.

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

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

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

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

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, $E[\psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)] = 0$, so it suffices to bound $E[\psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \hat{\eta}_k)|I_k^c]$. Suppressing other arguments in the score, define

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

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

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

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)-(g) used in the previous steps. Note that replacing ΔY with $\frac{T-\lambda}{\lambda(1-\lambda)}Y$, we can show claims (b),(e),(f),(g) using the same arguments as (a),(b),(c),(d) respectively in the repeated outcomes case. Hence it remains to show (a), (c), and (d).

First, recall that

$$(a): \quad \sup_{\substack{\lambda\in P_N,\ f\in F_N \\ \eta\in T_N}} E[|\partial_\lambda^2\psi_J(Z, \theta_{0J}, \lambda, f, \eta) - \partial_\lambda^2\psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2] \lesssim M_J^2\varepsilon_N^2.$$

In particular,

$$\partial_\lambda^2\psi_J(\lambda, f_d, \eta) = \frac{\partial^2}{\partial\lambda^2}\left(\frac{T-\lambda}{\lambda(1-\lambda)}\right)\mathbf{1}\{D=0\}\frac{f_J(d|X)}{f_d \cdot g(X)}.$$

Then by Taylor's theorem,

$$\partial_\lambda^2\psi_J(\lambda, f_d, \eta) - \partial_\lambda^2\psi_J(\lambda_0, f_d^0, \eta_0) = \partial_\lambda^2\psi_J(\lambda_0, f_d^0, \eta) - \partial_\lambda^2\psi_J(\lambda_0, f_d^0, \eta_0) \quad (\star)$$

$$
\begin{aligned}
&+ \partial_\lambda^2 \partial_f \psi_J(\bar{\lambda}, \bar{f}_d, \eta)(f_d - f_d^0) \quad (\star\star) \\
&+ \partial_\lambda^3 \psi_J(\bar{\lambda}, \bar{f}_d, \eta)(\lambda - \lambda_0) \quad (\star\star\star)
\end{aligned}
$$

where $\bar{\lambda} \in (\lambda, \lambda_0)$ and $\bar{f} \in (f_d, f_d^0)$. For the first term $(\star)$,

$$
\begin{aligned}
&\partial_\lambda^2 \psi_J(\lambda_0, f_d^0, \eta) - \partial_\lambda^2 \psi_J(\lambda_0, f_d^0, \eta_0) \\
&= \frac{\partial^2}{\partial \lambda^2} \left(\frac{T - \lambda_0}{\lambda_0(1-\lambda_0)} \right) \frac{Y \mathbf{1}\{D=0\}}{f_d^0} \left(\frac{f_J(d|X)}{g(X)} - \frac{f_J^0(d|X)}{g_0(X)} \right) \\
&= \frac{\partial^2}{\partial \lambda^2} \left(\frac{T - \lambda_0}{\lambda_0(1-\lambda_0)} \right) \frac{Y \mathbf{1}\{D=0\}}{f_d^0} \left(\frac{f_J(d|X)(g_0(X) - g(X)) - (f_J^0(d|X) - f_J(d|X))g(X)}{g(X) g_0(X)} \right)
\end{aligned}
$$

Moreover, by assumption B.1, for $\epsilon_N = o(N^{-1/4})$, $(\star\star)$ and $(\star\star\star)$ are of smaller order. Therefore, by the definition of (P_N, F_N, T_N), boundedness of the nuisance parameters, and triangle inequality, we have

$$
\begin{aligned}
&\sup_{\substack{\lambda \in P_N,\ f \in F_N \\ \eta \in T_N}} E[|\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda, f, \eta) - \partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim \sup_{\eta \in T_N} E[|\partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta) - \partial_\lambda^2 \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim \sup_{\eta \in T_N} \|f_J(d|X) - f_J^0(d|X)\|_{P,2}^2 + \|g(X) - g_0(X)\|_{P,2}^2 \\
&\lesssim M_J^2 \epsilon_N^2
\end{aligned}
$$

which shows (a). Similarly, by Taylor's theorem,

$$
\begin{aligned}
\partial_\lambda \partial_f \psi_J(\lambda, f_d, \eta) - \partial_\lambda \partial_f \psi_J(\lambda_0, f_d^0, \eta_0) &= \partial_\lambda \partial_f \psi_J(\lambda_0, f_d^0, \eta) - \partial_\lambda \partial_f \psi_J(\lambda_0, f_d^0, \eta_0) \\
&+ \partial_\lambda \partial_f^2 \psi_J(\bar{\lambda}, \bar{f}_d, \eta)(f_d - f_d^0) \\
&+ \partial_\lambda^2 \partial_f \psi_J(\bar{\lambda}, \bar{f}_d, \eta)(\lambda - \lambda_0)
\end{aligned}
$$

and (c) can be shown using similar arguments as (a).

Finally, we show (d):

$$
\sup_{\eta \in T_N} E[|\partial_\lambda \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta) - \partial_\lambda \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2] \lesssim M_J^2 \varepsilon_N^2.
$$

Note that

$$
\partial_\lambda \psi_J(\lambda, f_d, \eta) = \frac{\partial}{\partial \lambda} \left(\frac{T - \lambda}{\lambda(1-\lambda)} \right) \mathbf{1}\{D=0\} \frac{f_J(d|X)}{f_d \cdot g(X)}.
$$

which implies

$$
\begin{aligned}
&\partial_\lambda \psi_J(\lambda_0, f_d^0, \eta) - \partial_\lambda \psi_J(\lambda_0, f_d^0, \eta_0) \\
&= \frac{\partial}{\partial \lambda}\left(\frac{T-\lambda_0}{\lambda_0(1-\lambda_0)}\right)\frac{Y\mathbf{1}\{D=0\}}{f_d^0}\left(\frac{f_J(d|X)}{g(X)} - \frac{f_J^0(d|X)}{g_0(X)}\right) \\
&= \frac{\partial}{\partial \lambda}\left(\frac{T-\lambda_0}{\lambda_0(1-\lambda_0)}\right)\frac{Y\mathbf{1}\{D=0\}}{f_d^0}\left(\frac{f_J(d|X)(g_0(X)-g(X)) - (f_J^0(d|X) - f_J(d|X))g(X)}{g(X)g_0(X)}\right).
\end{aligned}
$$

Therefore, by the definition of T_N, boundedness of the nuisance parameters, and triangle inequality, we have

$$
\begin{aligned}
&\sup_{\eta \in T_N} E[|\partial_\lambda \psi_J(Z, \theta_{0J}, \lambda, f, \eta) - \partial_\lambda \psi_J(Z, \theta_{0J}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim \sup_{\eta \in T_N} \|f_J(d|X) - f_J^0(d|X)\|_{P,2}^2 + \|g(X) - g_0(X)\|_{P,2}^2 \lesssim M_J^2 \epsilon_N^2.
\end{aligned}
$$

This completes the proof for the auxiliary results.

Part III: Conclusion

Combining the results from I and II, we have

$$
\begin{aligned}
&\widehat{ATT(d)} - ATT(d) \\
&= \frac{1}{N}\sum_{i=1}^N \frac{K_h(D_i - d)Y_i^\lambda - E[(K_h(D-d)Y^\lambda]}{f_d^0} \qquad ① \\
&- \frac{E[Y^\lambda|D=d]}{f_d^0}\frac{1}{N}\sum_{i=1}^N (K_h(D_i - d) - E[K_h(D-d)]) \qquad ② \\
&- \frac{1}{N}\sum_{i=1}^N \psi_J(Z_i, \theta_{0J}, \lambda_0, f_d^0, \eta_0) \qquad ③ \\
&- \frac{1}{N}\sum_{i=1}^N S_f^0(K_h(D_i - d) - E[K_h(D_i - d)]) \qquad ④ \\
&+ o_p((Nh)^{-1/2}) + o_p(N^{-1/2}M_J) \qquad ⑤ \\
&+ \theta_0 - \theta_{0J} \qquad ⑥
\end{aligned}
$$

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. Moreover, by assumption, $E[(m_J^d(D))^2] \asymp \tilde{M}_J^2$ and $E[|m_J^d(D)|^3] \asymp \tilde{M}_J^3$, and using boundedness assumptions on the nuisance parameters, we have $E[\psi_J^2] \asymp \tilde{M}_J^2$ and $E[\psi_J^3] \asymp \tilde{M}_J^3$, then the Lyapunov condition is also satisfied for ③. Then by CLT, together with assumptions B.1 and B.2, we have

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

with σ_N defined by

$$\begin{aligned}\sigma_N^2 := E[\big(&\frac{1}{f_d^0}(K_h(D-d)Y^\lambda - E[K_h(D-d)Y^\lambda]) \\ &- \psi_J + (\frac{\theta_J}{f_d^0} - \frac{\mathcal{E}_{\lambda Y}^d}{f_d^0})(K_h(D-d) - E[K_h(D-d)])\big)^2]\end{aligned}$$

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

References

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

Acemoglu, D. and Finkelstein, A. (2008). Input and technology choices in regulated industries: evidence from the health care sector. *Journal of Political Economy* **116(5)**, 837–880.

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.

Ananat, E., Glasner, B., Hamilton, C., and Parolin, Z. (2022). Effects of the expanded Child Tax Credit on employment outcomes: evidence from real-world data from April to December 2021 (No. w29823). National Bureau of Economic Research.

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.

Bogachev, V. I. (2007a). *Measure theory (Vol. I)*. Berlin: Springer.

Bogachev, V. I. (2007b). *Measure theory (Vol. II)*. Berlin: Springer.

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.*

Candes, E. J. (2006). Modern statistical estimation via oracle inequalities. *Acta Numerica* **15**, 257–325.

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.

Cook, L. D., Jones, M. E., Logan, T. D., and Rosé, D. (2023). The evolution of access to public accommodations in the United States. *The Quarterly Journal of Economics* **138(1)**, 37–102.

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.

Folland, G. B. (1999). *Real analysis: modern techniques and their applications* (Vol. 40). John Wiley & Sons.

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.

Härdle, W. (1990). *Applied nonparametric regression* **(No.19)**. Cambridge university press.

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.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980v9.*

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.

Liebscher, E. (1990). Hermite series estimators for probability densities. *Metrika* **37(1)**, 321-343.

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

Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review* **81(1)**, 134–150.

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.

Rothfuss, J., Ferreira, F., Walther, S., and Ulrich, M. (2019). Conditional density estimation with neural networks: Best practices and benchmarks. *arXiv preprint arXiv:1903.00954.*

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.

Wahba, G. (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Annals of Statistics*, 15–29.

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.

Zeng, H. S., Danaher, B., and Smith, M. D. (2022). Internet governance through site shutdowns: the impact of shutting down two major commercial sex advertising sites. *Management Science* **68(11)**, 8234–8248.