UCLA
UCLA Electronic Theses and Dissertations

Title

Essays in Econometrics

Permalink

https://escholarship.org/uc/item/9163f0mg

Author

Zhang, Lucas

Publication Date

2024

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Essays in Econometrics

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Economics

by

Lucas Zhang

2024

© Copyright by
Lucas Zhang
2024

ABSTRACT OF THE DISSERTATION

Essays in Econometrics

by

Lucas Zhang
Doctor of Philosophy in Economics
University of California, Los Angeles, 2024
Professor Andres Santos, Co-Chair
Professor Denis Nikolaye Chetverikov, Co-Chair

This dissertation consists of three chapters that explore new methodologies and applications in econometrics.

In the first chapter, I propose a new class of functions defined by the so-called approximate sparsity condition. In general, functions in well-known classes can often be characterized by the rate of decay of their Fourier coefficients. The approximate sparsity condition generalizes this characterization by considering all sequences of such coefficients decreasing to zero at a certain rate while allowing for reordering. In particular, this generalization can potentially accommodate the modeling uncertainty of the unknown functions and aid estimation. For this new class of functions, I establish the metric entropy and minimax rate of convergence in terms of the estimation error. Moreover, I propose a data-driven density estimator based on a thresholding procedure and show this estimator can achieve the minimax rate up to a log term. A simulation study is also provided to demonstrate the performance of this estimator.

The second chapter focuses on the crucial role of conditional density in economic applications and introduces a data-driven nonparametric conditional density estimator suitable

for high-dimensional covariates. I first demonstrate that conditional density can be represented as a series, with each series term consisting of a known function multiplied by its conditional expectation. This structure is particularly beneficial in high-dimensional settings, where these conditional expectations can be flexibly estimated using various machine learning methods. Subsequently, I detail an algorithm that outlines the construction of my estimator based on this series formulation. Specifically, this procedure involves estimating a large number of conditional expectations and selecting the series cutoff through a data-driven procedure based on cross-validation. Lastly, I establish a general theory showing that this data-driven estimator is asymptotically optimal and can accommodate a wide range of machine learners under mild assumptions.

In the third chapter, I extend difference-in-differences to settings involving continuous treatments. Specifically, I identify the average treatment effect on the treated (ATT) at any level of continuous treatment intensity, using a conditional parallel trends assumption. In this framework, estimating the ATTs requires first estimating infinite-dimensional nuisance parameters, such as the conditional density of the continuous treatment, which can introduce significant biases. To address this challenge, I propose estimators for the causal parameters under the double/debiased machine learning framework. I demonstrate that these estimators are asymptotically normal and provide consistent variance estimators. To illustrate the effectiveness of my methods, I reexamine the study by Acemoglu and Finkelstein (2008), which assessed the effects of the 1983 Medicare Prospective Payment System (PPS) reform. By reinterpreting their research design using a difference-in-differences approach with continuous treatment, I nonparametrically estimate the treatment effects of the 1983 PPS reform, thereby providing a more detailed understanding of its impact.

The dissertation of Lucas Zhang is approved.

Rodrigo Ribeiro Antunes Pinto

Rosa Liliana Matzkin

Denis Nikolaye Chetverikov, Committee Co-Chair

Andres Santos, Committee Co-Chair

University of California, Los Angeles

2024

To my mom, in loving memory.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisors, Andres Santos, Denis Chetverikov, and Rosa Matzkin. Their wisdom, support, and guidance have been invaluable throughout the highs and lows of graduate school. I am profoundly grateful for the opportunity to learn from and collaborate with them.

Additionally, I extend my thanks to Rodrigo Pinto, Jinyong Hahn, Zhipeng Liao, Shuyang Sheng, Oscar Padilla, Mingli Chen, Kathleen McGarry, and all the participants at the econometrics proseminars at UCLA, for their insightful comments and enriching discussions during my academic journey.

Last but not least, my heartfelt thanks go to my friends and family, for their patience, support, and belief in me. Their constant encouragement has been a source of motivation and comfort throughout this journey.

VITA

2017	B.A. in Economics, Highest Distinction, UC Berkeley
2019	Ph.D. Candidate, UCLA
2019-2024	Teaching Assistant, Department of Economics, UCLA.
2020-2022	Instructor (Summer Sessions), Department of Economics, UCLA.
2021-2023	Teaching Assistant Consultant, Department of Economics, UCLA.

CHAPTER 1

Approximate Sparsity Class and Minimax Estimation

1.1 Introduction

First introduced by Čencov (1962); Kromal and Tarter (1968); Schwartz (1967); Watson (1969), density estimation using orthogonal series has since been extensively studied. Suppose we observe an i.i.d sample $\{X_i\}_{i=1}^n$ from the distribution of a random variable X on $[0,1]$ with probability density f_X. Let $\{\phi_j\}_{j=1}^\infty$ be an orthonormal basis of $L^2([0,1],\mu)$. If $f_X \in L^2([0,1],\mu)$, then it enjoys an expansion $f_X(\cdot) = \sum_{j=1}^\infty \theta_j \phi_j(\cdot)$, where $\theta_j = E[\phi_j(X)]$ is the j-th Fourier coefficients for all $j \geq 1$. A natural estimator then takes the form

$$\hat{f}_J(\cdot) := \sum_{j=1}^{J} \hat{\theta}_j \phi_j(\cdot), \quad \hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i). \tag{1.1}$$

When evaluating the performance of such estimator using MISE criteria, the cutoff value J plays the role of a tuning parameter, which, when chosen properly, balances the variance and bias and hence minimizes MISE. The optimal choices of the cutoff J have been discussed extensively in the literature, see, for example, Hall (1987); Hart (1985); Kromal and Tarter (1976); Watson (1969). A generalized version of such an estimator is based on thresholding

$$\hat{f}(\cdot) := \sum_{j=1}^{\infty} \omega_j \hat{\theta}_j \phi_j(\cdot), \tag{1.2}$$

where ω_j's are the so-called thresholding parameters and typically $\omega_j \in [0,1]$ plays a role of shrinking the estimated coefficients $\hat{\theta}_j$'s. Note that the previous estimator $\hat{f}_J(x)$ is a special

case of the thresholding estimator (1.2) when we use $\omega_j = \mathbf{1}\{j \leq J\}$. The thresholding estimator of the form (1.2) was first considered by Kromal and Tarter (1968), and many thresholding procedures have since been extensively studied, see for example, Buena et al. (2010); Chicken et al. (2005); Diggle and Hall (1986); Donoho et al. (1996); Efromovich (1986); Wahba (1981) and the references therein.

Various previous works have shown that when the thresholding parameters ω_j's are chosen properly, the orthogonal series estimators of the form (1.2) can achieve minimax rates of convergence over familiar function classes such as Sobolev ellipsoids and Besov spaces (commonly discussed in the context of wavelets thresholding), see for example, Buena et al. (2010); Chicken et al. (2005); Donoho et al. (1996, 1998); Efromovich (1986); Hall (1986); Härdle et al. (2012) and the references therein. We note that many function classes considered, such as the Sobolev ellipsoids, are characterized by the restrictions on the Fourier coefficients of the functions in those classes. Those restrictions on the Fourier coefficients in turn help researchers establish the statistical properties of estimators of the form (1.2).

The type of restriction on Fourier coefficients that we are interested in concerns how fast Fourier coefficients decay. It is motivated, for instance, by the discussion in Efromovich (2008); Hall (1986) that for a twice differentiable density f on $[0,1]$ with bounded second derivative, its Fourier expansion $f(\cdot) = \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot)$ given the cosine orthonormal basis $\{\phi_j\}_{j=1}^{\infty}$ has the property that the Fourier coefficients $\{\theta_j\}$ decay at rate j^{-2}. Similar results concerning the Hermite orthonormal basis were discussed in Schwartz (1967). In this paper, we generalize such restrictions. In particular, such a generalization allows for: (i) the non-increasing reordering (in absolute value) of the Fourier coefficients satisfies that the j-th largest Fourier coefficient decays at rate $|\theta_{(j)}| \leq Aj^{-k}$ for some constants A, k; (ii) the tail sum of the Fourier coefficients satisfies $\sum_{j=J+1}^{\infty} \theta_j^2 \leq CJ^{-2k+1}$ for some constant C. We call such class the "approximate sparsity class", motivated by the "approximate sparsity condition" from Belloni et al. (2018). This class is interesting for the following reasons. First, it generalizes various previously considered classes of functions that are characterized

by their Fourier coefficients, including but not limited to the Sobolev ellipsoids and Hölder classes (e.g. Katznelson (2004)). Second, in practice, researchers may be uncertain about the order (in terms of absolute magnitude) of the true coefficients, and our approximate sparsity class reflects such uncertainty while still maintaining the decaying properties of the re-ordered coefficients that are important for estimation purposes.

While the approximate sparsity class may seem complex, we show that such class is sandwiched in between two classes with simpler structures, and we will use sandwich arguments to formally establish the $L^2([0,1],\mu)$ ϵ-metric entropy of the approximate sparsity class and of its density subclass. To establish the upper bound on the entropy, we use an existing result from the full approximation set (Lorentz (1966)). On the other hand, we prove a lower bound using a volume-type argument inspired by Smolyak (1960). With these entropy bounds, we apply the results from Yang and Barron (1999) to establish the minimax rate of convergence (in terms of MISE) for both density estimation and nonparametric regression with Gaussian noise. Specifically, the minimax rate obtained is of order $n^{-(2k-1)/(2k)}$. As mentioned before, one can verify that the Sobolev ellipsoids are subsets of our approximate sparsity class, and as expected, this rate on approximate sparsity class is slower than the well-known minimax rate $n^{-2k/(2k+1)}$ for Sobolev ellipsoids. For a comprehensive review of the minimax estimation and the connections between metric entropy and minimax rates, see for example, Tsybakov (2008); Yang and Barron (1997, 1999) and the references therein.

With the obtained minimax rate in mind, we propose an adaptive density estimator based on a data-driven hard thresholding procedure. The main idea is as follows. We first pick a large cutoff J, potentially much larger than sample size n, and estimate the first J Fourier coefficients by the sample mean $\hat{\theta}_j = n^{-1}\sum_{i=1}^{n}\phi_j(X_i)$. However, including all the J terms in the series will inevitably lead to a large variance in estimation and hence a suboptimal rate. To overcome this issue, in the second step, we use a hard-thresholding procedure to select all the $\hat{\theta}_j$ above a certain data-driven threshold λ and penalize the rest to zero. Then

my estimator takes the following form:

$$\tilde{f}_J(\cdot) = \sum_{j \le J, |\hat{\theta}_j| \ge \lambda} \hat{\theta}_j \phi_j(\cdot). \tag{1.3}$$

In the final step, $\tilde{f}_J$ is projected onto the space of densities to obtain a bona-fide density.

Note that since $\{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis, the estimation errors will be characterized entirely by Fourier coefficients $\theta_j = E[\phi_j(X_i)]$ and estimated $\hat{\theta}_j = n^{-1}\sum_{i=1}^{n}\phi_j(X_i)$. This suggests that essentially we are dealing with the problem of estimating and selecting many approximate means. Following the ideas from Belloni et al. (2018), we can pick the threshold λ to be greater than the $(1-\alpha)$-quantile of $\max_{1\le j\le J}|\hat{\theta}_j - \theta_j|$, with $\alpha \downarrow 0$ as $n \to \infty$. Previous literature suggests that such λ can be approximated in a data-driven way using results from self-normalized moderate deviation theories or high-dimensional bootstrap, see, for example, Belloni et al. (2012, 2018) and the references therein. Although these constructions of λ can be used to show the non-asymptotic rate, unfortunately, they are not sufficient for establishing the rate in terms of MISE. Instead, we modify their constructions and propose an alternative data-driven λ, and we show that such λ has the desired property with the help of Talagrand's inequality (see e.g. Bousquet (2003)). We then show that if the true density f belongs to an approximate sparsity class, my estimator achieves the minimax rate up to a log factor. Moreover, the estimator itself does not depend on the assumptions of the parameters of the sparsity class and is therefore adaptive.

The remainder of the paper is organized as follows. In the next section, we introduce notations and formally define the approximate sparsity class. In Section 3 we establish the metric entropy and minimax rates for density estimation and nonparametric regression with Gaussian noise in such classes. In Section 4 we elaborate on the aforementioned adaptive density estimator and derive its rate of convergence. We then provide a specific example using the cosine basis for twice differentiable densities, in which we verify the assumptions and establish the rate of convergence by applying the results from the main theorem. We

conduct simulation studies in Section 5 to illustrate the performance of our estimator and conclude in Section 6. The proofs are deferred to the appendix.

1.2 Approximate Sparsity Class

Suppose $\Phi := \{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis of $L^2(\mathcal{X}, \mu)$ for $\mathcal{X} \subset \mathbf{R}$, and without loss of generality, let $\mathcal{X} = [0, 1]$. Then, for any $i \neq j$,

$$\int_0^1 \phi_j^2(x) d\mu(x) = 1, \quad \int_0^1 \phi_i(x)\phi_j(x) d\mu(x) = 0 \tag{1.4}$$

Moreover, for any $f \in L^2([0, 1], \mu)$, there is a representation

$$f(\cdot) = \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot) \quad \text{with} \quad \sum_{j=1}^{\infty} \theta_j^2 < \infty. \tag{1.5}$$

Here μ can be either the Lebesgue measure in the context of density estimation or known probability measures on $[0, 1]$ in the regression settings.

Restrictions on the Fourier coefficients $\{\theta_j\}_{j=1}^{\infty}$ lead to several familiar classes such as the *Sobolev ellipsoids* (e.g. Chapter 1.7.1 in Tsybakov (2008)) and *full approximation set* (e.g. Lorentz (1966)). Motivated by recent literature on high dimensional models (e.g. for a comprehensive review, see the handbook chapter by Belloni et al. (2018)), we introduce a new set of restrictions on the Fourier coefficients, and we call the resulting function class the *approximate sparsity class*:

Definition 1.2.1. *For given constants $A > 0, k > 1/2$ and $C > 0$, and for a given orthonor-*

mal basis $\Phi = \{\phi_j\}_{j=1}^{\infty}$*, the approximate sparsity class is defined as*

$$\Theta_k(\Phi, A, C) := \Bigg\{ f \in L^2([0,1], \mu) : f(\cdot) = \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot); \\ \textit{the non-increasing re-ordering } \{\theta_{(j)}\}_{j=1}^{\infty} \textit{ satisfies } |\theta_{(j)}| \leq A j^{-k}; \\ \forall J \geq 1, \textit{ the tail sum satisfies } \sum_{j=J+1}^{\infty} \theta_j^2 \leq C J^{-2k+1} \Bigg\}. \tag{1.6}$$

First, we note that the re-ordering in the definition refers to the re-ordering of the coefficients by the magnitude of their absolute values. Specifically, after re-ordering, $\theta_{(j)}$ will be the j-th largest element in $\{\theta_j\}_{j=1}^{\infty}$ by absolute magnitude. We also want to remark that the re-ordering requirement in the definition is equivalent to that $\forall J \geq 1$, the non-increasing re-ordering of $\{\theta_{(j)}\}_{j=1}^{J}$ satisfies $|\theta_{(j)}| \leq A j^{-k}$, which is a convenient characterization that will be used to establish various results in this paper. As we discussed in the introduction, it can be shown that the rate of decay of individual series coefficients is closely related to the smoothness of functions. The re-ordering condition relaxes such restriction by imposing less a priori knowledge on which series coefficients are important (as measured by magnitude).

Moreover, the restriction on the tail sum is a natural one. In particular, if the Fourier coefficients decay at Aj^{-k} without the re-ordering, the tail sum $\sum_{j=J+1}^{\infty} \theta_j^2$ can be shown to be bounded by CJ^{-2k+1} using integral bound. However, by allowing the re-ordering of the Fourier coefficients, the tail-sum restriction is a necessary one. In particular, the tail-sum restriction imposes structures on the re-ordering, which also helps us bound the estimation bias.

At first sight, the complexity of the approximate sparsity space seems difficult to characterize. We will introduce additional spaces that have simpler structures and "sandwich" the approximate sparsity space.

Definition 1.2.2. *For given constants* $A > 0, k > 1/2$*, and* $C > 0$*, and for a given or-*

thonormal basis $\Phi = \{\phi_j\}_{j=1}^{\infty}$*, define the following spaces*

$$\mathcal{E}_k(\Phi, A) := \left\{ f \in L^2([0,1], \mu) : \ f(\cdot) = \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot); \ |\theta_j| \leq A j^{-k} \ \forall j \geq 1 \right\} \tag{1.7}$$

$$\begin{aligned} \mathcal{A}_k(\Phi, A, C) := \Bigg\{ & f \in L^2([0,1], \mu) : f(\cdot) = \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot); \\ & |\theta_1| \leq A; \ \forall J \geq 1, \ \sum_{j=J+1}^{\infty} \theta_j^2 \leq C J^{-2k+1} \Bigg\}. \end{aligned} \tag{1.8}$$

The class $\mathcal{E}_k(\Phi, A)$ consists of all functions in $L^2([0,1], \mu)$ whose Fourier coefficients decay in an ordered manner in polynomial rates. For example, differentiable functions in $L^2([0,1], \mu)$ can be viewed as elements in such class. On the other hand, the set $\mathcal{A}_k(\Phi, A, C)$ is an example of the full-approximation set discussed in Lorentz (1966). A function $f \in \mathcal{A}_k(\Phi, A, C)$ can be approximated by the partial sums $\sum_{j=1}^{J} \theta_j \phi_j(\cdot)$, and the tail sum restriction in the definition of $\mathcal{A}_k(\Phi, A, C)$ can be understood as the restriction on the bias from such approximation. We show in the appendix (Lemma 1.8.1) that for appropriately chosen constant C, $\mathcal{E}_k(\Phi, A) \subseteq \Theta_k(\Phi, A, C) \subseteq \mathcal{A}_k(\Phi, A, C)$. In particular, note that if $|\theta_j| < A j^{-k}$, then

$$\sum_{j=J+1}^{\infty} \theta_j^2 \leq A^2 \int_J^{\infty} t^{-2k} dt = \frac{A^2}{2k-1} J^{-2k+1} \tag{1.9}$$

so we can simply take $C = A^2/(2k-1)$ in the definition of $\mathcal{A}_k(\Phi, A, C)$. The structures of $\mathcal{E}_k(\Phi, A)$ and $\mathcal{A}_k(\Phi, A, C)$ are much simpler and they will help us bound the metric entropy of the approximate sparsity class.

Throughout, we will use $M_2(\epsilon, \mathcal{F}) := \log N(\epsilon, \|\cdot\|_{L^2([0,1],\mu)}, \mathcal{F})$ to denote the Kolmogorov ϵ-entropy, where $N(\epsilon, \|\cdot\|_{L^2([0,1],\mu)}, \mathcal{F})$ is the cardinality of the largest ϵ-packing set of a set $\mathcal{F}$ under the $L^2([0,1], \mu)$ distance. Next, we formally introduce the definition of minimax rates, borrowing notation from Tsybakov (2008). Let $\{a_n\}$ and $\{b_n > 0\}$ be real sequences. We write $a_n \gtrsim b_n$ if $\liminf_{n\to\infty} a_n/b_n > 0$, and similarly, we write $a_n \lesssim b_n$ if $\limsup_{n\to\infty} a_n/b_n <$

∞. We use the notation "$\asymp$" as the asymptotic order symbol. That is, we write $a_n \asymp b_n$ if

$$0 < \liminf_{n\to\infty} \frac{a_n}{b_n} \leq \limsup_{n\to\infty} \frac{a_n}{b_n} < \infty. \tag{1.10}$$

Definition 1.2.3. *Given a set $\mathcal{F} \subseteq L^2([0,1],\mu)$ equipped with norm $\|\cdot\|_{L^2([0,1],\mu)}$, we say ψ_n is a minimax optimal rate of convergence on $(\mathcal{F}, \|\cdot\|_{L^2([0,1],\mu)})$ if*

$$\inf_{\hat{f}_n} \sup_{f\in\mathcal{F}} E_f \left[\|f - \hat{f}_n\|^2_{L^2([0,1],\mu)}\right] \asymp \psi_n^2 \tag{1.11}$$

where the infimum is taken over all possible estimators.

1.3 Entropy and Minimax Rate for Approximate Sparsity Class

As we have discussed in the previous section, for appropriately chosen constant C in the definitions of $\Theta_k(\Phi, A, C)$ and $\mathcal{A}_k(\Phi, A, C)$, we have $\mathcal{E}_k(\Phi, A) \subseteq \Theta_k(\Phi, A, C) \subseteq \mathcal{A}_k(\Phi, A, C)$. Therefore, understanding the metric entropy of $\mathcal{E}_k(\Phi, A)$ and $\mathcal{A}_k(\Phi, A, C)$ can help us control the metric entropy of the approximate sparsity class $\Theta_k(\Phi, A, C)$. In particular, the entropy on $\mathcal{E}_k(\Phi, A)$ and $\mathcal{A}_k(\Phi, A, C)$ will respectively give lower and upper bounds on the entropy of the sandwiched set $\Theta_k(\Phi, A, C)$, as shown in the next lemma.

Lemma 1.3.1. *The ϵ-entropy of $\mathcal{E}_k(\Phi, A))$ under the $L^2([0,1],\mu)$ distance satisfies*

$$M_2(\epsilon, \mathcal{E}_k(\Phi, A)) \gtrsim \epsilon^{-2/(2k-1)}. \tag{1.12}$$

Moreover, ϵ-entropy of $\mathcal{A}_k(\Phi, A, C))$ satisfies

$$M_2(\epsilon, \mathcal{A}_k(\Phi, A, C)) \asymp \epsilon^{-2/(2k-1)}. \tag{1.13}$$

The proof of this lemma is given in the appendix. As we remarked before, $\mathcal{A}_k(\Phi, A, C)$

is a special case of the full approximation set introduced in Lorentz (1966), and its entropy can be established using Theorem 3 in Lorentz (1966). On the other hand, to the best of our knowledge, the earliest reference of the class $\mathcal{E}_k(\Phi, A)$ can be traced back to Smolyak (1960) in which the trigonometric basis Φ is considered [1]. Inspired by Smolyak (1960), we adapt their arguments and extend similar results to $\mathcal{E}_k(\Phi, A)$ for more general orthonormal bases [2]. Next, we use Lemma 1.3.1 to establish the following theorem, which can be shown using a sandwich type of argument.

Theorem 1.3.1. *The ϵ-entropy of $\Theta_k(\Phi, A, C)$ under the $L^2([0,1], \mu)$ distance satisfies*

$$M_2(\epsilon, \Theta_k(\Phi, A, C)) \asymp \epsilon^{-2/(2k-1)}. \tag{1.14}$$

This theorem formally establishes bounds on the L^2 metric entropy of our approximate sparsity class $\Theta_k(\Phi, A, C))$with the help of $\mathcal{E}_k(\Phi, A)$ and $\mathcal{A}_k(\Phi, A, C)$. The following corollary establishes a similar entropy result on the density subset of $\Theta_k(\Phi, A, C))$.

Corollary 1.3.1. *Let $\tilde{\Theta}_k(\Phi, A, C) \subseteq \Theta_k(\Phi, A, C)$ be defined as the subset of all probability densities in $\Theta_k(\Phi, A, C)$. Moreover, assume that the basis $\Phi = \{\phi_j\}_{j=1}^{\infty}$ includes a constant term and $\Theta_k(\Phi, A, C)$ is uniformly bounded. Then*

$$M_2(\epsilon, \tilde{\Theta}_k(\Phi, A, C)) \asymp \epsilon^{-2/(2k-1)}. \tag{1.15}$$

We note that many familiar orthonormal bases on $L^2([0,1], \mu)$ contain a constant term. Moreover, the requirement that the functions in $\Theta_k(\Phi, A, C)$ are uniformly bounded, can

[1] The class $\mathcal{E}_k(\Phi, A)$ is sometimes referred to as the "hyperrectangles" in the literature, and minimax risks over these hyperrectangles has been studied in the context of Gaussian shift models, see, for example, Donoho et al. (1990) and the references therein.

[2] Prior to learning the existence of Smolyak (1960), we pursued an alternative route when establishing the entropy of $\mathcal{E}_k(\Phi, A)$. This alternative proof relies on an isometry between $\mathcal{E}_k(\Phi, A)$ and a generalized Hilbert cube (Kloeckner (2012)). Kloeckner (2012) establishes the bounds on the entropy of the Hilbert cubes with which one can infer lower bounds on the entropy of $\mathcal{E}_k(\Phi, A)$. While these lower bounds are sufficient for establishing the minimax rates, we opt to adopt the arguments used by Smolyak (1960) for the sake of clarity.

be satisfied, for example, when the orthonormal basis Φ is uniformly bounded and $k > 1$. Assuming uniform boundedness will allow us to find a set of the densities of the form $(f + M' + 1)/(\int f d\mu + M' + 1)$ for $f \in \mathcal{E}_k(\Phi, A)$ for some large constant M', which provides a lower bound for the entropy of $\tilde{\Theta}_k(\Phi, A', C')$ for some constants A' and C', which in turn is a lower bound for $\tilde{\Theta}_k(\Phi, A, C)$. Then applying the sandwich argument again gives us the results of the corollary. With the entropy results, we formally establish the minimax rates for nonparametric regression with Gaussian noise and density estimation on the approximate sparsity classes.

Theorem 1.3.2. *Let $\tilde{\Theta}_k(\Phi, A, C) \subseteq \Theta_k(\Phi, A, C)$ be the subset of all probability densities in $\Theta_k(\Phi, A, C)$ and assume $\Theta_k(\Phi, A, C)$ is uniformly bounded.*

(i) Suppose $\{X_i, Y_i\}_{i=1}^n$ is i.i.d with $X_i \sim P_X$ that admits a density. For nonparametric regression with Gaussian noise model $Y = f(X) + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ and $f \in \Theta_k(\Phi, A, C)$, the minimax rate of convergence satisfies

$$\inf_{\hat{f}_n} \sup_{f \in \Theta_k(\Phi, A, C)} E\left[\|\hat{f}_n - f\|^2_{L_2([0,1], P_X)}\right] \asymp n^{-\frac{2k-1}{2k}}. \tag{1.16}$$

(ii) Suppose $\{X_i\}_{i=1}^n$ is i.i.d with $X_i \sim X$ for $X \in [0, 1] \subseteq \mathbf{R}$ with density $f \in \tilde{\Theta}_k(\Phi, A, C)$, the minimax rate of convergence satisfies

$$\inf_{\hat{f}_n} \sup_{f \in \tilde{\Theta}_k(\Phi, A, C)} E_f\left[\|\hat{f}_n - f\|^2_{L_2([0,1], \mu)}\right] \asymp n^{-\frac{2k-1}{2k}}. \tag{1.17}$$

Theorem 1.3.2 can be seen as a direct consequence of Theorem 1.3.1 and Corollary 1.3.1 due to the tight connection between the entropy and minimax rates established in Yang and Barron (1999) and the references therein. In particular, the minimax rate of estimation on a particular class of functions $\mathcal{F}$ is closely related to the "critical separation" ϵ_n, which is determined via the identity $M_2(\epsilon_n, \mathcal{F}) = n\epsilon_n^2$. Then the results of the theorem should follow from our entropy results. We note that there are some caveats when applying Yang and

Barron (1999) directly on the density subset $\tilde{\Theta}_k(\Phi, A, C)$ since it is not convex nor bounded away from zero. However, we show in the proof that this subset is sandwiched in between two convex sets of densities with which we can establish the desired result.

1.4 Nearly Minimax Optimal Adaptive Density Estimator

1.4.1 Preliminaries

In this section, we propose a density estimator that is adaptive and achieves the minimax rate of convergence on $\tilde{\Theta}_k(\Phi, A, C)$ up to a log term. Suppose X is a random variable with probability density $f \in L^2([0,1], \mu)$ and let $\Phi = \{\phi_j\}_{j=1}^{\infty}$ be an orthonormal basis of $L^2([0,1], \mu)$ with μ being the Lebesgue measure. Then f has a unique representation:

$$f(\cdot) = \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot).$$

Moreover, f being a probability density and $\Phi = \{\phi_j\}_{j=1}^{\infty}$ being an orthonormal basis allow us to find expressions of θ_j's in terms of expectations:

$$\theta_j = \int_{[0,1]} \phi_j(x)(\sum_{j=1}^{\infty} \theta_j \phi_j(x)) d\mu(x) = \int_{[0,1]} \phi_j(x) f(x) d\mu(x) = E\left[\phi_j(X)\right]. \tag{1.18}$$

Given an i.i.d sample $\{X_i\}_{i=1}^n$ with $X_i \sim X$, the standard series estimator for the density f builds on identity (1.18) and takes the form

$$\hat{f}_J(x) = \sum_{j=1}^{J} \hat{\theta}_j \phi_j(x) \quad \text{where} \quad \hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i)$$

Note that the mean integrated squared error (MISE) of $\hat{f}_n$ equals

$$E\left[\|f-\hat{f}_J\|_{L^2([0,1],\mu)}\right]=\sum_{j=1}^{J}\frac{Var(\phi_j(X))}{n}+\sum_{j=J+1}^{\infty}\theta_j^2. \tag{1.19}$$

The choice of the cutoff J plays an essential role in the trade-off between variance and bias in (1.19), and yet in order to properly choose J in estimation, one often has to assume some prior knowledge about the smoothness of the unknown f.

We propose an alternative estimator that circumvents this issue. In particular, our estimator requires choosing a large cutoff J, and then uses LASSO to select the most "relevant" Fourier coefficients among $\{\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_J\}$. Note that many of the familiar bases, such as cosine and Legendre polynomial basis, contain a constant element. Therefore, we always assume the orthonormal basis of choice contains a constant term, which, without loss of generality, is assumed to be the first basis term ϕ_1. Then note that $\hat{\theta}_1=n^{-1}\sum_{i=1}^{n}\phi_1=\phi_1=E[\phi_1]=\theta_1$, that is, there's no estimation error for the first term. Therefore, we should always include $\hat{\theta}_1$ in our estimator in practice.

For a slight change of notation, let $\theta^J=\{\theta_1,\theta_2,\cdots,\theta_J\}\in\mathbf{R}^J$ denote the first J true but unknown series coefficients. Let $\hat{\theta}^J\in\mathbf{R}^J$ be an estimator of θ^J defined as follows

$$\hat{\theta}_j=\frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i);\quad \hat{\theta}^J:=(\hat{\theta}_1,\cdots,\hat{\theta}_J) \tag{1.20}$$

where each $\hat{\theta}_j$ is consistent by the weak law of large numbers under mild regularity conditions, and $\hat{\theta}^J$ can be shown to be consistent using Bernstein's inequality and maximal inequality under additional assumptions (e.g. if the orthonormal basis Φ is bounded or if the density is bounded). Given $\hat{\theta}^J$, the estimator we consider here is the so-called *hard-thresholding* estimator:

$$\tilde{\theta}_j=\omega_j\hat{\theta}_j\quad\text{where}\quad\omega_j=\mathbf{1}\{|\hat{\theta}_j|\geq\lambda\}\quad\text{for } 1\leq j\leq J \tag{1.21}$$

where ω_j's are the thresholding parameters depending on the penalty parameter λ. To be

consistent with the notation in (1.2), we let $\omega_j = 0$ for $j > J$. The hard thresholding estimator we consider here differs from the *soft-thresholding* estimator that has been considered in some previous literature, for example, Buena et al. (2010) and the references therein. Intuitively, for a properly chosen penalty parameter λ, we penalize "small" estimates of the series coefficients to 0 while keeping the rest unchanged. Let $T \subseteq \{1, \cdots, J\}$ denote the set of selected indices, that is,

$$T := \{j \in \{1, \cdots, J\} : |\hat{\theta}_j| \geq \lambda\}. \tag{1.22}$$

As a result, the new estimator we get is

$$\tilde{f}_J(x) := \sum_{j=1}^{J} \tilde{\theta}_j \phi_j(x) = \sum_{j=1}^{\infty} \omega_j \hat{\theta}_j \phi_j(x) = \sum_{j \in T} \hat{\theta}_j \phi_j(x). \tag{1.23}$$

As we will show later, the quality of this estimator depends on the penalty parameter and, to a lesser degree, the cutoff J.

Moreover, note that $\tilde{f}_J$ is not necessarily a probability density. We need to ensure that $\tilde{f}_J$ integrates to 1 and is nonnegative. The former is easily satisfied if μ is the Lebesgue measure on $[0, 1]$ and the first basis element $\phi_1 = 1$, which, by the definition of orthonormality, implies

$$\begin{aligned} \int_{[0,1]} \tilde{f}_J(x) d\mu(x) &= \int_{[0,1]} \phi_1^2 d\mu(x) + \sum_{j \in T \setminus \{1\}} \hat{\theta}_j \phi_j(x) d\mu(x) \\ &= \int_{[0,1]} \phi_1^2 d\mu(x) + \frac{1}{\phi_1} \int_{[0,1]} \phi_1 \sum_{j \in T \setminus \{1\}} \hat{\theta}_j \phi_j(x) d\mu(x) = 1. \end{aligned} \tag{1.24}$$

This is another reason why we should always include $\hat{\theta}_1 = \phi_1$ in our estimator.

On the other hand, some forms of post-processing are needed to ensure that the estimator is nonnegative. We will follow the *P-Algorithm* suggested by Gajek (1986), which is attractive in both its ease of implementation and the statistical properties of the resulting estimator. Here we review the P-Algorithm and illustrate with our estimator:

Definition 1.4.1. *The P-Algorithm is defined as follows:*

1. *Set* $f_0 = \tilde{f}_J$, *and* $k = 0$.
2. *Set* $f_{k+1} = \max\{0, f_k\}$, *and let* $C_{k+1} = \int_{[0,1]} f_{k+1}(x)d\mu(x)$. *Stop if* $C_{k+1} = 1$.
3. *Set* $f_{k+2} = f_{k+1} - (C_{k+1} - 1)$.
4. *Set* $k = k + 2$ *and go to step 2.*

Denote the resulting estimator from the algorithm as $\hat{f}^*$.

Gajek (1986) shows that the P-Algorithm converges both point-wise and in $\|\cdot\|_{L^2([0,1],\mu)}$ to a density $\hat{f}^* = \max\{0, \tilde{f}_J + c\}$, for some constant c. Moreover, $\hat{f}^*$ is the orthogonal projection of $\tilde{f}_J$ onto the space of densities, and $\hat{f}^*$ has at least the same rate of convergence (as measured in MISE) as $\tilde{f}_J$.

1.4.2 Main results

To estimate $\hat{f}^*$ and to establish its statistical properties, we need to consider a data-driven way of choosing the regularization parameter λ in (1.21). In particular, in order to control the penalization error in a uniform manner, we want to pick regularization parameter λ in a way such that

$$\lambda \quad \geq \quad (1-\alpha) - \text{quantile of } \|\hat{\theta}^J - \theta^J\|_\infty. \tag{1.25}$$

There are several ways of achieving this. For example, one can approximate such λ using the results from the self-normalized moderate deviation theory or high dimensional bootstrap literature, see, for example, Belloni et al. (2018) and the references therein. However, in order to establish the MISE, we need (1.25) to hold with $\alpha = \alpha_n$ going zero sufficiently fast as sample size n increases. To this end, we will modify the λ proposed by Belloni et al. (2018) based on the moderation deviation theories, and instead we will use Talagrand's inequality (see Bousquet (2003)) to establish the asymptotic result.

To facilitate the discussion, we borrow the following notations from Belloni et al. (2018): Let $\{X_i\}_{i=1}^n$ be an i.i.d random sample in $\mathbf{R}$ and let $\{\phi_j\}_{j=1}^\infty$ be an orthonormal basis of $L^2([0,1],\mu)$. With some abuse of notation, let Φ denote the CDF of the standard normal variable, and define

$$Z_{ij} := \phi_j(X_i) - E\left[\phi_j(X_i)\right] \tag{1.26}$$

and we let $\hat{Z}_{ij}$ be the sample analog of Z_{ij}

$$\hat{Z}_{ij} = \phi_j(X_i) - \frac{1}{n}\sum_{k=1}^{n} \phi_j(X_k). \tag{1.27}$$

We introduce the following regularity conditions.

Condition 1. *Let Z_{ij} be defined as in (1.26) and suppose that the orthonormal basis $\{\phi_j\}_{j=1}^\infty$ satisfies $\max_{1\le j\le J}\|\phi_j(\cdot)\|_\infty \le M_J$ for some M_J.*

(i) $n^{-1}\sum_{i=1}^n E\left[Z_{ij}^2\right] \ge 1$ *for all* $1 \le j \le J$;

(ii) $J = n^p$ *for some known* $p > 0$;

(iii) $M_J^2 \le n/\log(n)$;

(iv) $(JM_J/n)\alpha_n^{1/2} \le n^{-(2k-1)/(2k)}$ *for some positive sequence* $(\alpha_n) \downarrow 0$.

Condition 1 will be assumed to establish the main results in this section. We want to emphasize that all but (i) in Condition 1 can be verified, and we note that (i) is stated in the form that is convenient for the proof and the lower bound 1 in the statement can be replaced by any positive constant. In particular, (ii) is chosen by researchers, and (iii) can be checked for a given orthonormal basis. Note that when the chosen orthonormal basis is uniformly bounded, such as the cosine basis, we have $M_J^2 = M$ for some constant M, in which case (iii) is trivially satisfied, and we can potentially choose $J \gg n$. On the other hand, if we have a growing basis, the choices of J can be limited depending on how fast the basis grows.

For example, the Legendre basis grows with $M_J^2 = 2J + 1$, and we have to choose J such that $J = n^p$ for some $p < 1$ (p can be close to 1 for n large). The requirement (iv) can be viewed as the extra cost we pay when bounding the variance term in the MISE. In the next theorem, we propose a λ such that $\lambda \geq (1-\alpha_n)$-quantile of $\|\hat{\theta}^J - \theta^J\|_\infty$ with α_n goes to zero sufficiently fast so that the MISE can be established.

Theorem 1.4.1. *Let λ be defined as follows,*

$$\lambda := \sqrt{\frac{\log(J)}{n}}\Phi^{-1}\left(1 - \frac{1}{2\sqrt{2\pi}}\frac{1}{\sqrt{2\log(J)}}\frac{1}{J}\right)\max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{Z}_{ij}^2\right)^{\frac{1}{2}}. \tag{1.28}$$

Assume (i)-(iii) in Condition 1 are satisfied. Then there exists $n^* \in \mathbf{N}$ such that for all $n \geq n^*$,

$$\lambda \geq (1-\alpha_n) - \textit{quantile of } \|\hat{\theta}^J - \theta^J\|_\infty$$

with $\alpha_n = (Jn)^{-2} + 2n^{-3}$.

The proof of the theorem is given in the appendix and it can be easily adapted to allow other choices of α_n so that (iv) in Condition 1 can be satisfied. The expression of λ in the theorem bears some similarities to the ones given in Belloni et al. (2012) and in Belloni et al. (2018) (Theorem 2.4) using moderate deviation theories. Compared to their constructions, for $J = n^p$, our λ has an extra $\sqrt{\log(n)}$ multiplied (ignoring the constants). Roughly speaking, by Talagrand's inequality, this extra log term pushes our λ into the exponential tail and allows us to obtain α_n that is sufficiently fast for establishing MISE. We remark that the λ based on the moderate deviation theories can still be used to establish the non-asymptotic rate.

With this result, we can establish the second main result of this section. In particular, we are going to assume that the density of the random variable in question belongs to the approximate sparsity space $\tilde{\Theta}_k(\Phi, A, C)$, and we will show that the post-processed estimator $\hat{f}^*$ defined in 1.4.1 admits the minimax rate of convergence on $\tilde{\Theta}_k(\Phi, A, C)$ up to a log term.

Theorem 1.4.2. *Let $\{X_i\}_{i=1}^n \sim X$ be an i.i.d random sample with support $[0,1] \subset \mathbf{R}$. Assume that the true density f of X is in $\tilde{\Theta}_k(\Phi, A, C)$ uniformly bounded by some constant $\tilde{C}$ and that Condition 1 is satisfied. Let the regularization parameter λ be chosen as in (1.28) and let $\hat{f}^*$ be the estimator defined in 1.4.1. Then*

$$\sup_{f \in \tilde{\Theta}_k(\Phi, A, C)} E_f[\|f - \hat{f}^*\|_{L_2}^2] = O\left(\left(\frac{\log^2(n)}{n}\right)^{\frac{2k-1}{2k}}\right).$$

The proof of the theorem is given in the appendix. To bound the MISE, we first decompose the MISE into roughly the standard "variance" and "bias" components, and then we bound each separately with the help of Theorem 1.4.1. In the assumptions of the theorem, $f \in \tilde{\Theta}_k(\Phi, A, C)$ imposes restrictions on the Fourier coefficients of f, which will play a crucial role in establishing the rates. Moreover, we assume that $\tilde{\Theta}_k(\Phi, A, C)$ is uniformly bounded, which is also assumed when we establish the minimax rates. Note that for the cases when $k > 1$ and the orthonormal basis Φ is uniformly bounded, one can verify that $\tilde{\Theta}_k(\Phi, A, C)$ is uniformly bounded. In addition, Condition 1 is assumed so that we can utilize the results in Theorem 1.4.1, and as we commented before, the requirements in Condition 1 are easy to verify.

We also want to emphasize the adaptive nature of our estimator. In particular, once the researchers have decided on which orthonormal basis to use, the construction of the estimator $\hat{f}^*$ with data-driven λ does not depend on the assumptions on the approximate sparsity class (e.g. how fast the Fourier coefficients decay). Theorem 1.4.2 simply states that our estimator achieves the minimax rate on any approximate sparsity class. This is attractive in practice since the researchers do not have to assume the smoothness of the true data-generating density other than that it belongs to *some* approximate sparsity class.

1.4.3 Example Using Cosine Basis

In this section, we illustrate our previous results with the differentiable densities and the cosine basis. The standard cosine orthonormal basis $\{\phi_j\}_{j=1}^{\infty}$ on $L^2([0,1],\mu)$ is defined as follows:

$$\phi_1(x) = 1; \quad \phi_j(x) = \sqrt{2}\cos(\pi(j-1)x), \ j = 2, 3, \cdots \tag{1.29}$$

which is uniformly bounded. Suppose $f \in L^2([0,1],\mu)$ is twice differentiable. Then for $\{\phi_j\}_{j=1}^{\infty}$ being the cosine basis defined above, $f(\cdot) = \sum_{j=1}^{\infty}\theta_j\phi_j(\cdot)$ and there exists some constant c such that

$$|\theta_j| \leq cj^{-2}\int_0^1 |f^{(2)}(x)|\, dx, \quad j \geq 1$$

As noted in Efromovich (2008) (Section 2.2), unless additional (boundary) assumptions are made on the function f, the series coefficients can not decay faster than the rate j^{-2} (regardless of the smoothness of functions). This makes the set of twice differentiable functions with bounded second derivative a special example of the approximate sparsity class. In fact, this is an example of $\mathcal{E}_k(\Phi, A)$, which by itself is a special case of the approximate sparsity class without reordering. The next result follows directly from Theorem 1.4.2.

Corollary 1.4.1. *Suppose $\{X_i\}_{i=1}^n \sim X$ is an i.i.d sample of random variables with $X_i \in [0,1]$. Suppose the true density f of X is such that $f \in L^2([0,1],\mu)$ and is twice differentiable with bounded second derivative. Let $\{\phi_j\}_{j=1}^{\infty}$ be the cosine orthonormal basis defined as in (1.29). Assume requirement (i) in Condition 1 is satisfied and let $J = n^p$ for some $p > 0$. Let the regularization parameter λ be chosen as in (1.28) and let $\hat{f}^*$ be the estimator defined in 1.4.1. Then*

$$E\left[\|f - \hat{f}^*\|_{L_2}^2\right] = O\left(\left(\frac{\log^2(n)}{n}\right)^{\frac{3}{4}}\right).$$

Note that since the cosine basis is uniformly bounded, we can drop the assumption that the density is bounded and the requirements in Condition 1 can be verified. Moreover, in view of our results in Section 3, the rate in this corollary is minimax up to a log term. As

the approximate sparsity class and $\mathcal{E}_k(\Phi, A)$ class are new (and we commented previously that the familiar Sobolev ellipsoids are subsets of these classes), we have not yet considered the performances of other density estimators on such classes, and whether there are other adaptive rate-optimal estimators on these new classes remains an open question.

1.5 Simulations

In this section, we conduct simulation studies in which we compare the simulated MISE of our estimator introduced in Section 4 with an alternative estimator for which the series cutoff has to be properly specified. The true data-generating density is constructed to be in the approximate sparsity class.

1.5.1 Design density in approximate sparsity class

For the first simulation, we construct a density whose (cosine) Fourier coefficients satisfy approximate sparsity. Let

$$f_X(\cdot) = \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot) \tag{1.30}$$

where ϕ_j's are the cosine basis terms defined in (1.29). We specify the θ_j's in the following way:

- $\theta_1 = 1$, $\theta_j = Aj^{-2}$ for $j = 2, 3, 7, 8$;
- $\theta_5 = A10^{-2}$, $\theta_{11} = A4^{-2}$, $\theta_{13} = A6^{-2}$, $\theta_{14} = A5^{-2}$, $\theta_{15} = A9^{-2}$;
- $\theta_j = 0$ for $j = 4, 6, 9, 10, 12$ and for all $j \geq 16$.

As we've shown in section 4, since $\theta_1 = 1$, f_X integrates to 1. The constant $A = 2$ is chosen such that f_X in (1.30) is non-negative and hence a proper probability density.

1.5.2 Simulation procedures

To simulate MISE, we proceed in the following steps:

Step 1: We draw $B = 1000$ independent i.i.d. samples $\{X_i\}_{i=1}^N$ of size N. Here the N we consider will be $N = 5000, 10000, 15000, 20000$.

Step 2(a): For each sample $\{X_i\}_{i=1}^N$ from the true density f_X, we construct an estimator $\hat{f}^*$ described in section 4:

- We use cosine orthonormal basis defined in (1.29);
- The pre-processed estimator $\tilde{f}_J$ is constructed as in (1.23), with $J = 200$;
- The penalty parameter λ is chosen as in (1.28);
- We pass $\tilde{f}_J$ to the p-algorithm defined in (1.4.1):
  - The integral $C_{k+1} := \int_0^1 f_{k+1}(x)dx$ at each iteration is approximated using numerical integration, and we denote the approximated integral as $\hat{C}_{k+1}$;
  - The p-algorithm stops when $|\hat{C}_{k+1} - 1| < e^*$ for user specified e^*. This returns the positive estimator $\hat{f}^* = f_{k+1}$.

Step 2(b): For each sample $\{X_i\}_{i=1}^N$ from the true density f_X, we construct an alternative comparison estimator $\check{f}$ in the following way:

- We use cosine orthonormal basis defined in (1.29);
- We pick series cutoff J to be $N^{1/4}$ and construct the natural estimator

$$\hat{f}_J(\cdot) = \sum_{j=1}^{J} \hat{\theta}_j \phi_j(\cdot), \quad \text{where} \quad \hat{\theta}_j = \frac{1}{N} \sum_{j=1}^{n} \phi_j(X_i) \tag{1.31}$$

- We pass $\hat{f}_J$ to the p-algorithm defined in (1.4.1) in the same way as in Step 2(a), which returns a positive estimator $\check{f}$.

Step 3: For each $1 \leq b \leq B$ sample defined in *Step 1* and associated estimator from *Step 2(a)*, we estimate the integrated squared error (ISE) using numerical integration, and we use $\widehat{ISE}_b$ to denote the approximated ISE. We then calculate the estimated MISE

$$\widehat{MISE} := \frac{1}{B}\sum_{b=1}^{B} \widehat{ISE}_b \tag{1.32}$$

Repeat the same process for the estimator from Step 2(b).

Remark 1.5.1. Unlike our estimator in *Step 2(a)*, for the comparison estimator in *Step 2(b)* we need to properly specify the series cutoff J. In the simulation, for this comparison estimator, we choose $J = N^{1/4}$ for the following reason. When the true density is unknown, if the researcher is willing to make assumptions on the smoothness of the density, they can then determine the series cutoff based on such assumptions. For example, if the researcher assumes the true density is twice differentiable with bounded second derivative, as we discussed in Section 5, the Fourier coefficients θ_j decay at the rate j^{-2}, and one can show that a series cutoff $J = N^{1/4}$ minimizes the MISE under such assumption.

1.5.3 Simulation Results

The simulation design described above is coded directly using Python, with code available upon request. Since the design density is not a known density coded in Python packages, we use inverse transform sampling to generate random samples from our design density. We illustrate the performance of the inverse transform sampling in Figure (1.1), where the solid line is the design density and the normalized histogram below the solid line is constructed using a random sample of size 10000 from such inverse transform sampling.

We present the simulation results in Figure (1.2). Each dot represents the simulated MISE for a given sample size, where the red dots (labeled as "f_star") are the simulated MISEs for our estimator and the blue dots (labeled as "f_check") are for the comparison estimator. We make the following observations. First, as expected, for both our estimator

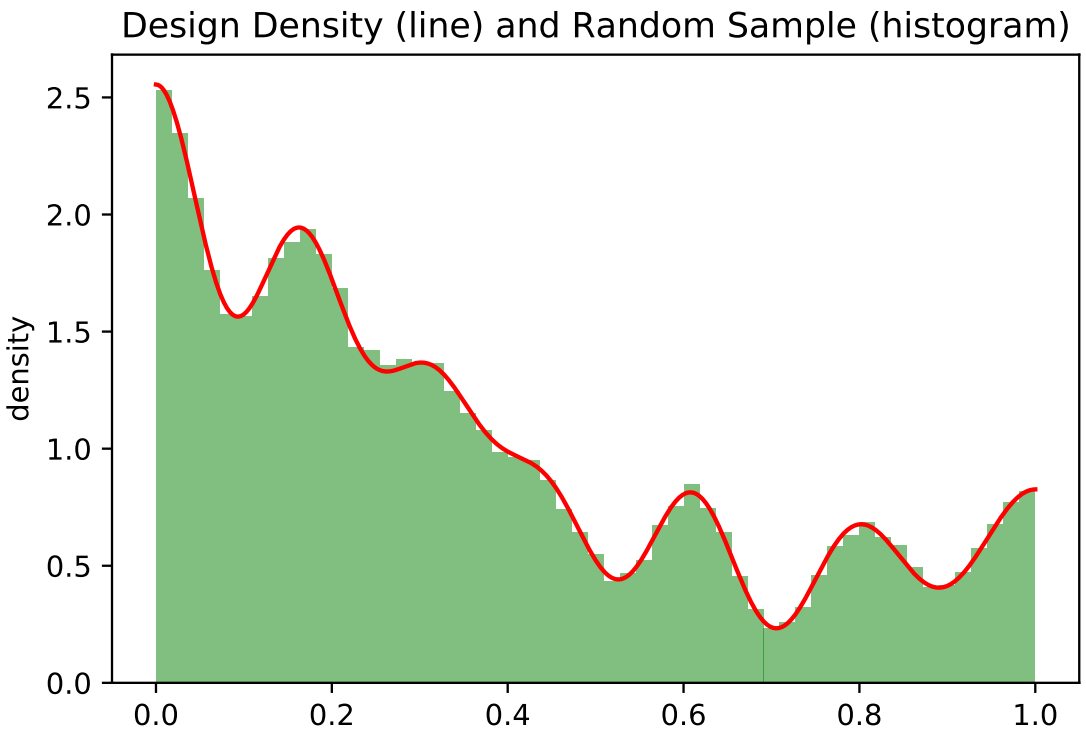

Figure 1.1: Design Density with Random Sample

and the comparison estimator, the simulated MISE decreases as the sample size increases. For the comparison estimator, this reflects the fact that such an estimator is using larger series cutoffs as sample size increases, which will eventually include all the nonzero terms of the series coefficients. On the other hand, our estimator requires choosing a large series cutoff ($J = 200$) to start with and then uses a data-driven hard-thresholding procedure to select the relevant terms. Second, for each of the sample sizes in our consideration, the simulated MISE of our estimator is smaller than that of the comparison estimator. This is likely due to the special feature of our design density, where the large Fourier coefficients show up in later series terms. Our estimator estimates the first $J = 200$ series terms and the data-driven hard-thresholding procedure is able to pick up the large series terms to reduce the bias. On the other hand, the comparison estimator has to specify the series cutoff $N^{1/4}$

and may fail to include the large Fourier coefficients that show up in later terms.

These simulation results demonstrate that our estimator performs well (as measured by the simulated MISE) for estimating densities in the approximate sparsity class. Moreover, the results showcase the adaptive nature of our estimator in comparison to an alternative estimator for which the researcher has to make potentially restrictive smoothness assumptions in order to determine the proper series cutoff. Even then, if the sample size is not large enough, such a comparison estimator may still miss large Fourier coefficients that show up in later series terms, which can result in larger estimation errors.

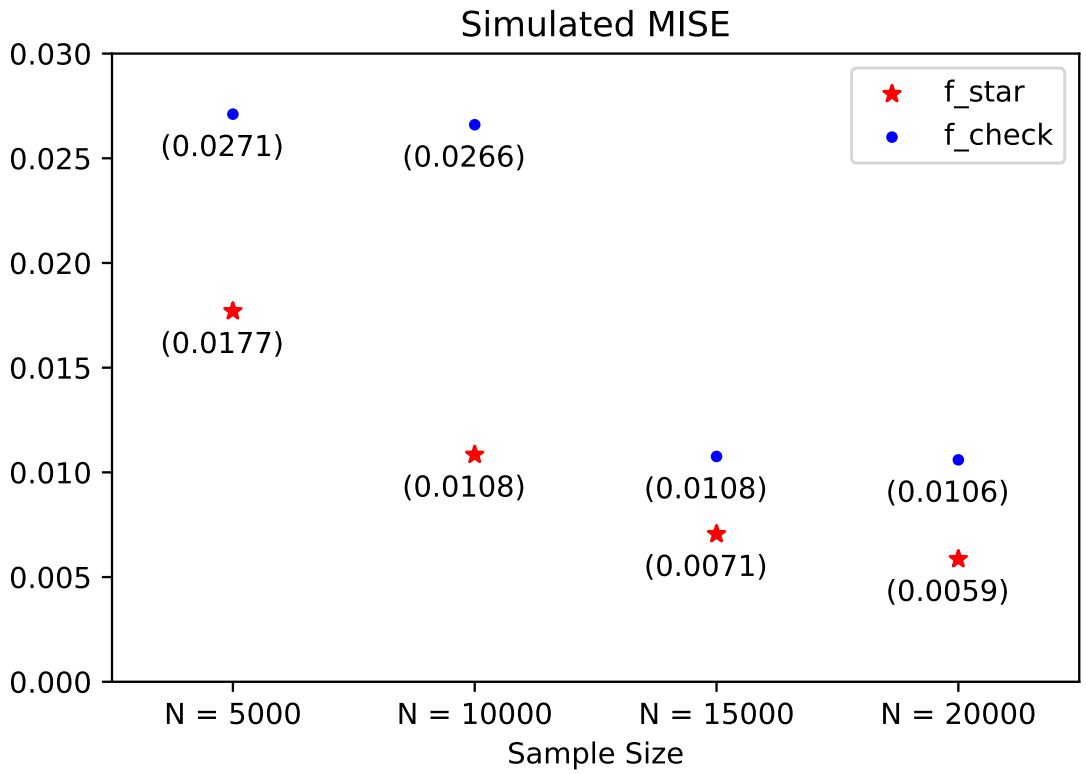

Figure 1.2: Simulated MISE With Various Sample Sizes

1.6 Conclusion

In this paper, we have studied a new class of functions, which we call the approximate sparsity class, that is characterized by a new set of restrictions on the Fourier coefficients of those functions for a given orthonormal basis. We have derived the upper and lower bounds on L^2 ϵ-metric entropy of the approximate sparsity class and we have established the minimax rates for nonparametric regression with Gaussian noise and for density estimation. We have shown that functions in such classes are natural candidates for the thresholding types of estimators and we proposed an adaptive density estimator that is nearly minimax optimal (up to a log term) over such class. For future research, we hope to study estimators for nonparametric regression for functions in the approximate sparsity class and we hope to generalize the approximate sparsity class to high-dimensional regression settings.

1.7 Proofs

1.7.1 Proof of Lemma 1.3.1

First, we establish the metric entropy of $\mathcal{A}_k(\Phi, A, C)$, which relies on Theorem 3 in Lorentz (1966). In particular, note that in the definition of $\mathcal{A}_k(\Phi, A, C)$, we have $\sum_{j=J+1}^{\infty} \theta_j^2 \leq CJ^{-2k+1}$. So following Lorentz (1966) notation, let $\delta_J^2 = CJ^{-2k+1}$. Moreover, simple calculation shows that $\delta_{2J} \leq c\delta_J$ for some constant $0 < c < 1$. Therefore, condition (13) in Lorentz (1966) holds.

Let $\epsilon > 0$ be given. Then by Theorem 3 of Lorentz (1966), the ϵ-metric entropy of $\mathcal{A}_k(\Phi, A, C)$ is of order

$$\min\{J : \delta_J^2 = CJ^{-2k+1} \leq \epsilon^2\}$$

solving which, we get $M_2(\epsilon, \mathcal{A}_k(\Phi, A, C)) \asymp \epsilon^{-2/(2k-1)}$.

Second, we establish the metric entropy of $\mathcal{E}_k(\Phi, A)$. Recall that the set $\mathcal{E}_k(\Phi, A)$ is

defined as

$$\mathcal{E}_k(\Phi, A) := \left\{ f \in L^2([0,1], \mu) : f(\cdot) = \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot);\ |\theta_j| \leq A j^{-k}\ \forall j \geq 1 \right\}.$$

which is isometric to the set

$$E_{k,A} := \left\{ (\theta_j)_{j=1}^{\infty} : |\theta_j| < A j^{-k} \right\}$$

i.e. for any two elements $f, g \in \mathcal{E}_k(\Phi, A)$, we can find two unique elements f^*, g^* in $E_{k,A}$ such that $\|f - g\|_{L^2([0,1],\mu)} = \|f^* - g^*\|_{\ell^2}$. This follows from the fact that Φ is an orthonormal basis. Define the following set

$$E_{k,A}^{Q} := \left\{ (\theta_j)_{j=1}^{\infty} : |\theta_j| < A j^{-k}\ \forall\ 1 \leq j \leq Q;\ \theta_j = 0\ \forall j \geq Q+1 \right\}.$$

which is a subset of $E_{k,A}$. It can be shown that $E_{k,A}^{Q}$ is isometric to the following set

$$H_{k,A}^{Q} := \left\{ (\theta_j)_{j=1}^{Q} : |\theta_j| < A j^{-k}\ \forall\ 1 \leq j \leq Q \right\} \subseteq \mathbf{R}^Q.$$

By the definition of isometry, the metric entropy of $\mathcal{E}_k(\Phi, A)$ is lower-bounded by the metric entropy of $H_{k,A}^{Q}$. We show that for a properly chosen Q, we can establish the desired lower bound.

Let V denote the volume of $H_{k,A}^{Q}$ and let v denote the volume of ϵ-ball in $\mathbf{R}^Q$. Then the ratio V/v provides a lower bound on the covering number of $H_{k,A}^{Q}$. By Sterling's formula, we have

$$v \asymp \frac{1}{Q\pi} \left(\frac{2\pi e}{Q} \right)^{\frac{Q}{2}} \epsilon^Q$$

Then we have

$$\frac{V}{v} \asymp \left(\prod_{j=1}^{Q} 2A j^{-k} \right) / \left(\frac{1}{Q\pi} \left(\frac{2\pi e}{Q} \right)^{\frac{Q}{2}} \epsilon^Q \right)$$

Taking log, we have

$$\begin{aligned}
\log\left(\frac{V}{v}\right) &\gtrsim \sum_{j=1}^{Q} \log(2Aj^{-k}) - \log\left(\frac{1}{Q\pi}\left(\frac{2\pi e}{Q}\right)^{\frac{Q}{2}}\epsilon^{Q}\right) \\
&\gtrsim -k\sum_{j=1}^{Q}\log(j) + Q\log(2A) - \log(Q^{-(Q+1)/2}\epsilon^{Q}) \\
&\gtrsim -k(Q\log(Q) - Q + \Theta(\log(Q))) + Q\log(2A) - \log(Q^{-(Q+1)/2}\epsilon^{Q}) \\
&\gtrsim (-k+\frac{1}{2})Q\log(Q) - Q\log(\epsilon) + (\log(2A)+k)Q
\end{aligned}$$

where the third line holds by Sterling's approximation. Then take Q to be such that $Q^{-k+1/2} = \epsilon^{-1}$, in which case $Q = \epsilon^{-2/(2k-1)}$, and we have

$$\log\left(\frac{V}{v}\right) \gtrsim \epsilon^{-2/(2k-1)}.$$

This gives a lower bound on the capacity (log of covering number) of $H^{Q}_{k,A}$. Using the convenient fact that the packing number is bounded below by the covering number (see for example, Lorentz (1966)), we conclude that

$$M_2(\epsilon, \mathcal{E}_k(\Phi, A)) \geq M_2(\epsilon, H^{Q}_{k,A}) \gtrsim \epsilon^{-2/(2k-1)}.$$

□

1.7.2 Proof of Theorem 1.3.1

By Lemma 1.8.1, for properly chosen constant C in the definition of $\mathcal{A}_k(\Phi, A, C)$, we have $\mathcal{E}_k(\Phi, A) \subseteq \Theta_k(\Phi, A, C) \subseteq \mathcal{A}_k(\Phi, A, C)$, and the results follow from Lemma 1.3.1. □

1.7.3 Proof of Corollary 1.3.1

We will use the sandwich argument again to establish this result. Note we are given that $\Theta_k(\Phi, A, C)$ are uniformly bounded by some constant M', which also holds for its subset $\mathcal{E}_k(\Phi, A)$. We will manipulate $\mathcal{E}_k(\Phi, A)$ to establish the lower bound, while the upper bound is still given by the full approximation set $\mathcal{A}_k(\Phi, A, C)$.

First, consider the following transformation

$$\begin{aligned}
\tilde{\mathcal{E}}_k(\Phi, A) &:= \left\{ \tilde{f} = \frac{f + M' + 1}{\int f d\mu + M' + 1} : f \in \mathcal{E}_k(\Phi, A) \right\} \\
&= \left\{ \tilde{f}(\cdot) = \frac{\theta_1 + M' + 1 + \sum_{j=2}^{\infty} \theta_j \phi_j(\cdot)}{\theta_1 + M' + 1} : f(\cdot) = \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot) \in \mathcal{E}_k(\Phi, A) \right\} \\
&= \left\{ \tilde{f}(\cdot) = 1 + \sum_{j=2}^{\infty} \frac{\theta_j}{\theta_1 + M' + 1} \phi_j(\cdot) : |\theta_j| < A j^{-k} \right\}.
\end{aligned}$$

This is a set of densities in $\mathcal{E}_k(\Phi, A')$, which is a subset $\tilde{\Theta}_k(\Phi, A', C')$ for some constant A'. Note that for M' large, $A' \leq A$ and $C' \leq C$, which implies that $\tilde{\Theta}_k(\Phi, A', C') \subseteq \tilde{\Theta}_k(\Phi, A, C)$. Therefore, it suffices to establish a lower bound for $\tilde{\Theta}_k(\Phi, A', C')$.

Second, consider a subset of $\tilde{\mathcal{E}}_k(\Phi, A)$, denote by

$$\mathcal{G}_1 := \left\{ g(\cdot) = 1 + \sum_{j=2}^{\infty} \theta_j \phi_j(\cdot) : |\theta_j| < \tilde{A} j^{-k} \right\}.$$

Note that $\mathcal{G}_1$ is a indeed a subset of $\tilde{\mathcal{E}}_k(\Phi, A)$ if, for example, $\tilde{A} = A/(A + M' + 1)$. Moreover, we define

$$\mathcal{G}_2 := \left\{ g(\cdot) = 1 + \sum_{j=1}^{\infty} \theta_j \phi_j(\cdot) : |\theta_j| < A^* j^{-k} \right\}$$

for some constant A^*. Note that we can change the index of θ_j by considering

$$|\theta_j| < A^* j^{-k} \implies |\theta_j| < A^* \left(\frac{j}{j+1} \right)^{-k} (j+1)^{-k} \leq 2^k A^* (j+1)^{-k}.$$

This implies that $\mathcal{G}_2$ is a subset of

$$\mathcal{G}_3 := \left\{ g(\cdot) = 1 + \sum_{j=2}^{\infty} \theta_j \tilde{\phi}_j(\cdot) : |\theta_j| < 2^k A^* j^{-k} \right\}$$

where $\tilde{\phi}_j = \phi_{j-1}$. Since $\{\phi_j\}$ is an orthonormal basis, it can be shown that the set $\mathcal{G}_1$ and $\mathcal{G}_3$ are isometric, which implies that they have the same order of entropy lower-bounded by the entropy of $\mathcal{G}_2$. Moreover, $\mathcal{G}_2$ and $\mathcal{E}_k(\Phi, A')$ are also isometric, so they also have the same order of $L^2([0,1],\mu)$ metric entropy.

Combining above results, we have found a subset of $\tilde{\Theta}_k(\Phi, A', C')$, $\mathcal{G}_1$, that has the same order of entropy as $\mathcal{E}_k(\Phi, A')$. Since $\tilde{\Theta}_k(\Phi, A', C') \subseteq \tilde{\Theta}_k(\Phi, A, C)$, by the results in Theorem 1.4.1, we conclude that

$$M_2(\epsilon, \tilde{\Theta}_k(\Phi, A, C)) \gtrsim \epsilon^{-2/(2k-1)}.$$

On the other hand, since $\tilde{\Theta}_k(\Phi, A, C) \subseteq \mathcal{A}_k(\Phi, A, C)$, the upper bound follows. □

1.7.4 Proof of Theorem 1.3.2

First, we show the claim on the nonparametric regression. Given the entropy bounds from Theorem 1.3.1, we can apply results from Yang and Barron (1997) (Theorem 9 and 10) and Yang and Barron (1999) (Theorem 6):

$$\inf_{\hat{f}_n} \sup_{f \in \Theta_k(\Phi, A, C)} E\left[\|f - \hat{f}_n\|_2^2 \right] \asymp n^{-\frac{2k-1}{2k}}.$$

Second, we establish the minimax rates on the density subset $\tilde{\Theta}_k(\Phi, A, C)$. Again, for notational simplicity, we use $\|\cdot\|_2$ to denote the $L^2([0,1],\mu)$ norm, where μ is the Lebesgue measure on $[0,1]$. Note that we can not apply Yang and Barron (1999)'s results directly on our set $\tilde{\Theta}_k(\Phi, A, C)$ since it may not be bounded away from zero and it is not convex (due to reordering restrictions on the Fourier coefficients). Nevertheless, since $\mathcal{E}_k(\Phi, A) \subseteq$

$\Theta_k(\Phi, A, C) \subseteq \mathcal{A}_k(\Phi, A, C)$, we have

$$\tilde{\mathcal{E}}_k(\Phi, A) \subseteq \tilde{\Theta}_k(\Phi, A, C) \subseteq \tilde{\mathcal{A}}_k(\Phi, A, C)$$

where (with some abuse of notation) $\tilde{\mathcal{E}}_k(\Phi, A)$, $\tilde{\Theta}_k(\Phi, A, C)$, and $\tilde{\mathcal{A}}_k(\Phi, A, C)$ are the density subsets of $\mathcal{E}_k(\Phi, A), \Theta_k(\Phi, A, C), \mathcal{A}_k(\Phi, A, C)$ respectively. Moreover, it can be verified from definition that $\tilde{\mathcal{E}}_k(\Phi, A)$ and $\tilde{\mathcal{A}}_k(\Phi, A, C)$ are both convex, so Yang and Barron (1999) applies to these two sets. In particular, the lower bound on the minimax rate of $\tilde{\Theta}_k(\Phi, A, C)$ is given by the a lower bound on $\tilde{\mathcal{E}}_k(\Phi, A)$ (the entropy of $\tilde{\mathcal{E}}_k(\Phi, A)$ is established in the proof of Corollary 1.3.1), and an upper bound is given by that of $\tilde{\mathcal{A}}_k(\Phi, A, C)$ (see Theorem 7 in Yang and Barron (1999)). Therefore, we have the following

$$\inf_{\hat{f}_n} \sup_{f \in \tilde{\Theta}_k(\Phi, A, C)} E\left[\|f - \hat{f}_n\|_2^2\right] \asymp n^{-\frac{2k-1}{2k}}.$$

□

1.7.5 Proof of Theorem 1.4.1

Recall that λ is defined as follows

$$\lambda := \sqrt{\frac{\log(J)}{n}} \Phi^{-1}\left(1 - \frac{1}{2\sqrt{2\pi}} \frac{1}{\sqrt{2\log(J)}} \frac{1}{J}\right) \max_{1 \le j \le J} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{Z}_{ij}^2\right)^{\frac{1}{2}}$$

where $\hat{Z}_{ij} := \phi_j(X_i) - n^{-1} \sum_{k=1}^{n} \phi_j(X_k)$. To establish the result, we need to bound λ from below. Using the property of normal CDF that $1 - \Phi(x) > 1/(2\pi)^{1/2}(x/(x^2+1)) \exp(-x^2/2)$ (see Lemma 1.8.2), we have

$$\Phi^{-1}\left(1 - \frac{1}{2\sqrt{2\pi}} \frac{1}{\sqrt{2\log(J)}} \frac{1}{J}\right) > \sqrt{2\log(J)}.$$

Now we bound $\max_{1\leq j\leq J}(n^{-1}\sum_{i=1}^n \hat{Z}_{ij}^2)^{1/2}$ from below. By triangle inequality, we have

$$\begin{aligned}\max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^n \hat{Z}_{ij}^2\right)^{\frac{1}{2}} &= \max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^n (\hat{Z}_{ij}-Z_{ij}+Z_{ij})^2\right)^{\frac{1}{2}} \\ &\geq \max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^n Z_{ij}^2\right)^{\frac{1}{2}} - \max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^n (\hat{Z}_{ij}-Z_{ij})^2\right)^{\frac{1}{2}}\end{aligned}$$

and we will bound each term separately.

First, following step 1 in the proof of Theorem 2.4 in Belloni et al. (2018), we have

$$P\left(\max_{1\leq j\leq J}\frac{1}{n}\sum_{i=1}^n Z_{ij}^2 \leq \frac{1}{2}\right) \leq P\left(\frac{1}{n}\sum_{i=1}^n Z_{ij}^2 \leq \frac{1}{2}\right)$$

where the inequality holds since $P(n^{-1}\sum_{i=1}^n Z_{ij}^2 \leq 1/2,\ \forall\ 1\leq j\leq J) \leq P(n^{-1}\sum_{i=1}^n Z_{ij}^2 \leq 1/2)$ for any $1\leq j\leq J$. Then for a generic $1\leq j\leq J$ and some constant $c'>0$,

$$\begin{aligned}P\left(\frac{1}{n}\sum_{i=1}^n Z_{ij}^2 \leq \frac{1}{2}\right) &\leq P\left(\frac{1}{n}\sum_{i=1}^n Z_{ij}^2 \leq \frac{1}{2n}\sum_{i=1}^n E\left[Z_{ij}^2\right]\right) \\ &= P\left(\sum_{i=1}^n Z_{ij}^2 - E\left[Z_{ij}^2\right] \leq -\frac{1}{2}\sum_{i=1}^n E\left[Z_{ij}^2\right]\right) \\ &\leq \exp\left(-\frac{(-\frac{1}{2}\sum_{i=1}^n E\left[Z_{ij}^2\right])^2}{2\sum_{i=1}^n E\left[Z_{ij}^4\right]}\right) \\ &\leq \exp\left(-\frac{n^2}{c'nM_J^2}\right)\end{aligned}$$

where the first inequality holds by the assumption that $E\left[Z_{ij}^2\right]\geq 1$, the second inequality holds by Bernstein inequality (see exercise 2.9 in Boucheron et al. (2013)), and the last inequality holds by that $E\left[Z_{ij}^4\right] = E\left[(\phi_j(X)-E\left[\phi_j(X)\right])^4\right] \leq (2M_J)^2 E\left[\phi_j^2(X)\right] \leq 4CM_J^2$. By assumption, $M_J^2 \leq n/\log(n)$ and $J=n^p$, which implies $n/(c'M_J^2)\geq 2\log(Jn)$. Then the

above result implies that

$$P\left(\frac{1}{n}\sum_{i=1}^{n} Z_{ij}^2 \leq \frac{1}{2}\right) \leq \exp(-2\log(Jn)) = (Jn)^{-2}$$

and in which case we have

$$P\left(\max_{1\leq j\leq J}\frac{1}{n}\sum_{i=1}^{n} Z_{ij}^2 \leq \frac{1}{2}\right) \leq (Jn)^{-2}.$$

Second, we bound the term $\max_{1\leq j\leq J}(n^{-1}\sum_{i=1}^{n}(\hat{Z}_{ij}-Z_{ij})^2)^{1/2}$. To simplify the notation, define

$$\mathbb{Z} := \max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{Z}_{ij}-Z_{ij})^2\right)^{\frac{1}{2}} = \max_{1\leq j\leq J}|\frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i) - E\left[\phi_j(X_i)\right]|.$$

By Bernstein inequality and maximal inequality (see Lemma 1.8.3 in the appendix),

$$E\left[\mathbb{Z}\right] \leq \frac{K(M_J\log(J)+\sqrt{n\log(J)})}{n}.$$

By Talagrand's inequality (see the version given by Bousquet (2003)),

$$P\left(\mathbb{Z} \geq E\left[\mathbb{Z}\right] + \sqrt{2\nu_n x} + \frac{Ux}{3}\right) \leq \exp(-x)$$

where U satisfies that $n^{-1}\|\phi_j(\cdot) - E\left[\phi_j(X_i)\right]\|_\infty \leq U < \infty$, and $\nu_n := 2UE\left[\mathbb{Z}\right] + n\sigma^2$ where $\sigma^2 := n^{-2}\max_{1\leq j\leq J}E\left[(\phi_j(X_i) - E\left[\phi_j(X_i)\right])^2\right]$. Note that we can take $U := 2M_J/n$ and $\sigma^2 \leq C/n^2$. Then we have

$$P\left(\mathbb{Z} \geq \frac{K(M_J\log(J)+\sqrt{n\log(J)})}{n} + \sqrt{2\left(\frac{2M_J}{n}\frac{K(M_J\log(J)+\sqrt{n\log(J)})}{n} + \frac{C}{n}\right)x + \frac{2M_J}{n}x}\right)$$

$$\leq P\left(\mathbb{Z} \geq E\left[\mathbb{Z}\right] + \sqrt{2\nu_n x} + \frac{Ux}{3}\right) \leq \exp(-x).$$

By Condition 1, $M_J^2 = o(n)$, which implies that there exists some $n_1 \in \mathbf{N}$ such that for all $n \geq n_1$ and for $x = 3\log(n)$ we have

$$P\left(\mathbb{Z} \geq \frac{1}{2\sqrt{2}}\right) \leq \exp(-3\log(n)) = n^{-3}.$$

Combining above, with probability at least $1 - (Jn)^{-2} - n^{-3}$, we have the following

$$\begin{aligned}
\max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{Z}_{ij}^2\right)^{\frac{1}{2}} &\geq \max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\right)^{\frac{1}{2}} - \max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{Z}_{ij}-Z_{ij})^2\right)^{\frac{1}{2}} \\
&= \max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\right)^{\frac{1}{2}} - \mathbb{Z} \\
&> \frac{1}{2\sqrt{2}}.
\end{aligned}$$

To see this, note that

$$\begin{aligned}
&P\left(\max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{Z}_{ij}^2\right)^{\frac{1}{2}} > \frac{1}{2\sqrt{2}}\right) \\
\geq &P\left(\max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\right)^{\frac{1}{2}} - \mathbb{Z} > \frac{1}{2\sqrt{2}}\right) \\
\geq &P\left(\max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\right)^{\frac{1}{2}} > \frac{1}{\sqrt{2}} \text{ and } \mathbb{Z} < \frac{1}{2\sqrt{2}}\right) \\
= &1 - P\left(\max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\right)^{\frac{1}{2}} \leq \frac{1}{\sqrt{2}} \text{ or } \mathbb{Z} \geq \frac{1}{2\sqrt{2}}\right) \\
\geq &1 - P\left(\max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}Z_{ij}^2\right)^{\frac{1}{2}} \leq \frac{1}{\sqrt{2}}\right) - P\left(\mathbb{Z} \geq \frac{1}{2\sqrt{2}}\right)
\end{aligned}$$

$$\geq 1 - (Jn)^{-2} - n^{-3}.$$

This implies that

$$\begin{aligned}\lambda &= \sqrt{\frac{\log(J)}{n}}\Phi^{-1}\left(1 - \frac{1}{2\sqrt{2\pi}}\frac{1}{\sqrt{2\log(J)}}\frac{1}{J}\right)\max_{1\leq j\leq J}(\frac{1}{n}\sum_{i=1}^{n}\hat{Z}_{ij}^2)^{1/2} \\ &\geq c''\sqrt{\frac{\log^2(J)}{n}}\end{aligned}$$

with probability at least $1 - (Jn)^{-2} - n^{-3}$ for $n \geq n_1$ and some constants $c'' > 0$.

Finally, to bound α_n, we once again appeal to Talagrand's inequality. By assumption, $M_J^2 = o(n)$, then there exists $n_2 \in \mathbf{N}$ such that for all $n \geq n_2$,

$$\begin{aligned}c''\sqrt{\frac{\log^2(J)}{n}} \geq& \frac{K(M_J\log(J) + \sqrt{n\log(J)})}{n} \\ &+ \sqrt{2\left(\frac{2M_J}{n}\frac{K(M_J\log(J) + \sqrt{n\log(J)})}{n} + \frac{C}{n}\right)x} + \frac{2M_J}{n}x \\ \geq& E\left[\mathbb{Z}\right] + \sqrt{2\nu_n x} + \frac{Ux}{3}\end{aligned}$$

for $J = n^p$ and $x = 3\log(n)$. Then for $n \geq n^* := \max\{n_1, n_2\}$,

$$\begin{aligned}P(\mathbb{Z} \geq \lambda) &\leq P\left(\mathbb{Z} \geq c''\sqrt{\frac{\log^2(J)}{n}}\right) + (Jn)^{-2} + n^{-3} \\ &\leq P\left(\mathbb{Z} \geq E\left[\mathbb{Z}\right] + \sqrt{2\nu_n\log(n^3)} + \frac{U\log(n^3)}{3}\right) + (Jn)^{-2} + n^{-3} \\ &\leq (Jn)^{-2} + 2n^{-3}.\end{aligned}$$

Recall that $\mathbb{Z} = \max_{1\leq j\leq J}|n^{-1}\sum_{i=1}^{n}\phi_j(X_i) - E\left[\phi_j(X_i)\right]| = \max_{1\leq j\leq J}|\hat{\theta}_j - \theta_j|$, so the above result suggests that we can take $\alpha_n = (Jn)^{-2} + 2n^{-3}$. Note there are many other permissible choices of α_n, and the proof can be modified accordingly if different α_n's are needed. □

1.7.6 Proof of Theorem 1.4.2

To establish the result for the post-processed $\hat{f}^*$ in 1.4.1, we first establish the result for $\tilde{f}_J$ in (1.23) and then use the results in Gajek (1986) to conclude. We keep J explicit throughout the proof and we will substitute $J = n^p$ at the end of the proof. Moreover, in order to bound the size of the selected set of indices T, for the convenience of the notation, we penalize using 2λ. Note that this is without loss of generality as we can multiply the original λ in Theorem 1.4.1 by $1/2$, and the results in Theorem 1.4.1 still hold (just with different constants and n^* in the proof).

The proof proceeds in six steps. In the first step, we decompose the MISE into the "variance" and "bias" components. We bound the cardinality of the set of selected indices in the second step. Next, we bound "variance" and "bias" separately by expressions of λ and α_n in the third and fourth steps respectively. In step 5 we establish that the λ and α_n we are using are the "correct" ones and we conclude in step 6.

Step 1: Decompose MISE. Let $f_J(\cdot) := \sum_{j=1}^{J} \theta_j \phi_j(\cdot)$ be the infeasible estimator for f, where θ_j's are the true but unknown Fourier coefficients. Then we have

$$
\begin{aligned}
& E_f\left[\|f - \tilde{f}_J\|_{L^2}^2\right] \\
=& E_f\left[\|f - f_J\|_{L^2}^2\right] + E_f\left[\|\tilde{f}_J - f_J\|_{L^2}^2\right] + 2E_f\left[\|(f - f_J)(\tilde{f}_J - f_J)\|_{L^2}\right] \\
=& E_f\left[\|f - f_J\|_{L^2}^2\right] + E_f\left[\|\tilde{f}_J - f_J\|_{L^2}^2\right] \\
=& \sum_{j=J+1}^{\infty} \theta_j^2 + E\left[\sum_{j\in T}(\hat{\theta}_j - \theta_j)^2 + \sum_{j\in T^c} \theta_j^2\right] \\
=& E\left[\sum_{j\in T}(\hat{\theta}_j - \theta_j)^2\right] + E\left[\sum_{j\in T^c} \theta_j^2 + \sum_{j=J+1}^{\infty} \theta_j^2\right]
\end{aligned}
$$

where the second and third equalities follow from orthonormality. Note that the last line corresponds to the variance-bias trade-off; however, the randomness of the set T of the selected indices no longer allows the interchange of expectation and summation.

Step 2: Cardinality of Selected Indices. Recall the definition of the selected indices T:

$$T = \{j \in \{1, \cdots, J\} : \hat{\theta}_j \geq 2\lambda\}$$

where $\lambda \geq (1-\alpha_n)$-quantile of $\|\hat{\theta}^J - \theta^J\|_\infty$. Then for $j \in T$, with probability at least $1-\alpha_n$,

$$2\lambda \leq |\hat{\theta}_j| \leq |\hat{\theta}_j - \theta_j| + |\theta_j| \leq \lambda + Aj^{-k}$$

where the last inequality holds by the definition of approximate sparsity set. This implies that for $j \in T$, with probability at least $1-\alpha_n$,

$$\lambda \leq Aj^{-k} \iff j \leq A^{\frac{1}{k}}\lambda^{-\frac{1}{k}}.$$

That is, $T \subseteq \{1 \leq j \leq J : j \leq A^{1/k}\lambda^{-1/k}\}$ with probability at least $1-\alpha_n$. This result establishes that with probability at least $1-\alpha_n$, the cardinality of the set of selected indices T satisfies

$$|T| \leq A^{\frac{1}{k}}\lambda^{-\frac{1}{k}}.$$

Step 3: Bound on Bias. To control the bias term $E\left[\sum_{j\in T^c}\theta_j^2 + \sum_{j=J+1}^{\infty}\theta_j^2\right]$, we need to control the random set T^c, the set of non-selected indices. By triangle inequality, we have

$$|\theta_j| \leq |\hat{\theta}_j| + |\hat{\theta}_j - \theta_j|$$

Note that since $\lambda \geq (1-\alpha_n)$-quantile of $\|\hat{\theta}^J - \theta^J\|_\infty$, we have $|\hat{\theta}_j - \theta_j| \leq \lambda$ with probability at least $1-\alpha_n$. Moreover, by definition, for $j \in T^c$, $|\hat{\theta}_j| < 2\lambda$. Combining the above, on T^c,

$$|\theta_j| \leq 3\lambda$$

with probability at least $1-\alpha_n$. This result will help us control the bias on the random set

T^c. Let E_n denote the event that

$$E_n := \{\lambda \geq \max_{j \leq J} |\hat{\theta}_j - \theta_j|\}$$

and note that the probability of this event E_n happening is at least $1-\alpha_n$, and on this event, $|\theta_j| \leq 3\lambda$ for $j \in T^c$. We will show that α_n can go to zero sufficiently fast for us to establish the minimax rate. Then we can establish the following: for some constants C_1, C_2, C_3,

$$\begin{aligned}
&E\left[\sum_{j\in T^c} \theta_j^2 + \sum_{j=J+1}^{\infty} \theta_j^2\right] \\
&\leq E\left[\sum_{j\in T^c \cap j\leq m} \theta_j^2 + \sum_{j=m+1}^{\infty} \theta_j^2\right] \\
&= E\left[\left(\sum_{j\in T^c \cap j\leq m} \theta_j^2 + \sum_{j=m+1}^{\infty} \theta_j^2\right) \cdot \mathbf{1}\{E_n\}\right] + E\left[\left(\sum_{j\in T^c \cap j\leq m} \theta_j^2 + \sum_{j=m+1}^{\infty} \theta_j^2\right) \cdot \mathbf{1}\{E_n^c\}\right] \\
&\leq E\left[9m\lambda^2 + C_1 m^{-2k+1}\right] + C_2 \cdot P(E_n^c)
\end{aligned}$$

where the first inequality holds for every $m \leq J$; the second inequality holds by that on the event E_n, $|\theta_j| \leq 3\lambda$, by Hölder's inequality, and by that the true density is bounded. Since the above expression holds for all $m \leq J$, we have the following

$$\begin{aligned}
&E\left[\sum_{j\in T^c} \theta_j^2 + \sum_{j=J+1}^{\infty} \theta_j^2\right] \\
&\leq E\left[\min_m 9m\lambda^2 + C_1 m^{-2k+1}\right] + C_2 \cdot P(E_n^c) \\
&\leq C_3 E\left[\lambda^{\frac{2k-1}{k}}\right] + \alpha_n C_2
\end{aligned}$$

where the last inequality holds by solving the minimization problem over m and by that the probability of event E_n^c is at most α_n. We will specify α_n and bound $E\left[\lambda^{\frac{2k-1}{k}}\right]$ explicitly.

Step 4: Bound on Variance. In this section, we establish bounds on $E\left[\sum_{j\in T}(\hat{\theta}_j - \theta_j)^2\right]$. Recall E_n is the event that $E_n = \{\lambda \geq \max_{j\in J} |\hat{\theta}_j - \theta_j|\}$, and in step 2 we have shown that

$|T| \leq A^{1/k}\lambda^{-1/k}$ on this event. Then

$$\begin{aligned}
&E\left[\sum_{j\in T}(\hat{\theta}_j-\theta_j)^2\right]\\
=&E\left[\sum_{j\in T}(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n\}\right]+E\left[\sum_{j\in T}(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n^c\}\right]\\
\leq&E\left[|T|\max_{j\in T}(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n\}\right]+\sum_{j=1}^{J}E\left[(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n^c\}\right]
\end{aligned}$$

where the inequality holds since (i) $\sum_{j\in T}(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n\} \leq |T|\max_{j\in T}(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n\}$ over the sample space; (ii) we can bound the sum over random $T \subseteq \{1,\cdots,J\}$ from above with the deterministic sum over $1 \leq j \leq J$, and then interchange the expectation and summation.

First, we bound $E\left[|T|\max_{j\in T}(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n\}\right]$ with the help of set E_n:

$$\begin{aligned}
&E\left[|T|\max_{j\in T}(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n\}\right]\\
\leq&A^{\frac{1}{k}}E\left[\lambda^{-\frac{1}{k}}\max_{j\in T}(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n\}\right]\\
\leq&A^{\frac{1}{k}}E\left[\lambda^{-\frac{1}{k}}\lambda^2\right]\\
=&A^{\frac{1}{k}}E\left[\lambda^{\frac{2k-1}{k}}\right]
\end{aligned}$$

where the first inequality holds since $|T| \leq A^{1/k}\lambda^{-1/k}$ on E_n and the second inequality holds since on E_n, $\max_{j\in J}|\hat{\theta}_j-\theta_j| \leq \lambda$.

Second, we bound $\sum_{j=1}^{J}E\left[(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n^c\}\right]$. By Cauchy-Schwarz,

$$\sum_{j=1}^{J}E\left[(\hat{\theta}_j-\theta_j)^2\mathbf{1}\{E_n^c\}\right] \leq \sum_{j=1}^{J}\left(E\left[(\hat{\theta}_j-\theta_j)^4\right]\right)^{\frac{1}{2}}(P(E_n^c))^{\frac{1}{2}}.$$

Moreover, for each j,

$$\begin{aligned}
E\left[(\hat{\theta}_j - \theta_j)^4\right] &= E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i) - E\left[\phi_j(X_i)\right]\right)^4\right] \\
&= n^{-4}E\left[\left(\sum_{i=1}^{n}\phi_j(X_i) - E\left[\phi_j(X_i)\right]\right)^4\right] \\
&\leq c_1 n^{-4}E\left[\left(\sum_{i=1}^{n}\left(\phi_j(X_i) - E\left[\phi_j(X_i)\right]\right)^2\right)^2\right] \\
&= c_1 n^{-2}E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left(\phi_j(X_i) - E\left[\phi_j(X_i)\right]\right)^2\right)^2\right] \\
&\leq c_1 n^{-2}E\left[\frac{1}{n}\sum_{i=1}^{n}\left(\phi_j(X_i) - E\left[\phi_j(X_i)\right]\right)^4\right] \\
&= c_1 n^{-2}E\left[\left(\phi_j(X_i) - E\left[\phi_j(X_i)\right]\right)^4\right]
\end{aligned}$$

where the first inequality holds by Marcinkiewicz-Zygmund inequality for some constant $c_1 > 0$, the second inequality holds by the convexity of the function $x \mapsto x^2$ and Jensen's inequality, and the last equality holds by i.i.d assumption. This derivation suggests that we need to bound the fourth central moment of $\phi_j(X_i)$. If the basis is uniformly bounded, the result is trivial. On the other hand, if the basis grows, i.e. $\max_{1\leq j\leq J}\|\phi_j(\cdot)\|_\infty \leq M_J$, we can bound $E\left[\phi_j^4(X_i)\right]$ explicitly. Note that for some constant $C > 0$,

$$E\left[\phi_j^4(X_i)\right] \leq M_J^2 E\left[\phi_j^2(X_i)\right] = M_J^2 \int \phi_j^2(x)f(x)dx \leq CM_J^2$$

where the last inequality holds by orthonormality and by that f is bounded. This gives us

$$E\left[(\hat{\theta}_j - \theta_j)^4\right] \leq C' M_J^2/n^2$$

for some constant $C' > 0$. Therefore, for some constant $c_2 > 0$,

$$\sum_{j=1}^{J} E\left[(\hat{\theta}_j - \theta_j)^2 \mathbf{1}\{E_n^c\}\right] \leq \sum_{j=1}^{J} \left(E\left[(\hat{\theta}_j - \theta_j)^4\right]\right)^{\frac{1}{2}} (P(E_n^c))^{\frac{1}{2}} \leq c_2 J M_J / n\alpha_n^{1/2}$$

where the second inequality holds by that $P(E_n^c) \leq \alpha_n$. We need $JM_J/n\alpha_n^{1/2}$ to go to zero sufficiently fast (at least as fast as the minimax rate), which is assumed and can be verified for a given orthonormal basis. Combining the above, we have an upper bound on the variance

$$E\left[\sum_{j\in T}(\hat{\theta}_j - \theta_j)^2\right] \leq c_1 E\left[\lambda^{\frac{2k-1}{k}}\right] + c_2 J M_J / n\alpha_n^{1/2}.$$

Step 5: The "Right" Penalization Parameter. We show the λ proposed in (1.28) will give us the "correct" minimax rate. Recall that λ is defined as follows

$$\lambda = \sqrt{\frac{\log(J)}{n}} \Phi^{-1}\left(1 - \frac{1}{2\sqrt{2\pi}} \frac{1}{\sqrt{2\log(J)}} \frac{1}{J}\right) \max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{Z}_{ij}^2\right)^{\frac{1}{2}}$$

where $\hat{Z}_{ij} := \phi_j(X_i) - n^{-1}\sum_{k=1}^{n}\phi_j(X_k)$. Then

$$\begin{aligned}
&E\left[\lambda^{\frac{2k-1}{k}}\right] \\
&= E\left[\left(\sqrt{\frac{\log(n)}{n}} \Phi^{-1}\left(1 - \frac{1}{2\sqrt{2\pi}} \frac{1}{\sqrt{2\log(J)}} \frac{1}{J}\right) \max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{Z}_{ij}^2\right)^{\frac{1}{2}}\right)^{\frac{2k-1}{k}}\right] \\
&\leq \left(\frac{\log(n)\log(2\sqrt{2\pi}\sqrt{2\log(J)}J)}{n}\right)^{\frac{2k-1}{2k}} E\left[\left(\max_{1\leq j\leq J}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{Z}_{ij}^2\right)^{\frac{1}{2}}\right)^{\frac{2k-1}{k}}\right] \\
&\leq \left(\frac{\log(n)\log(2\sqrt{2\pi}\sqrt{2\log(J)}J)}{n}\right)^{\frac{2k-1}{2k}} \left(E\left[\max_{1\leq j\leq J}\frac{1}{n}\sum_{i=1}^{n}\hat{Z}_{ij}^2\right]\right)^{\frac{2k-1}{2k}} \\
&\leq C'\left(\frac{\log(n)\log(J)}{n}\right)^{\frac{2k-1}{2k}} \left(E\left[\max_{1\leq j\leq J}\left|\frac{1}{n}\sum_{i=1}^{n}\phi_j^2(X_i) - E\left[\phi_j^2(X_i)\right]\right|\right] + C''\right)^{\frac{2k-1}{k}}
\end{aligned}$$

$$\leq C' \left(\frac{\log(n)\log(J)}{n} \right)^{\frac{2k-1}{2k}} \left(\frac{M_J^2 \log(J) + \sqrt{n M_J^2 \log(J)}}{n} + C'' \right)^{\frac{2k-1}{k}}$$
$$\leq C \left(\frac{\log(n)\log(J)}{n} \right)^{\frac{2k-1}{2k}}$$

where the first inequality holds by the property that $\Phi^{-1}(1-x) \leq \sqrt{2\log(1/x)}$ (see Lemma 1.8.2), the second inequality holds by Jensen's inequality, and the third inequality holds by the definition of $\hat{Z}_{ij}$ and by the fact that the density is bounded so that by orthonormality $E\left[\phi_j^2(X_i)\right] \leq C''$ for some constant $C'' > 0$, the fourth inequality holds by maximal inequality (see Lemma 1.8.3 in the appendix), and the last inequality holds by the assumption that $M_J^2 \leq n/\log(n)$ and $J = n^p$ for some constant $C > 0$.

Step 6: Conclusion. By Theorem 1.4.1, under Condition 1, there exists $n^* \in \mathbf{N}$ such that for all $n \geq n^*$,

$$\lambda \geq (1-\alpha_n) - \text{quantile of } \max_{1 \leq j \leq J} |\hat{\theta}_j - \theta_j|$$

with $\alpha_n = (Jn)^{-2} + 2n^{-3}$. Then combining results from step 1 to step 5, by the assumptions on J and M_J, we have for all $n \geq n^* = \max\{n_1, n_2\}$,

$$\begin{aligned}
&E_f\left[\|f - \tilde{f}_J\|_{L_2}^2\right] \\
&\leq E\left[\sum_{j \in T} (\hat{\theta}_j - \theta_j)^2\right] + E\left[\sum_{j \in T^c} \theta_j^2 + \sum_{j=J+1}^{\infty} \theta_j^2\right] \\
&\leq c_1 E\left[\lambda^{\frac{2k-1}{k}}\right] + c_2 \frac{JM_J}{n} \alpha_n^{\frac{1}{2}} + C_3 E\left[\lambda^{\frac{2k-1}{k}}\right] + \alpha_n C_2 \\
&\leq \tilde{C} \left(\frac{\log(n)\log(J)}{n} \right)^{\frac{2k-1}{2k}} + c_2 \frac{JM_J}{n} \left((Jn)^{-2} + 2n^{-3}\right)^{\frac{1}{2}} + C_2 \left((Jn)^{-2} + 2n^{-3}\right) \\
&= O\left(\frac{\log^2(n)}{n} \right)^{\frac{2k-1}{2k}}
\end{aligned}$$

where the last inequality holds by that $\alpha_n = (Jn)^{-2} + 2n^{-3} < n^{-(2k-1)/(2k)}$ and by the assumption that $(JM_J/n)\alpha_n^{1/2} \leq n^{-(2k-1)/(2k)}$. By Gajek (1986), the post-processed $\hat{f}^*$ has

smaller MISE, and note that our proof does not depend on a specific $f \in \tilde{\Theta}_k(\Phi, A, C)$, which proves the desired result

$$\sup_{f \in \tilde{\Theta}_k(\Phi,A,C)} E_f\left[\|f - \hat{f}^*\|_{L_2}^2\right] \leq \sup_{f \in \tilde{\Theta}_k(\Phi,A,C)} E_f\left[\|f - \tilde{f}_J\|_{L_2}^2\right] = O\left(\left(\frac{\log^2(n)}{n}\right)^{\frac{2k-1}{2k}}\right).$$

□

1.8 Additional Technical Results

Lemma 1.8.1. *Let $A > 0$ and $k > 1/2$ be some constants. Let the constant C in the definition of $\Theta_k(\Phi, A, C)$ and $\mathcal{A}_k(\Phi, A, C)$ be such that $C \geq A^2/(2k-1)$. Then*

$$\mathcal{E}_k(\Phi, A) \subseteq \Theta_k(\Phi, A, C) \subseteq \mathcal{A}_k(\Phi, A, C) \tag{1.33}$$

Proof of Lemma 1.8.1. First, note that by definition, $\mathcal{E}_k(\Phi, A)$ is a special case of $\Theta_k(\Phi, A, C)$ without reordering. In particular, in the definition of $\mathcal{E}_k(\Phi, A)$, since $|\theta_j| < Aj^{-k}$, then

$$\sum_{j=J+1}^{\infty} \theta_j^2 \leq A^2 \int_J^{\infty} t^{-2k} dt = \frac{A^2}{2k-1} J^{-2k+1}$$

This establishes $\mathcal{E}_k(\Phi, A) \subseteq \Theta_k(\Phi, A, C)$.

Moreover, since the restrictions on the tail sum $\sum_{j=J+1}^{\infty} \theta_j^2$ are identical in $\Theta_k(\Phi, A, C)$ and $\mathcal{A}_k(\Phi, A, C)$, the additional restrictions on individual θ_j makes $\Theta_k(\Phi, A, C)$ a subset of $\mathcal{A}_k(\Phi, A, C)$. □

Lemma 1.8.2. *Let Φ denote the CDF of the standard normal random variable, then for $x \geq 0$,*

$$\frac{1}{\sqrt{2\pi}} \frac{x}{x^2+1} \exp\left(-\frac{x^2}{2}\right) < 1 - \Phi(x) < \frac{1}{\sqrt{2\pi}} \frac{1}{x} \exp\left(-\frac{x^2}{2}\right).$$

Proof of Lemma 1.8.2. To show the upper bound, note that for $x \geq 0$

$$\begin{aligned} 1 - \Phi(x) &= \int_x^\infty \phi(s)ds \\ &= \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{s^2}{2}\right) ds \\ &< \frac{1}{\sqrt{2\pi}} \int_x^\infty \frac{s}{x} \exp\left(-\frac{s^2}{2}\right) ds \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{x} \exp\left(-\frac{x^2}{2}\right). \end{aligned}$$

To show the lower bound, let

$$h(x) := 1 - \Phi(x) - \frac{1}{\sqrt{2\pi}} \frac{x}{x^2+1} \exp\left(-\frac{x^2}{2}\right).$$

Since $h(0) > 0$, $h'(x) < 0$ for all $x \geq 0$, and $h(x) \to 0$ as $x \to \infty$, we must have $h(x) > 0$ for all $x \geq 0$. This gives us the lower bound. □

Lemma 1.8.3. *Let $\{X_i\}_{i=1}^n \sim X$ be an i.i.d sample with $X \in [0,1] \subseteq \mathbf{R}$. Suppose the density f of X is bounded above by some constant $\tilde{C}$ and that the orthonormal basis $\{\phi_j\}_{j=1}^\infty$ of $L^2([0,1], \mu)$ is such that $\max_{1 \leq j \leq J} \|\phi_j(\cdot)\|_\infty \leq M_J$ for some M_J that potentially grows with J for all J. Then for some constants K_1 and K_2,*

$$E\left[\max_{j \leq J} \left| \sum_{i=1}^n \phi_j(X_i) - E\left[\phi_j(X_i)\right] \right| \right] \leq K_1 \left(M_J \log(J) + \sqrt{n} \sqrt{\log(J)} \right)$$

and

$$E\left[\max_{j \leq J} \left| \sum_{i=1}^n \phi_j^2(X_i) - E\left[\phi_j^2(X_i)\right] \right| \right] \leq K_2 \left(M_J^2 \log(J) + \sqrt{n M_J^2} \sqrt{\log(J)} \right).$$

Proof of Lemma 1.8.3. First, since $\max_{1 \leq j \leq J} \|\phi_j(\cdot)\|_\infty \leq M_J$, we have

$$|\phi_j(X_i) - E\left[\phi_j(X_i)\right]| \leq 2M_J.$$

Since the density is bounded by $\tilde{C}$,

$$\max_{j \leq J} E\left[\phi_j^2(X)\right] = \max_{j \leq J} \int \phi_j^2(x) f(x) dx \leq \max_{j \leq J} \tilde{C} \int \phi_j^2(x) dx = \tilde{C}$$

where the last equality holds by orthonormality. This implies

$$Var\left(\sum_{i=1}^{n} \phi_j(X_i) - E\left[\phi_j(X_i)\right]\right) = \sum_{i=1}^{n} Var(\phi_j(X_i)) \leq nE\left[\phi_j^2(X_i)\right] \leq n\tilde{C}.$$

Then, by Bernstein's inequality (Lemma 2.2.9 in van der Vaart and Wellner (1996)),

$$P\left(\left|\sum_{i=1}^{n} \phi_j(X_i) - E\left[\phi_j(X_i)\right]\right| > x\right) \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{n\tilde{C} + \frac{2M_J x}{3}}\right)$$

and by maximal inequality (Lemma 2.2.10 in van der Vaart and Wellner (1996)),

$$E\left[\max_{j \leq J}\left|\sum_{i=1}^{n} \phi_j(X_i) - E\left[\phi_j(X_i)\right]\right|\right] \leq K_1\left(M_J \log(J) + \sqrt{n}\sqrt{\log(J)}\right)$$

for some fixed constant K_1.

The second part of the statement can be shown using similar arguments. By assumption, we have $\max_{1 \leq j \leq J} \|\phi_j^2\|_\infty \leq M_J^2$ for some potentially growing M_J for all J. Then

$$\begin{aligned} &Var\left(\sum_{i=1}^{n} \phi_j^2(X_i) - E\left[\phi_j(X_i)^2\right]\right) \\ &= nVar(\phi_j^2(X)) \leq nE\left[\phi_j^4(X)\right] \leq nM_J^2 E\left[\phi_j(X)^2\right] \leq n\tilde{C}M_J^2. \end{aligned}$$

Then by Bernstein's inequality

$$P\left(\left|\sum_{i=1}^{n} \phi_j^2(X_i) - E\left[\phi_j^2(X_i)\right]\right| > x\right) \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{n\tilde{C}M_J^2 + \frac{2M_J^2 x}{3}}\right)$$

which holds for all $1 \leq j \leq J$, and by Maximal inequality,

$$E\left[\max_{j\leq J}\left|\sum_{i=1}^{n}\phi_j^2(X_i) - E\left[\phi_j^2(X_i)\right]\right|\right] \leq K_2\left(M_J^2\log(J) + \sqrt{nM_J^2}\sqrt{\log(J)}\right)$$

for some fixed constant K_2. □

Bibliography

Belloni, A., Chernozhukov, V., Chetverikov, D., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80**, 2369–2429.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018). High-dimensional econometrics and regularized GMM. *arXiv:1806.01888*.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. *Stochastic Inequalities and Applications* 213–247. Birkhäuser, Basel.

Bunea, F., Tsybakov, A. B., Wegkamp, M. H., and Barbu, A. (2010). Spades and mixture models. *Annals of Statistics* **38**, 2525–2558.

Čencov, N. N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math. Dokl.* **3**, 1559–1562.

Chicken, E. and Cai, T. T. (2005). Block thresholding for density estimation: local and global adaptivity. *Journal of Multivariate Analysis* **95**, 76–106.

Diggle, P. J. and Hall, P. (1986). The selection of terms in an orthogonal series density estimator. *Journal of the American Statistical Association* **81**, 230–233.

Donoho, D.L., Liu, R.C., and MacGibbon, B. (1990). Minimax risk over hyperrectangles, and implications. *Annals of Statistics* **18**, 1416–1437.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *Annals of Statistics* **24**, 508–539.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics* **26** 879–921.

Efromovich, S. Y. (1986). Nonparametric estimation of a density of unknown smoothness. *Theory of Probability & Its Applications* **30**, 557–568.

Efromovich, S. Y. (2008). *Nonparametric Curve Estimation: Methods, Theory, and Applications*. Springer Science & Business Media.

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

Hall, P. (1986). On the rate of convergence of orthogonal series density estimators. *Journal of the Royal Statistical Society: Series B (Methodological)* **48**, 115–122.

Hall, P. (1987). Cross-validation and the smoothing of orthogonal series density estimators. *Journal of Multivariate Analysis* **21**, 189–206.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (2012). *Wavelets, Approximation, and Statistical Applications* **129**. Springer Science & Business Media.

Hart, J. D. (1985). On the choice of truncation point in Fourier series density estimation. *Journal of Statistical Computation and Simulation* **21**, 96–116.

Katznelson, Y. (2004). *An Introduction to Harmonic Analysis*. Cambridge University Press.

Kloeckner, B. (2012). A generalization of Hausdorff dimension applied to Hilbert cubes and Wasserstein spaces. *Journal of Topology and Analysis* **2**, 203–235.

Kronmal, R. and Tarter, M. (1986). The estimation of probability densities and cumulatives by Fourier series methods. *Journal of the American Statistical Association* **63**, 925–952.

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

Schwartz, S. C. (1967). Estimation of probability density by an orthogonal series. *Annals of Mathematical Statistics* **38**, 1261–1265.

Smolyak, S. A. (1960). The ϵ-entropy of classes $E_s^{\alpha,k}(B)$ and $W_s^{\alpha}(B)$ in metric L^2. *Dokl. Akad. Nauk SSSR* **131**, 30–33.

Tarter, M. and Kronmal, R. (1976). An introduction to the implementation and theory of nonparametric density estimation. *The American Statistician* **30**, 105–112.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Science & Business Media.

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

Wahba, G. (1981). Data-based optimal smoothing of orthogonal series density estimates. *Annals of Statistics* **9**, 146–156.

Watson, G. S. (1969). Density estimation by orthogonal series. *Annals of Mathematical Statistics* **40**, 1496–1498.

Yang, Y. and Barron, A. (1997). Information-theoretic determination of minimax rates of convergence. Technical Report 28, Dept. Statistics, Iowa State Univ.

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

CHAPTER 2

High-Dimensional Conditional Density Estimation

2.1 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.

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

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 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 converge 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 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

[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 have not been studied.

[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 of the related literature).

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 and Celisse (2010) for a comprehensive survey.

The rest of the paper is organized as follows. In section 2, we motivate by providing

a detailed review of the previously mentioned examples involving conditional densities. In section 3, we show the validity of the series representations of the conditional densities. We discuss the construction of our cross-validated estimator in detail in section 4 and establish the theoretical properties of our estimator in section 5. All proofs are provided in the appendix.

2.2 Examples

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

Example 2.2.1 (**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 2.2.2 (**Nonparametric Nonseparable Models**). In many nonparametric nonseparable models, the parameters of interest 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 relies 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. □

[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.

Example 2.2.3 (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 that $Y(t) \perp T|X$ for all $t \in [t_0, t_1]$, the average potential outcome at $T = t$ is identified as

$$E[Y(t)] = E\left[E\left[Y|T = t, f_{T|X}(t|X)\right]\right]. \tag{2.1}$$

The estimation of $E[Y(t)]$ based on the 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

$$E[Y(t)] = E[Y(\eta)|T = t]. \tag{2.2}$$

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

[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.

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 2.2.4 (**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

$$\partial_t E[Y(t)|X = x] = E[Y(\eta)|X = x] \tag{2.3}$$

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 (2.3) 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, estimate the conditional density as a univariate density. Nevertheless, the independence assumption $\epsilon \perp Z$ can be difficult 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 2.2.5 (**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

$$f^c_{Y_A}(y) = \int f_{Y_A|X_A}(y|x) \frac{dF_{X_B}(x)}{dF_{X_A}(x)} dF_{X_A}(x) \tag{2.4}$$

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 (2.4), 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 (2.4) 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. □

2.3 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 2.3.1. *(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 2.3.1 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

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

subsets $\mathbf{Y} \times \mathbf{X} \subseteq \mathbf{R} \times \mathbf{R}^d$, which, together with assumption 2.3.1 (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 2.3.1 (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 2.3.1. *Suppose Assumption 2.3.1 is satisfied. Then the following results hold:*

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

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

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

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

(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:

$$f_{Y|X}(y|x) = \sum_{j=1}^{\infty} E[\phi_j(Y)|X = x]\phi_j(y). \tag{2.5}$$

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 (2.5) 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.

2.4 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 2.3.1 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

$$\hat{f}_J(y|x) = \sum_{j=1}^{J} \hat{h}_j(x)\phi_j(y). \tag{2.6}$$

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 to 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

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

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

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

$$\|\hat{f} - f_{Y|X}\|_H^2 = \int \left(\hat{f}(y|x) - f_{Y|X}(y|x)\right)^2 d\nu_Y(y) dP_X(x) \tag{2.7}$$

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

$$\|\hat{f} - f_{Y|X}\|_H^2 - \|f_{Y|X}\|_H^2 = \int \hat{f}^2(y|x) - 2\hat{f}(y|x) f_{Y|X}(y|x) d\nu_Y(y) dP_X(x). \tag{2.8}$$

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.

[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 2.4.1. *Define a loss function*

$$Q((y,x),f) := \int f^2(t,x)d\nu_Y(t) - 2f(y,x) \tag{2.9}$$

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

$$R(\hat{f}) := E\left[Q((Y,X),\hat{f})\right] = E\left[\int \hat{f}^2(y|X)d\nu_Y(y) - 2\hat{f}(Y|X)\right]. \tag{2.10}$$

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

$$R(\hat{f}) = \|\hat{f} - f_{Y|X}\|_H^2 - \|f_{Y|X}\|_H^2. \tag{2.11}$$

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

$$R_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^n Q((Y_i,X_i),\hat{f}) = \frac{1}{n}\sum_{i=1}^n \int \hat{f}^2(y|X_i)d\nu_Y(y) - 2\hat{f}(Y_i|X_i). \tag{2.12}$$

We now have all the necessary 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.

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

[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.

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 (2.13) 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 $\tilde{h}_l(x) = K^{-1}\sum_{k=1}^K \hat{h}_l(D_k^{(n_T)})$.

$D^{(n_V)}$ 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 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 an empirical version of the risk on the validating sets. Specifically, we define the *K-fold empirical risk* of $\hat{f} \in \{\hat{f}_j\}_{j=1}^p$

[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 $D^{(m)}$.

as

$$R_{n,K}(\hat{f}) := \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_V}\sum_{i\in D_k^{(n_V)}} Q((Y_i,X_i),\hat{f}^{(n_T)}(D_k^{(n_T)})) \tag{2.13}$$

where recall $D_k^{(n_T)}$ is the k-th training set and $D_k^{(n_V)} = D^{(n)} \setminus D_k^{(n_T)}$ 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

$$\bar{f}^{(n)} := \frac{1}{K}\sum_{k=1}^{K}\hat{f}_{\hat{j}^*}^{(n_T)}(D_k^{(n_T)}) \quad \text{with} \quad \hat{j}^* = \arg\min_{1\le j\le p} R_{n,K}(\hat{f}_j). \tag{2.14}$$

Although $\bar{f}$ aggregates sub-sample estimators, it can still be expressed as 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 the 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 selecting $\hat{j}^*$.

[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.

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

$$C := \left\{ c \in \ell^2 : \sum_{j=2}^{\infty} c_j \phi_j(y) \geq -\phi_1 \right\}. \tag{2.15}$$

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.

2.5 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.

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

Theorem 2.5.1. *Let $\{(Y_i, X_i)\}_{i=1}^n$ be an i.i.d random sample distributed according to (Y, X) such that assumption 2.3.1 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 (2.6) are bounded by some constant M. Let $\bar{f}$ be the estimator defined in (2.14). Then for any constant $a > 0$, there exists a constant C that only depends on a such that*

$$E\left[\|\bar{f}^{(n)} - f_{Y|X}\|_H^2\right] \leq (1+a) \min_{1\leq j\leq p} E\left[\|\hat{f}_j^{(n_T)} - f_{Y|X}\|_H^2\right] + C\frac{\log p}{n_V}. \tag{2.16}$$

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 (2.16) 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 (2.16).

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 $f_{Y|X}$

[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.

are needed. Recall that the estimation error of $\hat{f}_J$ satisfies the bias-variance decomposition

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

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 (2.16), 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 2.5.2. *Suppose conditions in Theorem 2.5.1 are satisfied. Moreover, assume that*

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

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

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

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

Condition (i) in Theorem 2.5.2 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 the 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 er-

rors 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 2.3.1 (iii), the existence of the series expansion of the conditional density $f_{Y|X}$ requires that the tail sum satisfies $\lim_{J\to\infty}\sum_{j=J+1}^{\infty} E\left[h_j^2(X)\right] = 0$. 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 the 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 $E\left[(E[\phi_j(Y)|X])^2\right]$ 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 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

[16]For example, such conditions include but are not limited to the sparsity or approximate sparsity assumptions typically assumed in the literature.

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 2.5.2 can also be achieved in this point-wise case under the proposed conditions.

Theorem 2.5.3. *Suppose conditions in Theorem 2.5.1 and 2.5.2 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 $p \gtrsim n^{\delta/(\gamma+1)}$,

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

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 2.5.2, 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 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 2.5.1. 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 a 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 an estimator is left for future research.

2.6 Conclusion

In this paper, we introduce a data-driven conditional density estimator, designed to handle potentially high-dimensional conditioning variables. This estimator leverages a cross-validation procedure, and we have demonstrated an oracle inequality for its estimation error. Notably, this data-driven approach can integrate any new machine learning methods for estimating the conditional expectation in each series term. Consequently, our estimator can facilitate a better understanding of the dependence relationships between the economic variables, especially given the increasingly rich data sources and the growing complexity of the economic models.

2.7 Proofs

2.7.1 Proof of Proposition 2.3.1

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[17], 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.[18]

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,

$$\int f_{Y|X}^2(y|x) d\nu_Y dP_X < \infty \implies P_X(x \in \mathbf{X} : \int f_{Y|X}^2(y|x) d\nu_Y < \infty) = 1. \tag{2.17}$$

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 $f_{Y|X}(\cdot|x) \in L^2(\nu_Y)$,

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

[18] 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.

there exists $\{h_j(x)\}_{j=1}^{\infty} \in \ell^2$ such that

$$\lim_{J\to\infty} \sum_{j=J+1}^{\infty} h_j^2(x) = \lim_{J\to\infty} \int \left(f_{Y|X}(y|x) - \sum_{j=1}^{J} h_j(x)\phi_j(y) \right)^2 d\nu_Y = 0 \tag{2.18}$$

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

$$h_j(x) := \int \phi_j(y) f_{Y|X}(y|x) d\nu_Y. \tag{2.19}$$

Since (2.19) 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

$$P\left(h_j(X) = E[\phi_j(Y)|X]\right) = 1. \tag{2.20}$$

Then the claim follows from (2.18) and (2.20).

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

$$\lim_{J\to\infty} \sum_{j=1}^{J} E\left[(E[\phi_j(Y|X)])^2\right] < \infty. \tag{2.21}$$

Then by Fatou's Lemma,

$$E\left[\lim_{J\to\infty} \sum_{j=1}^{J} (E[\phi_j(Y|X)])^2\right] \leq \lim_{J\to\infty} \sum_{j=1}^{J} E\left[(E[\phi_j(Y|X)])^2\right] < \infty \tag{2.22}$$

which also implies that

$$P\left(\lim_{J\to\infty} \sum_{j=1}^{J} (E[\phi_j(Y|X)])^2 < \infty\right) = 1. \tag{2.23}$$

By orthonormality,

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

By $f_{Y|X} \in L^2(\nu_Y)$ and (2.22), $M(X) \in L^1(P_X)$. Therefore, by the second claim in the theorem, applying the dominated convergence theorem, we have

$$\lim_{J\to\infty} E\left[\int \left(f_{Y|X}(y|X) - \sum_{j=1}^{J} E[\phi_j(Y)|X]\phi_j(y) \right)^2 d\nu_Y \right] = 0. \tag{2.24}$$

To show the other direction, assume (2.24) holds. Note that by orthonormality,

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

Then by $f_{Y|X} \in L^2(\nu_Y \otimes P_X)$ and (2.24),

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

This concludes the proof. □

2.7.2 Proof of Lemma 2.4.1

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

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

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

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

□

2.7.3 Proof of Theorem 2.5.1

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 bound the expected loss in $\|\cdot\|_H$ norm by the sum of the "oracle" and a shifted empirical process. Finally, we use the 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

$$Q(\lambda f_1 + (1-\lambda) f_2) = \int (\lambda f_1 + (1-\lambda) f_2)^2 d\nu_Y(y) - 2(\lambda f_1 + (1-\lambda) f_2)$$

$$\leq \int \lambda f_1^2 + (1-\lambda) f_2^2 d\nu_Y(y) - 2(\lambda f_1 + (1-\lambda) f_2)$$
$$= \lambda Q(f_1) + (1-\lambda) Q(f_2)$$

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,

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

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

$$\begin{aligned} & R(\bar{f}^{(n)}) - R(f_{Y|X}) \\ &= (1+a)(R^*_{n,K} - R_{n,K}(f_{Y|X})) + (R(\bar{f}^{(n)} - R(f_{Y|X})) - (1+a)(R^*_{n,K} - R_{n,K}(f_{Y|X})) \\ &\leq (1+a)(R_{n,K}(\hat{f}_j) - R_{n,K}(f_{Y|X})) + (R(\bar{f}^{(n)}) - R(f_{Y|X})) - (1+a)(R^*_{n,K} - R_{n,K}(f_{Y|X})) \end{aligned} \tag{2.25}$$

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

$$\begin{aligned}
&E\left[R_{n,K}(\hat{f}_j) - R_{n,K}(f_{Y|X})\right] \\
&= E\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_V}\sum_{i\in D_k^{(n_V)}} Q((Y_i,X_i),\hat{f}_j^{(n_T)}(D_k^{(n_T)})) - Q((Y_i,X_i),f_{Y|X})\right] \\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_V}\sum_{i\in D_k^{(n_V)}} E\left[Q((Y_i,X_i),\hat{f}_j^{(n_T)}(D_k^{(n_T)}))\right] - E[Q((Y_i,X_i),f_{Y|X})] \\
&= E_{D^{(n_T)}}\left[R(\hat{f}_j^{(n_T)}(D^{(n_T)}))\right] - R(f_{Y|X})
\end{aligned} \tag{2.26}$$

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

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

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

$$
\begin{aligned}
& E\left[(R(\bar{f}^{(n)}) - R(f_{Y|X})) - (1+a)(R^*_{n,K} - R_{n,K}(f_{Y|X}))\right] \\
\leq & E\left[\frac{1}{K}\sum_{k=1}^{K} E_P\left[Q((Y,X), \hat{f}^{(n_T)}_{\hat{j}^*}(D^{(n_T)}_k))\right] - E_P[Q((Y,X), f_{Y|X})]\right. \\
& \left. - (1+a)\left(\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_V}\sum_{i\in D^{(n_V)}_k} Q((Y_i,X_i), \hat{f}^{(n_T)}_{\hat{j}^*}(D^{(n_T)}_k)) - Q((Y_i,X_i), f_{Y|X}))\right)\right] \\
= & \frac{1}{K}\sum_{k=1}^{K} E\left[E_P\left[(Q((Y,X), \hat{f}^{(n_T)}_{\hat{j}^*}(D^{(n_T)}_k)) - Q((Y,X), f_{Y|X}))\right]\right. \\
& \left. - \frac{1+a}{n_V}\sum_{i\in D^{(n_V)}_k} Q((Y_i,X_i), \hat{f}^{(n_T)}_{\hat{j}^*}(D^{(n_T)}_k)) - Q((Y_i,X_i), f_{Y|X}))\right] \\
\leq & E\left[\max_{1\leq j\leq p}(P - (1+a)P_{n_V})\left(Q((Y,X), \hat{f}^{(n_T)}_j(D^{(n_T)})) - Q((Y,X), f_{Y|X}))\right)\right].
\end{aligned}
\tag{2.27}
$$

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 $D^{(n_V)}_k$ 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 (2.25), (2.26), and (2.27), we have

$$
\begin{aligned}
& E\left[\|\bar{f}^{(n)} - f_{Y|X}\|^2_H\right] \\
= & E\left[R(\bar{f}^{(n)}) - R(f_{Y|X})\right] \\
\leq & \min_{1\leq j\leq p}(1+a)E_{D^{(n_T)}}\left[R(\hat{f}^{(n_T)}_j(D^{(n_T)}))\right] - R(f_{Y|X}) \\
& + E\left[\max_{1\leq j\leq p}(P - (1+a)P_{n_V})\left(Q((Y,X), \hat{f}^{(n_T)}_j(D^{(n_T)})) - Q((Y,X), f_{Y|X}))\right)\right] \\
\leq & \min_{1\leq j\leq p}(1+a)E\left[\|\hat{f}^{(n_T)}_j - f_{Y|X}\|^2_H\right] \\
& + E\left[\max_{1\leq j\leq p}(P - (1+a)P_{n_V})\left(Q((Y,X), \hat{f}^{(n_T)}_j(D^{(n_T)})) - Q((Y,X), f_{Y|X}))\right)\right]
\end{aligned}
\tag{2.28}
$$

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 maximum of the shifted empirical process term in (2.28) 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,

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

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

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

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

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

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

$$P\left[\max_{1\le j\le p} E[G_j(Z)] - (1+a)\frac{1}{n}\sum_{i=1}^{n} G_j(Z_i) \ge x\right] \le p\exp\left(-C_1 n\frac{x}{\tilde{M}}\right). \tag{2.29}$$

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

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

where the first inequality holds since $E[X] = \int_{\mathbf{R}} 1_{x\ge 0}(x) - F_X(x)dx$; the second inequality holds since the probability is bounded above by one; the third inequality holds by (2.29); the last inequality holds using the fact that $\int_u^\infty \exp(-Bt)dt \le \exp(-Bu)/B$ (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 $x(p) \le \log(ep)$. Let $u = \tilde{M}x(p)/(nC_1)$, we have

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

Therefore, we conclude that, for some constant C_2 that only depends on a and C_1,

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

Note that throughout the derivation, we kept the constant $\tilde{M}$ explicit to accommodate the

possibility of $\tilde{M}$ potentially growing with p.[19]

Step 4: Bound on Shifted Empirical Process

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

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

where Q is the loss defined in 2.9. 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

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

[19]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.

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

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

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

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

where the second line holds by the law of iterated expectation and the fourth line holds by

boundedness of $f_{Y|X}$. Therefore, by combining the above results, 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

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

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. □

2.7.4 Proof of Theorem 2.5.2

First, given that V is fixed, the training sample size n_T and testing/validating sample size n_V are in 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

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

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

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 $E\left[(\hat{h}_k(X)-h_k(X))^2\right] \asymp n^{-\delta}$ and bias $\sum_{k=j+1}^{\infty}E\left[h_k^2(X)\right] \lesssim j^{-\gamma}$, which implies

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

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

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

Combining this result with the oracle inequality in 2.5.1, we have the desired result. □

2.7.5 Proof of Theorem 2.5.3

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 $J \ge 1$,

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

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

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

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

Now consider the first term $E\left[\int \sum_{j=1}^{J} (h_j(x) - \hat{h}_j(x))^2 \phi_j^2(y) dP_X(x)\right]$. 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

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

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

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

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

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

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

2.5.2, we conclude that

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

□

Bibliography

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

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

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.

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

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*.

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*.

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.

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.

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.

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.

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

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

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

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

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

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

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

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.*

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.

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

CHAPTER 3

Difference-in-Differences With Continuous Treatment Under Double Machine Learning

3.1 Introduction

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 further 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. (2024); D'Haultfoeuille et al. (2021); de Chaisemartin et al. (2022). For instance, Callaway et al. (2024) 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 parameters: the average treatment effect on the treated (ATT) at any given treatment intensity. As noted in Abadie (2005), the (unconditional) parallel trends assumption[1] 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 that 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 function 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

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

and the causal parameter using the same sample, another source of bias due 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 provide consistent variance estimators based on cross-fitting. In addition, to illustrate the usefulness of our method, we revisit Acemoglu and Finkelstein (2008) which studies the impact of the 1983 Medicare payment system reform on the heavily regulated healthcare industry.

3.2 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[2]: 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 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 treat-

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

ment dose/intensity $D = d \in [d_L, d_H]$ according to some continuous distribution. Moreover, we will assume throughout that assumption 2.3.1 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 3.2.1. To formalize the mixture distribution of the treatment variable, consider a measure $\nu = \delta_0 + \lambda$, 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 3.2.1 (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 3.2.2 (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 observation* 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 3.2.3 (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) $0 < P(D = 0) < 1$ *and* D *admits a strictly positive probability density* f_D *on* (d_L, d_H).

Assumption 3.2.4 (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 3.2.1 and 3.2.2 are standard in the DiD literature. In particular, Assumption 3.2.1 does not allow the covariates to vary over time, while Assumption 3.2.2(ii) requires that the sample is not stratified by the outcome, treatment, or covariates.[3] Moreover, Assumption 3.2.3 describes the requirements for 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 3.2.4, 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, the 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 dynam-

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

ics. Moreover, covariates allow 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* $d \in (d_L, d_H)$:

$$ATT(d) := E[Y_t(d) - Y_t(0)|D = d]. \tag{3.1}$$

The interpretation of this parameter is analogous to the cases with discrete treatment variables: the expected effect of 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[4].

The following theorem presents the main results of this section, in which we establish the identifications of $ATT(d)$ for both repeated outcomes and repeated cross-sections settings.

Theorem 3.2.1 (Identification of ATT)**.**

- *(Repeated Outcomes) Suppose Assumptions 3.2.1, 3.2.3, and 3.2.4 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 3.2.2, 3.2.3, and 3.2.4 hold. Then,*

[4]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. (2024) Section 3.3 for details.

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 begin 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

$$E[Y_t(d) - Y_{t-1}(0)|D=d] = E[Y_t - Y_{t-1}|D=d]. \tag{3.2}$$

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

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

$$= E\left[(Y_t - Y_{t-1})\mathbf{1}\{D=0\}\frac{f_{D|X}(d|X)}{f_D(d)P(D=0|X)}\right] \tag{3.3}$$

Subtracting (3.3) from (3.2), we obtain the desired result. In particular, in the third equality in (3.3), 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 3.2.1, one can build estimators for $ATT(d)$ using the estimated sample analogs. 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[5], which makes such "plug-in" estimators less desirable. One way to alleviate such biases is to consider alternative estimating equations that reduce the influence of the nuisance parameters on the causal parameters. We formalize this idea in detail in the next section.

3.3 Orthogonal Scores

In this section, we use the repeated outcomes case as our main example for illustration as the analogous discussion on repeated cross-sections only requires minor modifications.

We begin by introducing the notion of *Neyman orthogonality*. For simplicity, consider the following notations: let $\theta_0 \in \Theta \subset \mathbf{R}$ be the low-dimensional parameter of interest, e.g., $ATT(d)$ in our case; let $\rho_0 \in \mathcal{H}$ denote the true low-dimensional nuisance parameters, e.g., 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[6] $f_{D|X}(d|X)$ 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

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

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

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[7].

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

Definition 3.3.1 (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, (i) says that the score ψ identifies the parameter of interests θ_0 while (ii) ensures that the first-order bias from estimating the *infinite-dimensional* nuisance parameters is zero.

Recall that in the repeated outcomes case,

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

where $\Delta Y := Y_t - Y_{t-1}$. This expression has two features that are worth noting. First, if $f_D(d)$ is estimated nonparametrically, 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, a score based on this expression does not satisfy Neyman orthogonality, and an adjustment term has to be added.

[7]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.

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 $E[\mathbf{1}\{D=0\}|X]$, the conditional density $f_{D|X}(d|X)$ presents additional challenges. To address this issue, we rely on the following observation (e.g., Fan et al. (1996)):

$$f_{D|X}(d|x) = \lim_{h\to 0} E[K_h(D-d)|X=x], \quad K_h(u) := \frac{1}{h}K\left(\frac{u}{h}\right) \tag{3.4}$$

where $K(\cdot)$ is a kernel function. The proof of this result can be established under mild regularity conditions and is given in the appendix. Replacing $f_{D|X}(d|x)$ by $E[K_h(D-d)|X=x]$, we define $ATT_h(d)$ as follows:

$$\begin{aligned} ATT_h(d) :=& E\left[\Delta Y \frac{K_h(D-d)}{f_D(d)}\right] - E\left[\Delta Y \mathbf{1}\{D=0\}\frac{E[K_h(D-d)|X]}{f_D(d)P(D=0|X)}\right] \\ =& E\left[\Delta Y \frac{K_h(D-d)P(D=0|X) - \mathbf{1}\{D=0\}E[K_h(D-d)|X]}{f_D(d)P(D=0|X)}\right], \end{aligned} \tag{3.5}$$

which is an expression that consists of only conditional expectations. Notably, it can be shown that

$$ATT(d) = \lim_{h\to 0} ATT_h(d),$$

which suggests that we can work with $ATT_h(d)$ instead. In particular, define the bias $B_h(d) := ATT_d - ATT_h(d)$, one can show that $B_h(d) = O(h^2)$, and we defer the formal result and proof to the next section.

For notation simplicity, we now formally define $ATT_h(d)$ in both settings.

Definition 3.3.2 (Repeated Outcomes)**.**

$$ATT_h(d) = E\left[\frac{K_h(D-d)g(X) - \mathbf{1}\{D=0\}f_h(d|X)}{f_D(d)g(X)}\left(\Delta Y - \mathcal{E}_{\Delta Y}(X)\right)\right] \tag{3.6}$$

where $\Delta Y = Y_t - Y_{t-1}$.

Definition 3.3.3 (Repeated Cross-Sections)**.**

$$ATT_h(d) = E\left[Y^\lambda \frac{K_h(D-d)P(D=0|X) - \mathbf{1}\{D=0\}E[K_h(D-d)|X]}{f_D(d)P(D=0|X)}\right] \tag{3.7}$$

where $Y^\lambda := \frac{T-\lambda}{\lambda(1-\lambda)}Y$.

Our goal is to construct a score that satisfies Neyman orthogonality for each h, and then take the limit as $h \to 0$. The next lemma presents such scores. To simplify the expressions, denote: $g(X) := P(D=0|X)$; $f_h(d|X) := E[K_h(D-d)|X]$; $\mathcal{E}_{\Delta Y}(X) := E[\Delta Y|X, D=0]$; $\mathcal{E}_{\lambda Y}(X) := E\left[\frac{T-\lambda}{\lambda(1-\lambda)}Y|X, D=0\right]$ with $\lambda = P(T=1)$; $f_d := f_D(d)$.

Lemma 3.3.1. *Suppose there exists* $M_h^{(1)} \in L^1(P_{Y_{t-1},Y_t,D,X})$ *and* $M_h^{(2)} \in L^1(P_{Y,T,D,X})$ *such that* $|\psi_h^{(1)}| \leq M_h^{(1)}$ *and* $|\psi_h^{(2)}| \leq M_h^{(2)}$ *almost surely. Then the scores* $\psi_h^{(1)}$ *and* $\psi_h^{(2)}$ *satisfy Neyman orthogonality defined in (3.3.1), where*

(i) for the repeated outcomes setting,

$$\psi_h^{(1)} := \frac{K_h(D-d)g(X) - \mathbf{1}\{D=0\}f_h(d|X)}{f_D(d)g(X)}\left(\Delta Y - \mathcal{E}_{\Delta Y}(X)\right) - ATT_h(d); \tag{3.8}$$

(ii) for the repeated cross-sections setting,

$$\psi_h^{(2)} := \frac{K_h(D-d)g(X) - \mathbf{1}\{D=0\}f_h(d|X)}{f_D(d)g(X)}\left(\frac{T-\lambda}{\lambda(1-\lambda)}Y - \mathcal{E}_{\lambda Y}(X)\right) - ATT_h(d). \tag{3.9}$$

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 3.3.1. The assumption on the existence of integrable functions $M_h^{(1)}$ and $M_h^{(2)}$ 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 $\psi_h^{(1)}$ and $\psi_h^{(2)}$ whenever the context is clear.

We note that in these new scores, the infinite-dimensional nuisance parameters are $f_h(d|X)$, $g(X)$, $\mathcal{E}_{\Delta Y}(X)$, 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 $ATT_h(d)$'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 $ATT_h(d)$'s using these scores and establish their asymptotic properties.

3.4 Estimation and Inference

As mentioned in the introduction, constructing DML estimators involves two main steps. In the previous section, we addressed the first step by establishing scores that satisfy Neyman orthogonality, as detailed in Lemma 3.3.1. These scores are then utilized in conjunction with the second critical aspect of DML estimators — the cross-fitting techniques. These techniques aim to reduce the overfitting bias that arises when estimating nuisance parameters using machine learning methods. With these key components in place, we can construct DML estimators following the procedure proposed by CCDDHNR (2018).

First, we partition the random sample I_N 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 $I_k^c := I_N \setminus I_k$ to estimate the nuisance parameters with the preferred machine learning methods. Next, we compute sample averages according to (3.6)/(3.7) using the estimated nuisance parameters evaluated at the sample I_k to obtain the k-th estimate $\widehat{ATT}_k(d)$ for $ATT(d)$. Finally, we average through the K estimates to obtain the final estimator. The following algorithms summarize the procedure.

Algorithm 3.4.1 (CDID Estimator, Repeated Outcomes)**.** *Let* $\{I_k\}_{k=1}^K$ *denote a 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* $I_k^c := I_N \setminus I_k$ *denote the complement.*

- *Step 1: for each* k*, construct*

$$\widehat{ATT}_k(d) := \frac{1}{n} \sum_{i \in I_k} \frac{K_h(D_i - d)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{h,k}(d|X_i)}{\hat{f}_{d,k}\hat{g}_k(X_i)} \left(\Delta Y_i - \hat{\mathcal{E}}_{\Delta Y,k}(X_i)\right)$$

  where $\hat{f}_{d,k}, \hat{f}_{h,k}, \hat{g}_k, \hat{\mathcal{E}}_{\Delta Y,k}$ *are the estimators of* $f_d, f_h(d|X), g(X)$ *and* $\mathcal{E}_{\Delta Y}(X)$ *respectively using the rest of the sample* I_k^c*. In particular,* $\hat{f}_{d,k}$ *is a kernel density estimator, and* $\hat{f}_{h,k}$*,* $\hat{g}_k$ *and* $\hat{\mathcal{E}}_{\Delta Y,k}$ *are estimated using ML methods (e.g. random forests or deep neural networks).*

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

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

Algorithm 3.4.2 (CDID Estimator, Repeated Cross-Sections)**.** *Let* $\{I_k\}_{k=1}^K$ *denote a 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* $I_k^c := I_N \setminus I_k$ *denote the complement.*

- *Step 1: for each* k*, construct*

$$\widehat{ATT}_k(d) := \frac{1}{n} \sum_{i \in I_k} \frac{K_h(D_i - d)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{h,k}(d|X_i)}{\hat{f}_{d,k}\hat{g}_k(X_i)} \times \left(\frac{T_i - \hat{\lambda}_k}{\hat{\lambda}_k(1 - \hat{\lambda}_k)} Y_i - \hat{\mathcal{E}}_{\lambda Y,k}(X_i) \right)$$

  where $\hat{f}_{d,k}, \hat{f}_{h,k}, \hat{g}_k, \hat{\mathcal{E}}_{\lambda Y,k}$ *are the estimators of* $f_d, f_h(d|X), g(X)$ *and* $\mathcal{E}_{\lambda Y}(X)$ *respectively using the rest of the sample* I_k^c*. In particular,* $\hat{f}_{d,k}$ *is a kernel density estimator, and* $\hat{f}_{h,k}$*,* $\hat{g}_k$ *and* $\hat{\mathcal{E}}_{\Delta Y,k}$ *are estimated using ML methods (e.g. random forests or deep neural networks).*

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

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

Remark 3.4.1. It is important to note that at each $k = 1, \cdots, K$, the nuisance parameters and the $\widehat{ATT}_k(d)$ 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.

Next, we state regularity conditions that allow us to prove the asymptotic normality of our DML estimators.

Assumption 3.4.1 (Kernel)**.** *The kernel function $K(\cdot)$ satisfies:*

(i) $K(\cdot)$ is bounded and differentiable;

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

Moreover, for notational simplicity, define $K_h(u) := h^{-1}K(u/h)$.

Assumption 3.4.2 (Bounds and Smoothness, Repeated Outcomes)**.**

(i) For some constants $0 < c < 1$ and $0 < C < \infty$, $f_d > c$, $|Y_{t-1}| < C$, $|Y_t| < C$, $|f_h(d|X)| < C$, and $|\mathcal{E}_{\Delta Y}(X)| < C$ almost surely;

(ii) for some constants $0 < \kappa < \frac{1}{2}$ and for all $h > 0$, $\kappa < g(X) < 1 - \kappa$ almost surely;

(iii) f_d is twice continuously differentiable at $D = d \in (d_L, d_H)$ with bounded second derivatives;

(iv) $f_{D|X}(d|x)$ *is twice continuously differentiable at* $d \in (d_L, d_H)$ *with bounded second derivatives uniformly in* $supp(X)$*;*

(v) joint density $f_{\Delta Y,D}(t,d)$ *is twice continuously differentiable in its first argument with uniformly bounded second derivatives over* $supp(\Delta Y)$ *for each* $d \in (d_L, d_H)$*.*

Assumption 3.4.3 (Rates, Repeated Outcomes)**.**

(i) The kernel bandwidth h *is a deterministic sequence that depends on* N *and satisfies* $Nh \to \infty$ *and* $\sqrt{Nh^5} = o(1)$*;*

(ii) there exists a sequence $\varepsilon_N \to 0$ *such that* $h^{-1}\varepsilon_N^2 = o(1)$*;*

(iii) with probability tending to 1, $\|\hat{f}_h(d|X) - f_h(d|X)\|_{P,2} \leq h^{-1/2}\varepsilon_N$, $\|\hat{g}(X) - g(X)\|_{P,2} \leq \varepsilon_N$, $\|\hat{\mathcal{E}}_{\Delta Y}(X) - \mathcal{E}_{\Delta Y}(X)\|_{P,2} \leq \varepsilon_N$*;*

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

Assumption 3.4.4 (Bounds and Smoothness, Repeated Cross-Sections)**.**

(i) For some constants $0 < c < 1$ *and* $0 < C < \infty$, $f_d > c$, $c < \lambda < 1 - c$, $|Y| < C$, $|f_h(d|X)| < C$, *and* $|\mathcal{E}_{\lambda Y}(X)| < C$ *almost surely;*

(ii) for some constants $0 < \kappa < \frac{1}{2}$ *and for all* $h > 0$, $\kappa < g(X) < 1 - \kappa$ *almost surely;*

(iii) f_d *is twice continuously differentiable at* $D = d \in (d_L, d_H)$ *with bounded second derivatives;*

(iv) $f_{D|X}(d|x)$ *is twice continuously differentiable at* $d \in (d_L, d_H)$ *with bounded second derivatives uniformly in* $supp(X)$*;*

(v) joint density $f_{Y^\lambda,D}(t,d)$ *is twice continuously differentiable in its first argument with uniformly bounded second derivatives over* $supp(Y^\lambda)$ *for each* $d \in (d_L, d_H)$*.*

Assumption 3.4.5 (Rates, Repeated Cross-Sections)**.**

(i) *The kernel bandwidth h is a deterministic sequence that depends on N and satisfies $Nh \to \infty$ and $\sqrt{Nh^5} = o(1)$;*

(ii) *there exists a sequence $\varepsilon_N \to 0$ such that $h^{-1}\varepsilon_N^2 = o(1)$;*

(iii) *with probability tending to 1, $\|\hat{f}_h(d|X) - f_h(d|X)\|_{P,2} \leq h^{-1/2}\varepsilon_N$, $\|\hat{g}(X) - g(X)\|_{P,2} \leq \varepsilon_N$, $\|\hat{\mathcal{E}}_{\lambda Y}(X) - \mathcal{E}_{\lambda Y}(X)\|_{P,2} \leq \varepsilon_N$;*

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

Kernel plays a central role in our analysis. Not only do we use the kernel to estimate the low-dimensional parameter $f_D(d)$ given its well-established theoretical properties, but we also use the kernel to approximate the point mass at $D = d$ as well as the conditional density $f_{D|X}(d|X)$. In Assumption 3.4.1, we assume the standard regularity conditions for the kernel function, which are crucial for establishing the asymptotic normality of our estimator and are easy to verify. Moreover, Assumptions 3.4.2, 3.4.4 impose bounds and smoothness conditions on the outcome variable and distributions/conditional distributions. In addition, Assumptions 3.4.3, 3.4.5 impose restrictions on the kernel bandwidth as well as the quality of the nonparametric estimators of nuisance parameters: (i) and (ii) require the kernel bandwidth h to be under-smoothing (but not too much) so that the bias vanishes asymptotically (otherwise asymptotic normality still holds but not centered at θ_0); (iii) assumes the estimators of the nuisance parameters to satisfy certain rates of convergence; and (iv) ensures that such estimators are bounded.

Remark 3.4.2. First, if $\varepsilon_N = o(N^{-1/4})$, then Assumption 3.4.3 (i) and (ii) together imply that $h = o(N^{-1/5})$ and $h \geq O(N^{-1/2})$. However, as we will see shortly, in order to show the consistency of the variance estimator, we need to assume additionally that $h^{-2}\varepsilon_N^2 = o(1)$. This suggests that we need an under-smoothing kernel bandwidth h but we cannot under-

smooth too much. Second, while the DML literature typically assumes that the estimators of the nuisance parameters converge at rate $\varepsilon_N = o(N^{-1/4})$, see CCDDHNR (2018) for example, we allow the conditional density $\hat{f}_h$ to converge at an even slower rate. This does not contradict the existing DML literature since the estimator for our target parameter can not achieve $\sqrt{N}$ rate due to the presence of a continuous treatment variable.

Before stating the main theorems of this section, we first introduce a lemma that characterizes the bias $ATT(d) - ATT_h(d)$ in terms of the kernel bandwidth h.

Lemma 3.4.1 (Bias of $ATT_h(d)$)**.** *Suppose assumptions 3.4.1, 3.4.2, 3.4.3 hold for the repeated outcomes case, and assumptions 3.4.1, 3.4.4, 3.4.5 hold for the repeated cross-sections case. Then $B_h(d) := ATT(d) - ATT_h(d)$ satisfies $B_h(d) = O(h^2)$ for any $d \in (d_L, d_H)$.*

The proof is given in the appendix. The results suggest that the bias from the kernel approximation can be controlled by choosing an appropriate bandwidth. In particular, for an under-smoothing bandwidth, the bias does not affect the asymptotic distribution of our estimators.

The next theorem is the main result of this section that establishes the asymptotic normality of our estimators for $ATT(d)$.

Theorem 3.4.1 (Asymptotic Normality)**.**

- *(Repeated Outcomes) Suppose assumptions 3.2.1, 3.2.3, 3.2.4, 3.4.1, 3.4.2, and 3.4.3 hold. Then*

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

where

$$\sigma_{1N}^2 := E\left[\left(\psi_h^{(1)}(Z, \theta_{0h}, f_d^0, \eta_0) - \frac{\theta_{0h}}{f_d^0}\left(K_h(D-d) - E[K_h(D-d)]\right)\right)^2\right] \tag{3.10}$$

for $\theta_{0h} := ATT_h(d)$ *defined as in (3.6) and* $\psi_h^{(1)}$ *defined as in (3.8).*

- *(Repeated Cross-Sections) Suppose assumptions 3.2.2, 3.2.3, 3.2.4, 3.4.1, 3.4.4, and 3.4.5 hold. Then*

$$\frac{\widehat{ATT}(d) - ATT(d)}{\sigma_{2N}/\sqrt{N}} \quad \to^d \quad N(0,1)$$

where

$$\sigma_{2N}^2 := E\left[\left(\psi_h^{(2)}(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0) - \frac{\theta_{0h}}{f_d^0}\left(K_h(D-d) - E[K_h(D-d)]\right)\right)^2\right] \tag{3.11}$$

for $\theta_{0h} := ATT_h(d)$ *defined as in (3.7) and* $\psi_h^{(2)}$ *defined as in (3.9).*

The proof follows the general framework for DML estimators studied in CCDDHNR (2018) with modifications to accommodate the presence of kernel functions. The asymptotic variance roughly consists of two parts that contribute to the slower than $\sqrt{N}$ rate: the part from the orthogonal score ψ_h that grows with h and the part from the kernel used to nonparametrically estimate the density $f_D(d)$. Under the assumptions, our estimator $\widehat{ATT}(d)$ converges at rate $\sqrt{Nh}$, which is slower than the parametric rate $\sqrt{N}$ but comparable to the optimal rate for 1-dimensional nonparametric estimation.

In practice, to establish a point-wise confidence interval for $ATT(d)$, we need consistent estimators for the asymptotic variances established in the theorem. Following CCDDHNR (2018) and Chang (2020), we consider the following cross-fitted variance estimators. For notational simplicity, we use $\hat{\theta}_h := \widehat{ATT}(d)$ to denote our cross-validated estimators, and $E_{n,k}$ to denote the empirical average using the subsample I_k.

Definition 3.4.1 (Cross-fitted Variance Estimator)**.**

- *(Repeated Outcomes)*

$$\hat{\sigma}_{1N}^2 := \frac{1}{K}\sum_{k=1}^{K} E_{n,k}\left[\left(\psi_h^{(1)}(Z, \hat{\theta}_h, \hat{f}_{d,k}, \hat{\eta}_k) - \frac{\hat{\theta}_h}{\hat{f}_{d,k}}\left(K_h(D-d) - \hat{f}_{d,k}\right)\right)^2\right] \tag{3.12}$$

- *(Repeated Cross-Sections)*

$$\hat{\sigma}_{2N}^2 := \frac{1}{K}\sum_{k=1}^{K} E_{n,k}\left[\left(\psi_h^{(2)}(Z,\hat{\theta}_h,\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k) - \frac{\hat{\theta}_h}{\hat{f}_{d,k}}\left(K_h(D-d)-\hat{f}_{d,k}\right)\right)^2\right] \tag{3.13}$$

Then, with these variance estimators, 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. The following theorem establishes the consistency of the cross-fitted variance estimators for both settings.

Theorem 3.4.2. *Suppose assumptions in Theorem 3.4.1 hold. In addition, assume that* $h^{-2}\varepsilon_N^2 = o(1)$. *Then*

- *(Repeated Outcomes)*

$$\hat{\sigma}_{1N}^2 = \sigma_{1N}^2 + o_p(1)$$

  where $\hat{\sigma}_{1N}^2$ *is defined as in (3.12) and* σ_{1N}^2 *is defined as in (3.10);*

- *(Repeated Cross-Sections)*

$$\hat{\sigma}_{2N}^2 = \sigma_{2N}^2 + o_p(1)$$

  where $\hat{\sigma}_{2N}^2$ *is defined as in (3.13) and* σ_{2N}^2 *is defined as in (3.11).*

Alternatively, we can consider a multiplier bootstrap procedure to construct confidence intervals. Notably, such procedure 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 $\{Y_{i,t-1}, Y_{i,t}, D_i, X_i\}_{i=1}^N$ for repeated outcomes case, or independent of $\{Y_i, T_i, D_i, X_i\}_{i=1}^N$ for repeated cross-sections case, 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 estimates based on the following expressions.

Definition 3.4.2 (Multiplier Bootstrap)**.**

- *(Repeated Outcomes)*

$$\widehat{ATT}(d)_b^* := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\xi_i \frac{K_h(D_i - d)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{h,k}(d|X_i)}{\hat{f}_{d,k}\hat{g}_k(X_i)} \times \left(\Delta Y_i - \hat{\mathcal{E}}_{\Delta Y,k}(X_i)\right). \tag{3.14}$$

- *(Repeated Cross-Sections)*

$$\widehat{ATT}(d)_b^* := \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\xi_i \frac{K_h(D_i - d)\hat{g}_k(X_i) - \mathbf{1}\{D_i = 0\}\hat{f}_{h,k}(d|X_i)}{\hat{f}_{d,k}\hat{g}_k(X_i)} \times \left(\frac{T_i - \hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)}Y_i - \hat{\mathcal{E}}_{\lambda Y,k}(X_i)\right). \tag{3.15}$$

Let $\hat{c}_\alpha$ denote the α's quantile of $\{\widehat{ATT}(d)_b^* - \widehat{ATT}(d)\}_{b=1}^B$, a $1-\alpha$ confidence interval can be constructed 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.

In the next section, we put our theory into practice by applying our methods to a notable study by Acemoglu and Finkelstein (2008) in which the research design can be reframed as continuous DiD.

3.5 Empirical Application: Acemoglu and Finkelstein (2008)

3.5.1 Background

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[8] 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.

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

$$Y_{i,t} = \alpha_i + \gamma_t + X'_{i,t}\eta + \beta \cdot (D_i \cdot \text{Post}_t) + \varepsilon_{i,t}, \tag{3.16}$$

where $Y_{i,t}$ denotes either the capital-labor ratio or the total number of medical facilities[9] 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,

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

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

$X_{i,t}$ 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.

3.5.2 Setup as a Continuous DiD

Notably, regression in (3.16) 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. (2024), with continuous treatment, the coefficient β in (3.16) can be expressed as a weighted average of the $ATT(d)$ overall the treatment intensities with potentially *negative* weights[10], 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 (3.16): Y can be either the capital-labor ratio or some measures of technological adoption.

[10] See Proposition 10 in Callaway et al. (2024).

- We assume a conditional parallel trends assumption:

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

  That is, on average, in the absence of the PPS reform, the outcome variables of hospitals with share $D = d$ should have experienced similar changes over time as hospitals with no Medicare patients (shares $D = 0$), *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[11]. 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 (3.16), 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 the 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

[11] 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 not to 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.

should experience a greater increase in the capital-labor ratio would hold if $ATT(d)$ increases 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 is 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 setting. Therefore, we demonstrate our methods in both cases, specifically[12]:

- 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 are the capital-labor ratio and a measure of technological adoption (number of medical facilities)[13].

- In addition, to allow for a direct comparison with Acemoglu and Finkelstein (2008) in the repeated outcomes setting, we also consider outcome variables Y_{t-1} averaged over the pre-treatment years (1980-1983) and Y_t averaged over the post-treatment years (1984-1986 for capital-labor ratio and 1984-1985 for tech adoption).

- 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.

[12] I would like to thank Kathleen McGarry, Daron Acemoglu, Amy Finkelstein, and the National Bureau of Economic Research for making it possible to access the data source in Acemoglu and Finkelstein (2008).

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

3.5.3 Results

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.

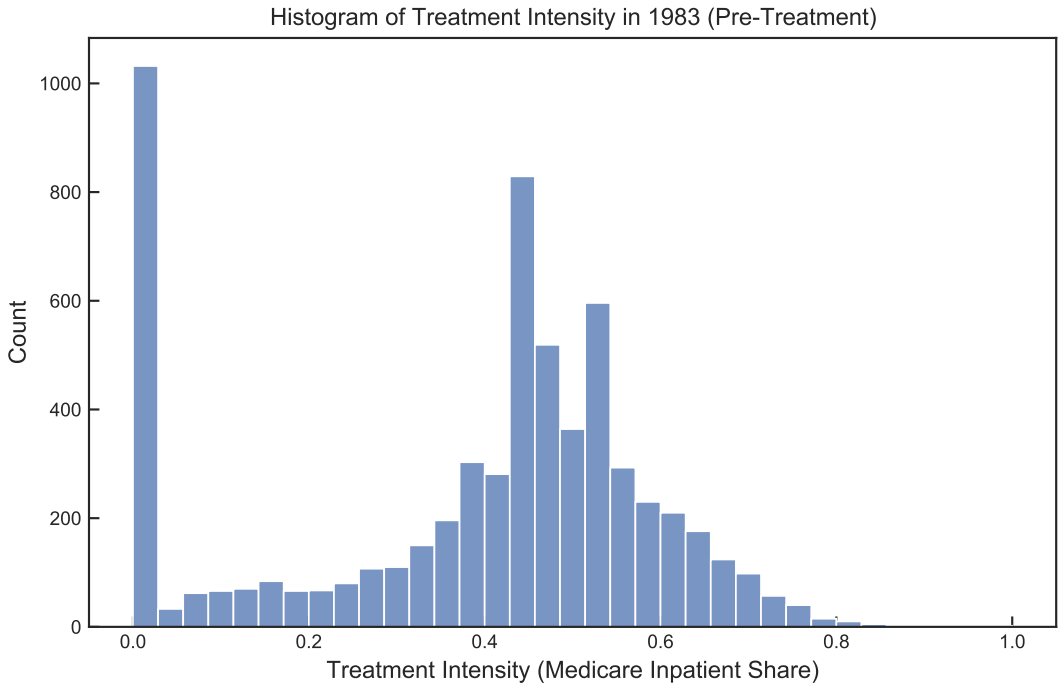

Figure 3.1: Histogram of Treatment Intensity (Medicare Share in 1983)

Figure 3.1 depicts the histogram of D for 1983, which suggests that this treatment variable is suitable for our continuous DiD framework. Specifically, we note that a significant number of hospitals register at $D = 0$, enabling us to consider these hospitals as the control group. Moreover, the positive Medicare shares ($D > 0$) vary 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 year, we estimate the causal parameter ATT(d) at various intensities d ranging from 0.1 to

0.8 for each $t = 1984, 1985, 1986$. The results are shown in figures 3.2, 3.3, 3.4 and Table 3.1. In the table, we provide standard errors in parentheses as well as bootstrap confidence intervals.

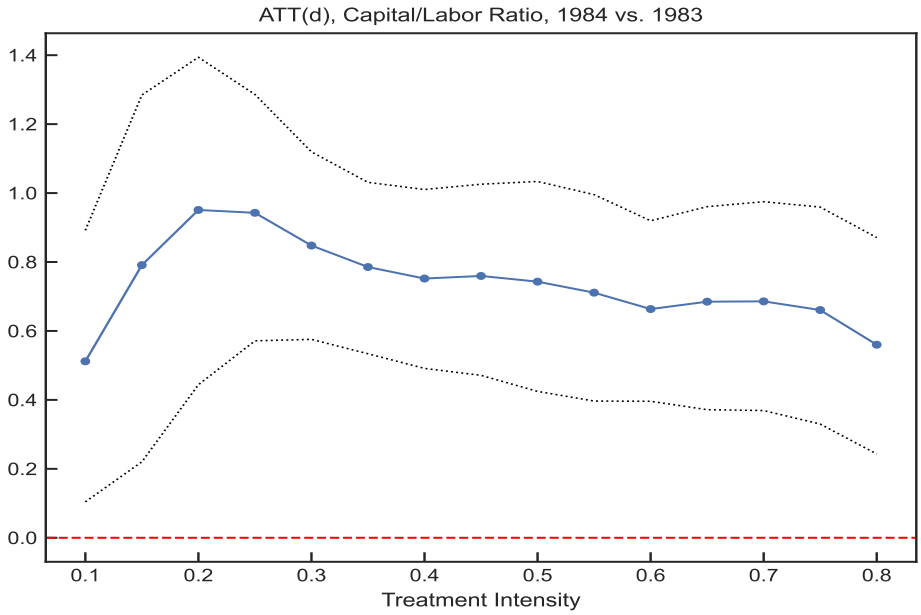

Figure 3.2: $\widehat{ATT}(d)$ for Capital-Labor Ratio (Panel Data), 1984 vs. 1983

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 $t = 1986$. Interestingly, 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 was incremental over time.

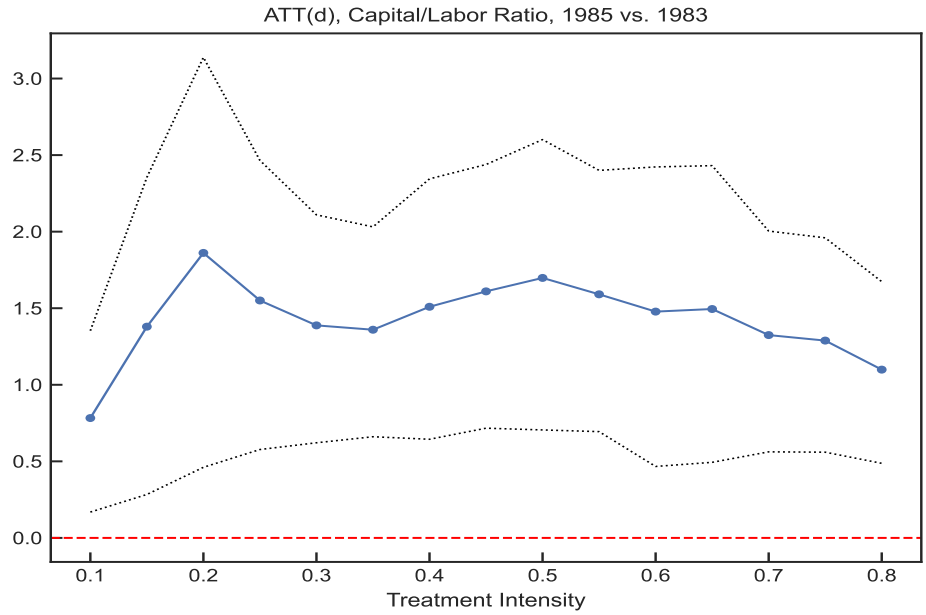

Figure 3.3: $\widehat{ATT}(d)$ for Capital-Labor Ratio (Panel Data), 1985 vs. 1983

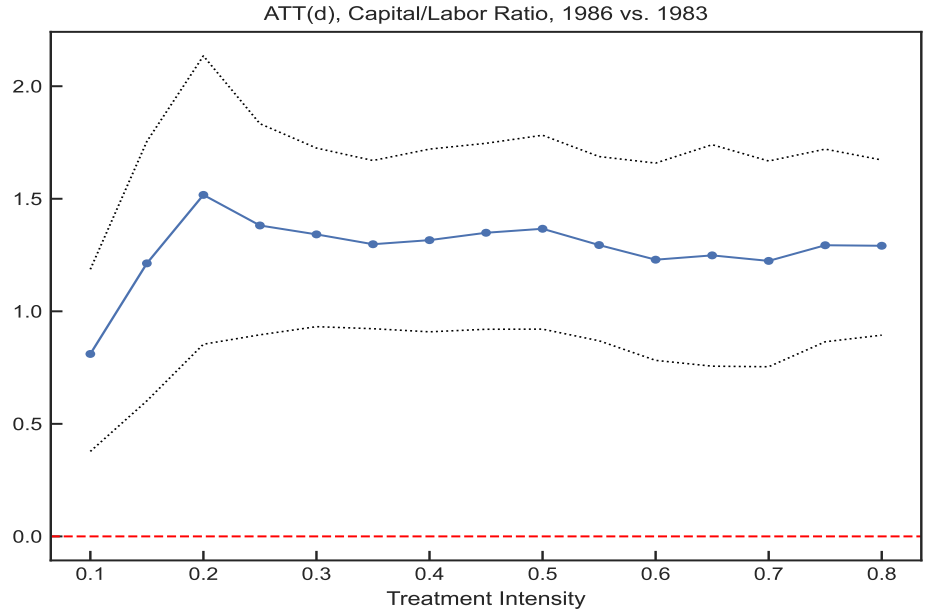

Figure 3.4: $\widehat{ATT}(d)$ for Capital-Labor Ratio (Panel Data), 1986 vs. 1983

Finally, for all three years, the estimated ATTs vary across treatment intensities and don't display increasing trends, which is inconsistent 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 a pattern. Notably, 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 3.2 and Figures 3.5 and 3.6.

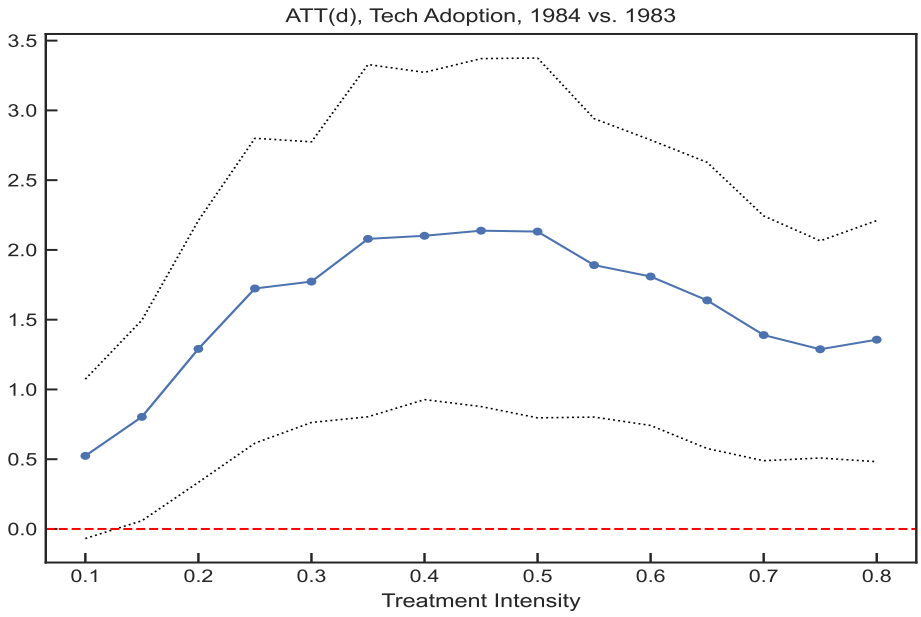

Figure 3.5: $\widehat{ATT}(d)$ for Tech Adoption (Panel Data), 1984 vs. 1983

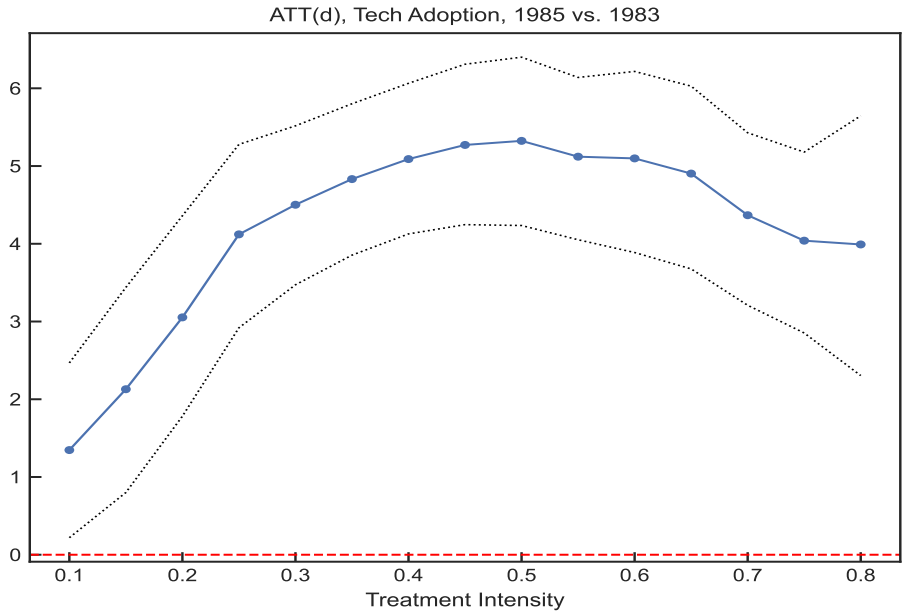

Figure 3.6: $\widehat{ATT}(d)$ for Tech Adoption (Panel Data), 1985 vs. 1983

Figures 3.5 and 3.6 further reveal 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 technological adoption. Moreover, similar to the findings for the capital-labor ratio, the 1985 estimates are much larger in magnitude compared to their 1984 counterparts, further suggesting that the impact of the PPS reform is staggered over time. Finally, for both years, the estimates are increasing for lower treatment intensities and decreasing for higher treatment intensities, which is especially evident for $t = 1986$. This pattern is inconsistent with the theoretical prediction that hospitals with higher Medicare inpatient shares should experience a bigger increase in technological adoption following the PPS reform [14].

Since Acemoglu and Finkelstein (2008) use all the available data between 1980 and 1986

[14]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.

in their linear specification, we can alternatively apply our methods to averaged outcomes over all the available periods pre and post-treatment. The results are presented in Table 3.4 and depicted in Figures 3.7 and 3.8 for the capital-labor ratio and technology adoption respectively. Notably, the estimated ATTs for the capital-labor ratio are consistently positive and display an upward trend. This aligns with the hypothesis that hospitals with a higher proportion of Medicare inpatients are likely to see more significant increases in their capital-labor ratios. However, the estimated ATTs for tech adoption, although large and positive, do not display an increasing trend.

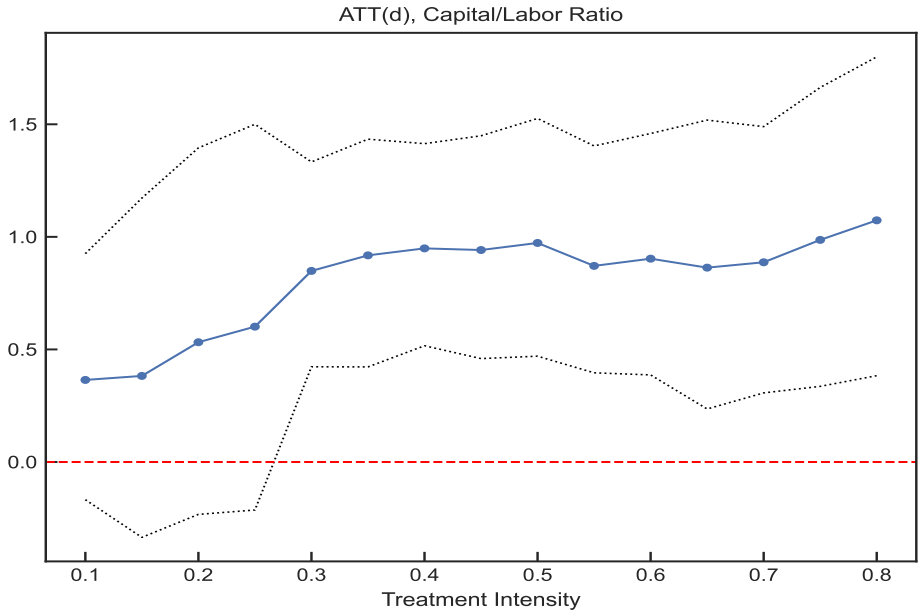

Figure 3.7: $\widehat{ATT}(d)$ for Capital-Labor Ratio (Panel Data), Average

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 acknowledge 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 1986, as displayed in Figure 3.9, reveals

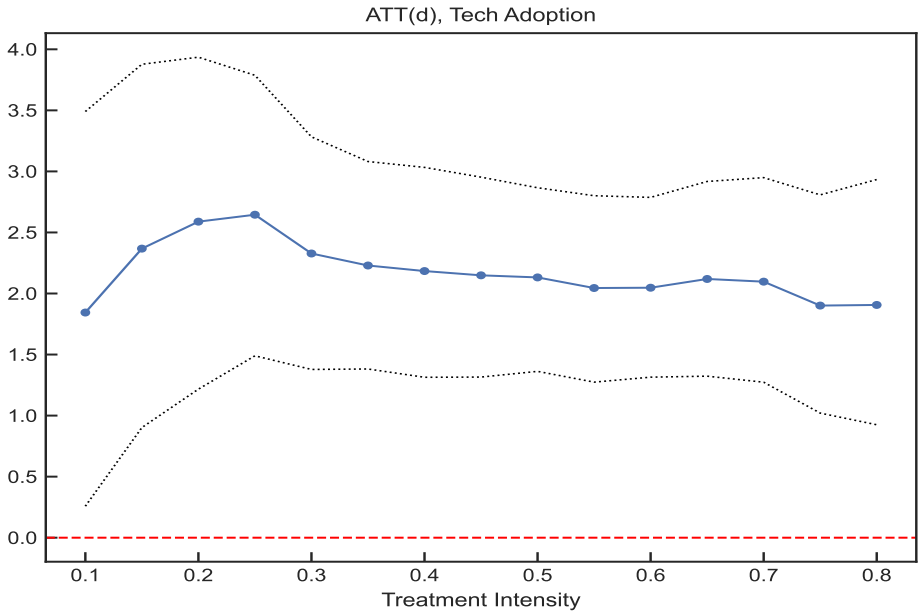

Figure 3.8: $\widehat{ATT}(d)$ for Tech Adoption (Panel Data), Average

a leftward shift in the distribution.

Therefore, to account for the changes in the treatment intensity (Medicare share), we can alternatively 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 $t = 1984, 1985, 1986$. The results from our repeated cross-sections methods, as shown in Table 3.3, differ considerably from those in the panel setting. As a highlight, we plot the results for 1986 in Figure 3.12.

Notably, most of the estimates for the year 1984 are not significantly different from zero. On the other hand, for the year 1985, 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 holds for the year 1986, as shown in Figure 3.12. This pattern markedly differs from what we see in the panel

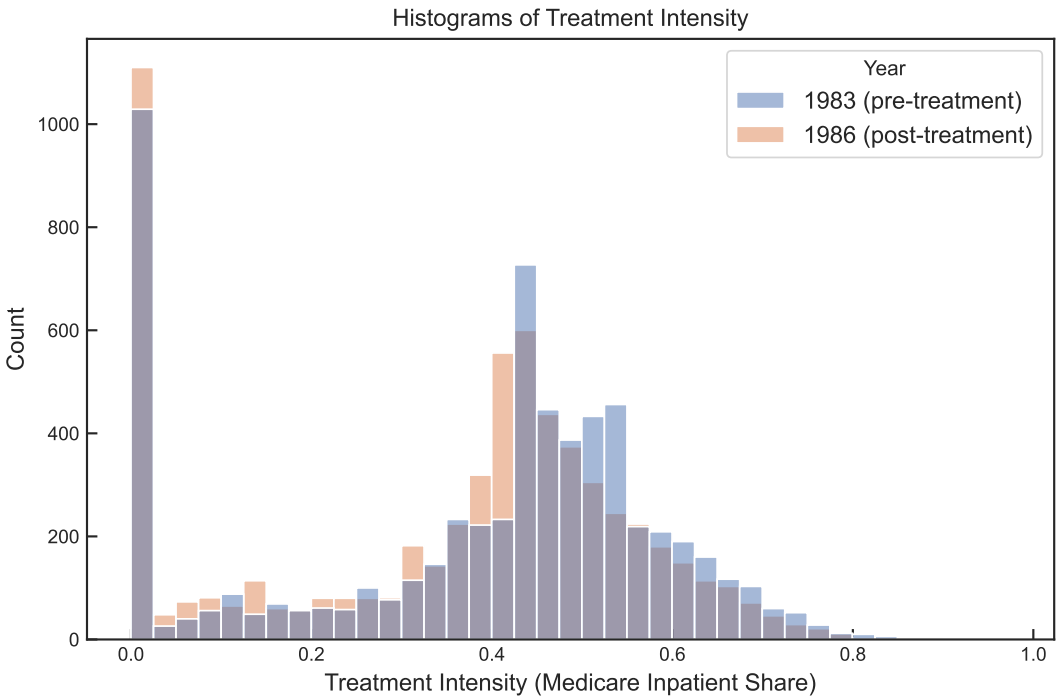

Figure 3.9: Histograms of Treatment Intensity (1983 vs. 1986)

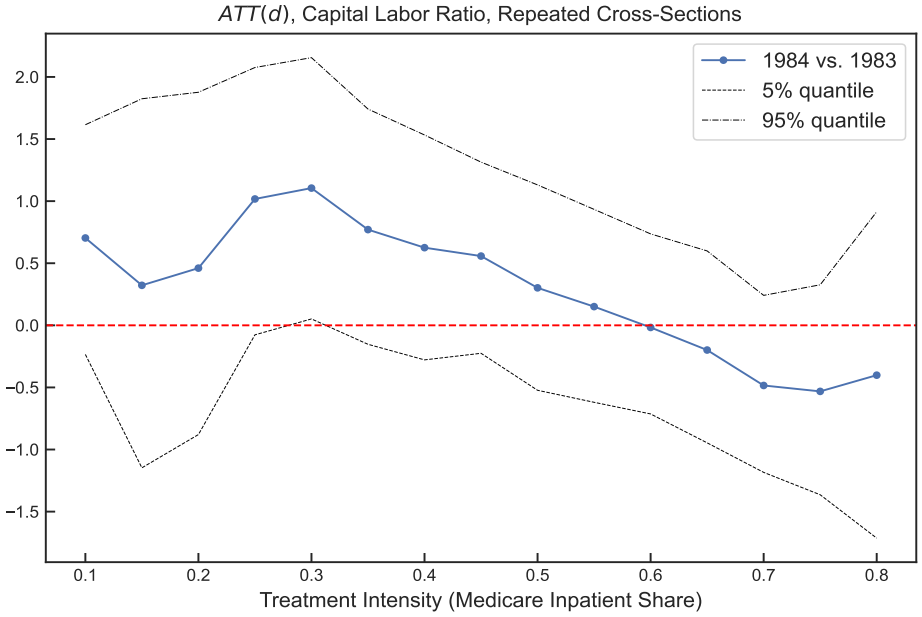

Figure 3.10: $\widehat{ATT}(d)$ for Capital-Labor Ratio (Repeated Cross-Sections), 1984 vs. 1983

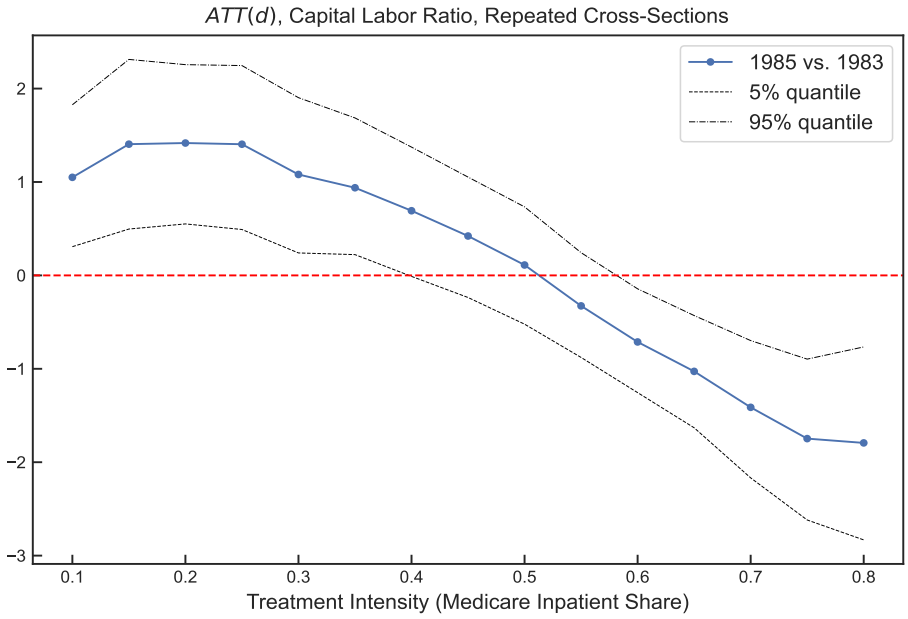

Figure 3.11: $\widehat{ATT}(d)$ for Capital-Labor Ratio (Repeated Cross-Sections), 1985 vs. 1983

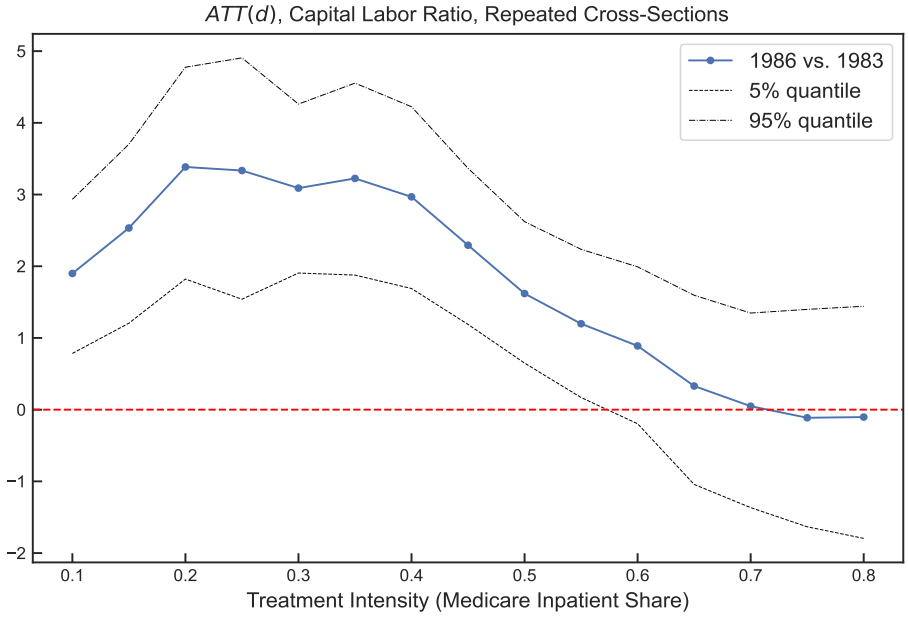

Figure 3.12: $\widehat{ATT}(d)$ for Capital-Labor Ratio (Repeated Cross-Sections), 1986 vs. 1983

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 a 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 3.5.1. 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[15].
- The second-order Gaussian kernel with bandwidth $h = O(N^{-1/4})$ is used to construct the estimator and to estimate the density $f_D(d)$ and the conditional mean $E[K_h(D - d)|X]$.
- The infinite-dimensional nuisance parameters are estimated using the Random Forest (RF). We use the scikit-learn RF packages from Python with default settings. 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 standard errors are calculated using the cross-fitted estimator defined in (3.12) and (3.13). In addition, we also present 90-percent bootstrap confidence intervals constructed using the multiplier bootstrap procedure defined in (3.14) and (3.15): Gaus-

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

sian multipliers $\{\xi_i\}_{i=1}^N$ are drawn from a normal distribution with $E[\xi_i] = Var[\xi_i] = 1$ for $B = 1000$ repetitions.

3.6 Conclusion

This paper studies difference-in-differences models with continuous treatments. Our identification results are based on a conditional parallel trends assumption, allowing researchers to control for a rich set of covariates. Under the double/debiased machine learning framework, we develop estimators for the causal parameters and establish their asymptotic properties. To illustrate the practical application of our methodologies, we revisit the research questions posed in Acemoglu and Finkelstein (2008), applying our estimators to their dataset and deriving new research insights. The extension of difference-in-differences models to the continuous treatment setting has important implications in empirical research. Our methods provide researchers with new tools for examining the impacts of continuous treatment variables.

Table 3.1: 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	0.5119 (0.2263)	[0.1043, 0.8920]	0.7827 (0.3702)	[0.1687, 1.3531]	0.8105 (0.2517)	[0.3776, 1.1868]
d = 0.15	0.7911 (0.3090)	[0.2204, 1.2845]	1.3795 (0.6281)	[0.2837, 2.3546]	1.2130 (0.3592)	[0.6022, 1.7548]
d = 0.2	0.9510 (0.2812)	[0.4440, 1.3939]	1.8611 (0.8210)	[0.4610, 3.1380]	1.5172 (0.3919)	[0.8537, 2.1348]
d = 0.25	0.9427 (0.2110)	[0.5713, 1.2862]	1.5505 (0.5833)	[0.5772, 2.4654]	1.3813 (0.2912)	[0.8956, 1.8335]
d = 0.3	0.8478 (0.1668)	[0.5756, 1.1195]	1.3880 (0.4581)	[0.6213, 2.1097]	1.3417 (0.2414)	[0.9318, 1.7253]
d = 0.35	0.7855 (0.1536)	[0.5338, 1.0310]	1.3595 (0.4211)	[0.6607, 2.0312]	1.2980 (0.2282)	[0.9227, 1.6700]
d = 0.4	0.7520 (0.1573)	[0.4913, 1.0103]	1.5094 (0.5170)	[0.6442, 2.3448]	1.3162 (0.2501)	[0.9089, 1.7206]
d = 0.45	0.7596 (0.1685)	[0.4712, 1.0257]	1.6101 (0.5286)	[0.7165, 2.4385]	1.3490 (0.2501)	[0.9202, 1.7464]
d = 0.5	0.7430 (0.1828)	[0.4247, 1.0336]	1.6979 (0.5826)	[0.7051, 2.6013]	1.3667 (0.2628)	[0.9208, 1.7823]
d = 0.55	0.7111 (0.1828)	[0.3966, 0.9955]	1.5908 (0.5283)	[0.6940, 2.4002]	1.2943 (0.2532)	[0.8691, 1.6875]
d = 0.6	0.6635 (0.1649)	[0.3960, 0.9196]	1.4779 (0.5839)	[0.4667, 2.4228]	1.2293 (0.2724)	[0.7820, 1.6585]
d = 0.65	0.6849 (0.1826)	[0.3716, 0.9606]	1.4948 (0.5979)	[0.4936, 2.4310]	1.2485 (0.3054)	[0.7562, 1.7402]
d = 0.7	0.6858 (0.1895)	[0.3691, 0.9749]	1.3246 (0.4452)	[0.5620, 2.0039]	1.2239 (0.2876)	[0.7537, 1.6677]
d = 0.75	0.6607 (0.1969)	[0.3299, 0.9594]	1.2886 (0.4359)	[0.5598, 1.9597]	1.2935 (0.2654)	[0.8646, 1.7206]
d = 0.8	0.5601 (0.1994)	[0.2432, 0.8709]	1.0990 (0.3710)	[0.4870, 1.6736]	1.2912 (0.2359)	[0.8940, 1.6723]

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 3.2: Estimated ATT(d) for Technological Adoption (Panel)

	$(t = 1984)$ ATT(D=d)	Bootstrap CI	$(t = 1985)$ ATT(D=d)	Bootstrap CI
d = 0.1	0.5240 (0.3541)	[-0.0683, 1.0739]	1.3464 (0.6621)	[0.2192, 2.4675]
d = 0.15	0.8035 (0.4492)	[0.0589, 1.4966]	2.1292 (0.7826)	[0.8004, 3.4388]
d = 0.2	0.9510 (0.2812)	[0.4440, 1.3939]	3.0533 (0.7594)	[1.7817, 4.3604]
d = 0.25	0.9427 (0.2110)	[0.5713, 1.2862]	4.1212 (0.7150)	[2.9192, 5.2785]
d = 0.3	0.8478 (0.1668)	[0.5756, 1.1195]	4.5021 (0.6252)	[3.4722, 5.5170]
d = 0.35	0.7855 (0.1536)	[0.5338, 1.0310]	4.8321 (0.6251)	[3.8539, 5.8011]
d = 0.4	0.7520 (0.1573)	[0.4913, 1.0103]	5.0896 (0.6054)	[4.1269, 6.0641]
d = 0.45	0.7596 (0.1685)	[0.4712, 1.0257]	5.2714 (0.6439)	[4.2462, 6.3094]
d = 0.5	0.7430 (0.1828)	[0.4247, 1.0336]	5.3237 (0.6799)	[4.2349, 6.4002]
d = 0.55	0.7111 (0.1828)	[0.3966, 0.9955]	5.1203 (0.6437)	[4.0518, 6.1395]
d = 0.6	0.6635 (0.1649)	[0.3960, 0.9196]	5.0983 (0.7095)	[3.8875, 6.2169]
d = 0.65	0.6849 (0.1826)	[0.3716, 0.9606]	4.9029 (0.7314)	[3.6766, 6.0275]
d = 0.7	0.6858 (0.1895)	[0.3691, 0.9749]	4.3674 (0.6848)	[3.2089, 5.4279]
d = 0.75	0.6607 (0.1969)	[0.3299, 0.9594]	4.0398 (0.7158)	[2.8524, 5.1793]
d = 0.8	0.5601 (0.1994)	[0.2432, 0.8709]	3.9913 (0.9975)	[2.3046, 5.6463]

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 3.3: 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	0.7033 (0.5416)	[-0.2344, 1.6152]	1.0499 (0.4538)	[0.3075, 1.8266]	1.8990 (0.6866)	[0.7839, 2.9331]
d = 0.15	0.3221 (0.8935)	[-1.1473, 1.8242]	1.4049 (0.5416)	[0.4962, 2.3123]	2.5333 (0.7822)	[1.2058, 3.7020]
d = 0.2	0.4605 (0.8390)	[-0.8795, 1.8767]	1.4170 (0.5105)	[0.5512, 2.2562]	3.3844 (0.9248)	[1.8212, 4.7757]
d = 0.25	1.0175 (0.6380)	[-0.0765, 2.0759]	1.4043 (0.5174)	[0.4914, 2.2464]	3.3348 (1.0201)	[1.5400, 4.9057]
d = 0.3	1.1050 (0.6381)	[0.0515, 2.1554]	1.0803 (0.4943)	[0.2404, 1.9028]	3.0897 (0.7121)	[1.9054, 4.2620]
d = 0.35	0.7707 (0.5560)	[-0.1527, 1.7415]	0.9383 (0.4464)	[0.2224, 1.6858]	3.2249 (0.8235)	[1.8767, 4.5541]
d = 0.4	0.6253 (0.5413)	[-0.2782, 1.5327]	0.6922 (0.4144)	[-0.0106, 1.3734]	2.9678 (0.7748)	[1.6895, 4.2243]
d = 0.45	0.5577 (0.4582)	[-0.2252, 1.3142]	0.4216 (0.3859)	[-0.2366, 1.0526]	2.2934 (0.6716)	[1.1890, 3.3662]
d = 0.5	0.3019 (0.4932)	[-0.5230, 1.1306]	0.1114 (0.3800)	[-0.5221, 0.7335]	1.6193 (0.5831)	[0.6496, 2.6220]
d = 0.55	0.1507 (0.4700)	[-0.6197, 0.9337]	-0.3263 (0.3376)	[-0.8784, 0.2433]	1.1984 (0.6158)	[0.1696, 2.2362]
d = 0.6	-0.0171 (0.4502)	[-0.7135, 0.7363]	-0.7125 (0.3392)	[-1.2549, -0.1437]	0.8891 (0.6643)	[-0.1967, 1.9915]
d = 0.65	-0.1989 (0.4806)	[-0.9461, 0.5991]	-1.0275 (0.3680)	[-1.6313, -0.4297]	0.3298 (0.7843)	[-1.0422, 1.5955]
d = 0.7	-0.4843 (0.4350)	[-1.1839, 0.2412]	-1.4125 (0.4569)	[-2.1677, -0.6978]	0.0484 (0.8190)	[-1.3661, 1.3474]
d = 0.75	-0.5318 (0.5186)	[-1.3642, 0.3263]	-1.7474 (0.5290)	[-2.6185, -0.8964]	-0.1131 (0.8910)	[-1.6335, 1.3985]
d = 0.8	-0.4013 (0.8252)	[-1.7134, 0.9144]	-1.7933 (0.6282)	[-2.8317, -0.7658]	-0.1034 (0.9394)	[-1.7952, 1.4421]

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 3.4: Estimated ATT(d) Average (Panel)

	(capital-labor ratio) ATT(D=d)	Bootstrap CI	(tech adoption) ATT(D=d)	Bootstrap CI
d = 0.1	0.3645 (0.2847)	[-0.1674, 0.9254]	1.8441 (0.9930)	[0.2580, 3.4905]
d = 0.15	0.3823 (0.3948)	[-0.3348, 1.1722]	2.3676 (0.8988)	[0.9020, 3.8775]
d = 0.2	0.5320 (0.4316)	[-0.2327, 1.3953]	2.5885 (0.8456)	[1.2165, 3.9355]
d = 0.25	0.6012 (0.4526)	[-0.2132, 1.5000]	2.6452 (0.7149)	[1.4897, 3.7872]
d = 0.3	0.8487 (0.2472)	[0.4229, 1.3328]	2.3276 (0.5846)	[1.3780, 3.2829]
d = 0.35	0.9178 (0.2685)	[0.4224, 1.4339]	2.2293 (0.5170)	[1.3818, 3.0810]
d = 0.4	0.9488 (0.2353)	[0.5163, 1.4140]	2.1838 (0.5010)	[1.3138, 3.0329]
d = 0.45	0.9415 (0.2606)	[0.4593, 1.4489]	2.1489 (0.4807)	[1.3156, 2.9529]
d = 0.5	0.9729 (0.2729)	[0.4700, 1.5259]	2.1317 (0.4349)	[1.3623, 2.8661]
d = 0.55	0.8710 (0.2630)	[0.3962, 1.4039]	2.0449 (0.4414)	[1.2741, 2.8009]
d = 0.6	0.9030 (0.2751)	[0.3868, 1.4591]	2.0473 (0.4365)	[1.3147, 2.7872]
d = 0.65	0.8633 (0.3224)	[0.2355, 1.5186]	2.1190 (0.4675)	[1.3230, 2.9176]
d = 0.7	0.8871 (0.3004)	[0.3069, 1.4896]	2.0970 (0.4874)	[1.2735, 2.9491]
d = 0.75	0.9864 (0.3206)	[0.3360, 1.6636]	1.9011 (0.5301)	[1.0205, 2.8078]
d = 0.8	1.0734 (0.3493)	[0.3834, 1.7995]	1.9062 (0.5947)	[0.9250, 2.9333]

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 capital-labor ratio, the pre-treatment period outcomes are averaged over year $1980, 1981, 1982, 1983$, and post-treatment period outcomes are averaged over year $1984, 1985, 1986$;(v)for tech adoption, the pre-treatment period outcomes are averaged over year $1980, 1981, 1982, 1983$, and post-treatment period outcomes are averaged over year $1984, 1985$; (vi) all the nuisance parameters are estimated nonparametrically using random forests.

3.7 Proofs

3.7.1 Proof of Theorem 3.2.1

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,

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

where the first equality holds by the law of total probability, the second equality holds by the 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 the above results, we have

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

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

Next, for repeated cross-sections, we have

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

where the first equality holds by law of iterated expectation, the second equality holds by definition, and the last two equalities hold by assumption 3.2.2.

3.7.2 Proof of Lemma 3.3.1

First, consider the repeated outcomes case. Define the unadjusted score φ_h as

$$
\varphi_h := \Delta Y \frac{K_h(D-d)g_0(X) - \mathbf{1}\{D=0\}f_h^0(d|X)}{f_d^0 g_0(X)} - ATT_h(d)
$$

where we use the following notation: $\Delta Y = Y_t - Y_{t-1}$, $f_d^0 := f_D(d)$, $f_h^0(d|X) := E[K_h(D-d)|X]$, $g_0(X) := P(D=0|X)$. 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.

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

$$c_h := (K_h(D-d) - f_h^0(d|X))E[\partial_1\varphi_h|X] + (\mathbf{1}\{D=0\} - g_0(X))E[\partial_2\varphi_h|X]$$

where ∂_1 and ∂_2 denote the partial derivatives with respect to $f_h^0(d|X)$ and $g_0(X)$ respectively. Then, we have

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

where $\mathcal{E}_{\Delta Y}^0(X) = E[\Delta Y \mathbf{1}\{D=0\}|X]/g_0(X) = E[\Delta Y|D=0,X]$. In particular, note that ψ_h in the lemma satisfies $\psi_h = \varphi_h + c_h$.

Now it remains to show the new score ψ_h satisfies Neyman orthogonality w.r.t. the nuisance parameters, $f_h^0(d|X)$, $g_0(X)$, and $\mathcal{E}_{\Delta Y}^0(X)$. First, we need to check the moment condition $E[\psi_h] = 0$. Since $E[\varphi_h] = 0$, we only need to check $E[c_h] = 0$:

$$\begin{aligned}
E[c_h] &= E\left[\frac{\mathbf{1}\{D=0\}f_h^0(d|X) - K_h(D-d)g_0(X)}{f_d^0 \cdot g_0(X)}\mathcal{E}_{\Delta Y}^0(X)\right] \\
&= E\left[\frac{E[\mathbf{1}\{D=0\}|X]f_h^0(d|X) - E[K_h(D-d)|X]g_0(X)}{f_d^0 \cdot g_0(X)}\mathcal{E}_{\Delta Y}^0(X)\right] \\
&= E\left[\frac{g_0(X)f_h^0(d|X) - f_h^0(d|X)g_0(X)}{f_d^0 \cdot g_0(X)}\mathcal{E}_{\Delta Y}^0(X)\right] \\
&= 0
\end{aligned}$$

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

Second, we need to show the Gateaux derivative of the score w.r.t. the nuisance parameters $\eta_0 := (f_h^0(d|X), g_0(X), \mathcal{E}_{\Delta Y}^0(X))$ vanishes at zero, that is, we need to show

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

We use the notation η without the subscript 0 to denote generic nuisance parameters in the set $\mathcal{T}_n$. 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 $f_h(d|X)$:

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

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

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

$$\partial_r E[\psi_h(g_0(X) + r(g(X) - g_0(X)))]|_{r=0}$$

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

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

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

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

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

This shows that the score ψ_h is Neyman orthogonal w.r.t. the infinite-dimensional nuisance parameters. The proof for the repeated cross-sections case follows the same argument by replacing ΔY with $\frac{T-\lambda}{\lambda(1-\lambda)}Y$. □

3.7.3 Proof of Lemma 3.4.1

We focus on the repeated outcomes case. The bias $B_h(d)$ is defined as

$$\begin{aligned}
B_h(d) :=& ATT(d) - ATT_h(d) \\
=& E[\Delta Y|D=d] - E\left[\Delta Y \mathbf{1}\{D=0\}\frac{f_{D|X}(d|X)}{f_D(d)P(D=0|X)}\right] \\
& - E\left[\Delta Y \frac{K_h(D-d)P(D=0|X) - \mathbf{1}\{D=0\}E[K_h(D-d)|X]}{f_D(d)P(D=0|X)}\right] \\
=& \left(E[\Delta Y|D=d] - E\left[\Delta Y \frac{K_h(D-d)}{f_D(d)}\right]\right) \\
& - E\left[\Delta Y \mathbf{1}\{D=0\}\frac{f_{D|X}(d|X) - E[K_h(D-d)|X]}{f_D(d)P(D=0|X)}\right].
\end{aligned}$$

First, note that

$$\begin{aligned}
& E[\Delta Y|D=d] - E\left[\Delta Y \frac{K_h(D-d)}{f_D(d)}\right] \\
=& \int t \frac{f_{\Delta Y,D}(t,d)}{f_D(d)} dt - \int t \frac{1}{f_D(d)} \int \frac{1}{h} K\left(\frac{s-d}{h}\right) f_{\Delta Y,D}(t,s) ds dt \\
=& \int C_1 \frac{t}{f_D(d)} h^2 f^{(2)}_{\Delta Y,D}(t,d) dt + o(h^2) \\
=& O(h^2)
\end{aligned}$$

where the first equality holds by definition, the second equality holds by change of variables and Taylor expansion (see Lemma 5.1 in Fan and Yao (2003)), and the last equality holds by assumption.

Second, by the same argument using the change of variables and Taylor expansion, we have

$$\begin{aligned}
E[K_h(D-d)|X=x] &= \int \frac{1}{h} K\left(\frac{d-s}{h}\right) f_{D|X}(s|x) ds \\
&= \int K(u) f_{D|X}(d+hu|x) du
\end{aligned}$$

$$= f_{D|X}(d|x) + C_2 h^2 f^{(2)}_{D|X}(d|x) + o(h^2).$$

Then by the uniform boundedness of $f^{(2)}_{D|X}(d|x)$ and assumptions on $\Delta Y, f_D(d), P(D=0|X)$, applying the dominated convergence theorem, we have

$$\begin{aligned}
&E\left[\Delta Y \mathbf{1}\{D=0\}\frac{f_{D|X}(d|X) - E[K_h(D-d)|X]}{f_D(d)P(D=0|X)}\right] \\
=&C_3 h^2 E\left[\Delta Y \mathbf{1}\{D=0\}\frac{f^{(2)}_{D|X}(d|X)}{f_D(d)P(D=0|X)}\right] + o(h^2) \\
=&O(h^2).
\end{aligned}$$

Combining the two results, we have $B_h(d) = O(h^2)$, which completes the proof. The proof for the repeated cross-sections case follows the same argument by replacing ΔY with $\frac{T-\lambda}{\lambda(1-\lambda)}Y$. □

3.7.4 Proof of Theorem 3.4.1 (Repeated Outcomes)

Let T_N be the set of square integrable nuisance parameters $\eta := (f_h(d|X), g(X), \mathcal{E}_{\Delta Y}(X))$ such that assumption 3.4.3 holds. Let F_N be the set of $f > 0$ such that $|f - f^0_d| \leq (Nh)^{-1/2}$. Then assumption 3.4.3 implies that, with probability tending to 1, $\hat{\eta}_k \in T_N$ and $\hat{f}_{d,k} \in F_N$. Throughout the proof, we use N to denote the sample size and $n := N/K$ to denote the size of any of the subsamples. In particular, since K is fixed, $n \asymp N$.

To simplify notation, let θ_0 denote the true $ATT(d)$, θ_{0h} denote the true $ATT_h(d)$, and $\hat{\theta}_h$ denote our cross-fitted estimator. In particular, recall that our estimator is

$$\hat{\theta}_h := \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i\in I_k}\frac{K_h(D_i-d)\hat{g}_k(X_i) - \mathbf{1}\{D_i=0\}\hat{f}_{h,k}(d|X_i)}{\hat{f}_{d,k}\hat{g}_k(X_i)}\left(\Delta Y_i - \hat{\mathcal{E}}_{\Delta Y,k}(X_i)\right).$$

Then we can decompose the following difference as

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

where (†) will be our main focus while the bias term (††) is shown in Lemma 3.4.1 to be $O(h^2)$ and asymptotically negligible by the assumption of the under-smoothing bandwidth.

By definition,

$$\sqrt{N}(\hat{\theta}_h - \theta_{0h}) = \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi_h(Z_i, \theta_{0,h}, \hat{f}_{d,k}, \hat{\eta}_k)] \tag{3.17}$$

where ψ_h is defined as in (3.8), and $E_{n,k}(f) = \frac{1}{n}\sum_{i\in I_k} f(Z_i)$ denotes the empirical average of a generic function f over the set I_k. Then we have the following decomposition, using Taylor's theorem:

$$\sqrt{N}(\hat{\theta}_h - \theta_{0h}) = \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi_h(Z, \theta_{0h}, f_d^0, \hat{\eta}_k)] \tag{3.18}$$

$$+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\partial_f \psi_h(Z, \theta_{0h}, f_d^0, \hat{\eta}_k)](\hat{f}_{d,k} - f_d^0) \tag{3.19}$$

$$+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\partial_f^2 \psi_h(Z, \theta_{0h}, \bar{f}_k, \hat{\eta}_k)](\hat{f}_{d,k} - f_d^0)^2 \tag{3.20}$$

where $\bar{f}_k \in (f_d^0, \hat{f}_{d,k})$. This decomposition provides a roadmap for the remainder of the proof. There are roughly four steps. In the first step, we show the second-order term (3.20) vanishes rapidly and does not contribute to the asymptotic variance. In the second step, we bound the first-order term (3.19), which potentially contributes to the asymptotic variance. In step 3, we expand (3.18) 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 under our assumptions. In the final step, we verify the results used in the first two steps and conclude.

Before we start the main proof, we state two well-known results that will be used in the proof. For an i.i.d. sample $\{D_i\}_{i=1}^n$, the kernel estimator for the density $f_D(d) := f_d^0$ in our setting is defined as

$$\hat{f}_d := \frac{1}{n}\sum_{i=1}^{n} K_h(D_i - d).$$

Then,

$$\hat{f}_d - f_d^0 = \hat{f}_d - E[K_h(D-d)] - (f_d^0 - E[K_h(D-d)]).$$

One can show that (see for example, Härdle (1990))

$$\begin{aligned}\hat{f}_d - E[K_h(D-d)] &= O_p((nh)^{-1/2}) \\ f_d^0 - E[K_h(D-d)] &= O(h^2).\end{aligned}$$

Therefore, for an under-smoothing $h = o(n^{-1/5})$, we have $\hat{f}_d - f_d^0 = O_p((nh)^{-1/2})$ and $(\hat{f}_d - f_d^0)^2 = O_p((nh)^{-1})$.

Step 1: Second Order Terms

First, we consider (3.20). By triangle inequality, we have

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

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

M_h,

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

which implies that

$$\begin{aligned}
E[J_{2k}^2] &= E\left[\left(\frac{1}{n}\sum_{i\in I_k} \partial_f^2 \psi_h(Z, \theta_{0J}, f_d^0, \eta_0) - E[\partial_f^2 \psi_h(Z, \theta_{0h}, f_d^0, \eta_0)]\right)^2\right] \\
&= E\left[\left(\frac{1}{n}\sum_{i\in I_k} \partial_f^2 \psi_h(Z, \theta_{0h}, f_d^0, \eta_0)\right)^2\right] - \left(E[\partial_f^2 \psi_h(Z, \theta_{0h}, f_d^0, \eta_0)]\right)^2 \\
&\leq \frac{1}{n} E[(\partial_f^2 \psi_h(Z, \theta_{0h}, f_d^0, \eta_0))^2] \\
&\lesssim E[K_h^2(D-d)]/N \\
&\lesssim (hN)^{-1},
\end{aligned}$$

where the third line holds by Cauchy-Schwarz inequality and Jensen's inequality, the fourth line holds by the boundedness assumption on the components of the score, and the last line holds by the assumption on the kernel function K. Then by the Markov's inequality, we have $J_{2k} \leq O_p\left((Nh)^{-1/2}\right)$.

Next, for J_{1k}, we have

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

where the second line holds by Cauchy-Schwarz inequality and the definition of supremum over the sets F_N and T_N, and the third line holds since the supremum does not depend on the sample I_k^c. Then by conditional Markov's inequality, $J_{1k} \leq O_p(h^{-1/2}\varepsilon_N)$. Using the

previous result that $(\hat{f}_{d,k} - f_d^0)^2 = O_p((Nh)^{-1})$, we conclude that (3.20) $= o_p(1)$. We will verify (a) at the end of this section.

Step 2: First-Order Terms

To bound (3.19), we first use the triangle inequality to obtain the decomposition

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

We first bound J_{4k}: By definition, we have

$$\partial_f \psi_h(Z, \theta_{0h}, f_d^0, \eta_0) = -\frac{1}{f_d^0}(\psi_h(Z, \theta_{0h}, f_d^0, \eta_0) + \theta_{0h}).$$

By the boundedness assumption,

$$E[J_{4k}^2] \leq \frac{1}{N} E[(\partial_f \psi_h(Z, \theta_{0h}, f_d^0, \eta_0))^2] \lesssim (Nh)^{-1}.$$

Then by Markov's inequality, we have $J_{4k} \leq O_p((Nh)^{-1/2})$. With the assumption that $Nh \to \infty$, we have $J_{4k} = o_p(1)$.

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

$$\begin{aligned}
E[J_{3k}^2 | I_k^c] &= E[|E_{n,k}[\partial_f \psi_h(Z, \theta_{0h}, f_d^0, \hat{\eta}_k)] - E_{n,k}[\partial_f \psi_h(Z, \theta_{0h}, f_d^0, \eta_0)]|^2 | I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_f \psi_h(Z, \theta_{0h}, f_d^0, \eta) - \partial_f \psi_h(Z, \theta_{0h}, f_d^0, \eta_0)|^2 | I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_f \psi_h(Z, \theta_{0h}, f_d^0, \eta) - \partial_f \psi_h(Z, \theta_{0h}, f_d^0, \eta_0)|^2] \\
&\lesssim h^{-1} \varepsilon_N^2 \qquad \text{(b)}
\end{aligned}$$

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 $J_{3k} = o_p(1)$ provided that $h^{-1}\varepsilon_N^2 = o(1)$, which is satisfied for an under-smoothing bandwidth h already assumed for valid inference. We will show (b) at the end of this section. Therefore,

$$E_{n,k}[\partial_f \psi_h(Z, \theta_{0h}, f_d^0, \hat{\eta}_k)] = \underbrace{E[\partial_f \psi_h(Z, \theta_{0h}, f_d^0, \eta_0)]}_{:=S_f^0} + o_p(1).$$

Note that the kernel density estimator satisfies $(\hat{f}_{d,k} - f_d^0) = O_p((Nh)^{-1/2})$, so we can rewrite (3.19) as

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

where the last equality holds by the definition that

$$\hat{f}_{d,k} - f_d^0 = (N - n)^{-1}\sum_{i \in I_k^c} K_h(D_i - d) - E[K_h(D - d)] + O(h^2)$$

with $N - n$ the sample size of each auxiliary subsample used to estimate the nuisance parameters, h being an under-smoothing bandwidth, 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 (3.18), which we can rewrite as

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

Since K is fixed, $n = O(N)$, it suffices to show that $R_{nk} = o_p(N^{-1/2}h^{-1})$, 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

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

with $G_{nk}(f) = \sqrt{n}(P_n - P)(f)$ denote the empirical process, i.e., $G_{nk}(f) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E[f(Z)]$, 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,

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

First, we consider R_{1k}. For simplicity, let's suppress other arguments in ψ and denote $\psi_\eta^i := \psi_h(Z_i,\theta_{0h},f_d^0,\eta)$. Then, by the definition of the empirical process, we have

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

In particular, it can be shown that $E[\Delta_{ik}\Delta_{jk}|I_k^c] = 0$ for all $i \neq j$ using the law of iterated expectation, the i.i.d. assumption of the data, and the fact that the nuisance parameter $\hat{\eta}_k$ is estimated using the auxiliary sample I_k^c. Then, we have

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

and using the conditional Markov's inequality, we conclude that $R_{1k} = O_p(h^{-1/2}\varepsilon_N)$.

Now we bound R_{2k}. Note that by definition of the score, $E[\psi_h(Z, \theta_{0h}, f_d^0, \eta_0)] = 0$, so it suffices to bound $E[\psi_h(Z, \theta_{0h}, f_d^0, \hat{\eta}_k)|I_k^c]$. Suppressing other arguments in the score, define

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

where by definition $h_k(0) = E[\psi_h(\eta_0)|I_k^c] = 0$ and $h_k(1) = E[\psi_h(\hat{\eta}_k)|I_k^c]$. 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,

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

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

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

Combining the above results, we conclude that

$$\sqrt{N}R_{n,k} \lesssim h^{-1/2}\varepsilon_N + \sqrt{N}h^{-1/2}\varepsilon_N^2,$$

and for $\varepsilon_N = o(N^{-1/4})$, we have $\sqrt{N}R_{n,k} = o_p(h^{-1/2})$.

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

$$(c): \qquad \sup_{\eta\in T_N} E[(\psi_\eta - \psi_{\eta_0})^2] \lesssim h^{-1}\varepsilon_N^2.$$

By definition,

$$\begin{aligned} \psi_\eta - \psi_{\eta_0} =& \frac{K_h(D-d)g(X) - \mathbf{1}\{D=0\}f_h(d|X)}{f_d^0 g(X)}\left(\Delta Y - \mathcal{E}_{\Delta Y}(X)\right) \\ &- \frac{K_h(D-d)g_0(X) - \mathbf{1}\{D=0\}f_h^0(d|X)}{f_d^0 g_0(X)}\left(\Delta Y - \mathcal{E}_{\Delta Y}^0(X)\right) \\ =& \frac{K_h(D-d)}{f_d^0}(\mathcal{E}_{\Delta Y}^0(X) - \mathcal{E}_{\Delta Y}(X)) \\ &- \frac{\mathbf{1}\{D=0\}}{f_d^0}\left(\frac{f_h(d|X)}{g(X)}(\Delta Y - \mathcal{E}_{\Delta Y}(X)) - \frac{f_h^0(d|X)}{g_0(X)}(\Delta Y - \mathcal{E}_{\Delta Y}^0(X))\right) \\ =& \frac{K_h(D-d)}{f_d^0}(\mathcal{E}_{\Delta Y}^0(X) - \mathcal{E}_{\Delta Y}(X)) \\ &- \frac{\mathbf{1}\{D=0\}}{f_d^0}\left(\frac{f_h(d|X)}{g(X)} - \frac{f_h^0(d|X)}{g_0(X)}\right)\Delta Y \end{aligned}$$

$$+\frac{\mathbf{1}\{D=0\}}{f_d^0}\left(\frac{f_h(d|X)}{g(X)}\mathcal{E}_{\Delta Y}(X)-\frac{f_h^0(d|X)}{g_0(X)}\mathcal{E}_{\Delta Y}^0(X)\right)$$
$$\lesssim C_1(f_h(X)-f_h^0(X))+C_2(g(X)-g_0(X))+C_3K_h(D-d)(\mathcal{E}_{\Delta Y}(X)-\mathcal{E}_{\Delta Y}^0(X))$$

where the last line can be shown using the "plus-minus" trick with C_1, C_2, C_3 being some constants. Then by the definition of T_N, the assumptions on the rate of convergence of the nuisance parameters, and $E[K_h^2(D-d)] = O(h^{-1})$, we have

$$\begin{aligned}
&\sup_{\eta\in T_N} E[(\psi_\eta-\psi_{\eta_0})^2]\\
&\lesssim \|f_h-f_h^0\|_{P,2}^2+\|g-g_0\|_{P,2}^2+\|K_h(D-d)\|_{P,2}^2\|\mathcal{E}_{\Delta Y}-\mathcal{E}_{\Delta Y}^0\|_{P,2}^2\\
&+\|f_h-f_h^0\|_{P,2}\|g-g_0\|_{P,2}+\|K_h(D-d)\|_{P,2}\|f_h-f_h^0\|_{P,2}\|\mathcal{E}_{\Delta Y}-\mathcal{E}_{\Delta Y}^0\|_{P,2}\\
&+\|K_h(D-d)\|_{P,2}\|g-g_0\|_{P,2}\|\mathcal{E}_{\Delta Y}-\mathcal{E}_{\Delta Y}^0\|_{P,2}\\
&\lesssim h^{-1}\varepsilon_N^2.
\end{aligned}$$

This shows (c).

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

$$(a): \quad \sup_{f\in F_N,\eta\in T_N} E[|\partial_f^2\psi_h(Z,\theta_{0h},f,\eta)-\partial_f^2\psi_h(Z,\theta_{0h},f_d^0,\eta_0)|^2]\lesssim h^{-1}\varepsilon_N^2$$

By definition,

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

Then using Taylor's theorem expand $\partial_f^2\psi_h(Z,\theta_{0h},f,\eta)$ around f_d^0, we have

$$\partial_f^2\psi_h(Z,\theta_{0h},f,\eta)-\partial_f^2\psi_h(Z,\theta_{0h},f_d^0,\eta_0)$$

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

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, by assumption, $(\star\star) = O((Nh)^{-1/2})$, which is dominated by $(\star)$. Therefore we conclude that

$$
\sup_{f \in F_N, \eta \in T_N} E[|\partial_f^2 \psi_h(Z, \theta_{0h}, f, \eta) - \partial_p^2 \psi_h(Z, \theta_{0h}, f_d^0, \eta_0)|^2] \lesssim h^{-1} \varepsilon_N^2.
$$

Similarly, by definition,

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

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

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

$$
\sup_{r \in (0,1), \eta \in T_N} |\partial_r^2 E[\psi_h(\eta_0 + r(\eta - \eta_0))]| \lesssim h^{-1/2} \varepsilon_N^2.
$$

By definition,

$$
\begin{aligned}
&\psi_h(\eta_0 + r(\eta - \eta_0)) \\
&= \frac{K_h(D - d)}{f_d^0} \left(\Delta Y - (\mathcal{E}_{\Delta Y}^0(X) + r(\mathcal{E}_{\Delta Y}(X) - \mathcal{E}_{\Delta Y}^0(X))) \right) - \\
&\frac{\mathbf{1}\{D = 0\}(f_h^0(d|X) + r(f_h(d|X) - f_h^0(d|X)))}{f_d^0(g_0(X) + r(g(X) - g_0(X)))} \left(\Delta Y - \mathcal{E}_{\Delta Y}^0(X) - r(\mathcal{E}_{\Delta Y}(X) - \mathcal{E}_{\Delta Y}^0(X)) \right)
\end{aligned}
$$

and we take the second-order partial derivatives w.r.t. r term by term. For simplicity, we

omit the derivations, and we have

$$\begin{aligned}&\partial_r^2\psi_h(\eta_0 + r(\eta - \eta_0))\\ &\asymp \tilde{C}_1\Delta_f\Delta_g + \tilde{C}_2\Delta_{\mathcal{E}}\Delta_g + \tilde{C}_3\Delta_f\Delta_{\mathcal{E}} + \tilde{C}_4(\Delta_g)^2\end{aligned}$$

where $\Delta_f := f_h - f_h^0$, $\Delta_g := g - g_0$, and $\Delta_{\mathcal{E}} := \mathcal{E}_{\Delta Y} - \mathcal{E}^0_{\Delta Y}$ and $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3, \tilde{C}_4$ are some constants. Then by triangle inequality, Cauchy-Schwarz, and the assumption on the space of nuisance parameters T_N, we have

$$\begin{aligned}E[|\partial_r^2\psi_h(\eta_0 + r(\eta - \eta_0))|] &\lesssim \|f_h - f_h^0\|_{P,2}\|g - g_0\|_{P,2} + \|f_h - f_h^0\|_{P,2}\|\mathcal{E}_{\Delta Y} - \mathcal{E}^0_{\Delta Y}\|_{P,2}\\ &\quad + \|g - g_0\|_{P,2}\|\mathcal{E}_{\Delta Y} - \mathcal{E}^0_{\Delta Y}\|_{P,2} + \|g - g_0\|^2_{P,2}\\ &\lesssim h^{-1/2}\varepsilon_N^2.\end{aligned}$$

Then (d) follows by Jensen's inequality.

Combining previous results, we have

$$\begin{aligned}&\widehat{ATT}(d) - ATT(d)\\ &= \frac{1}{N}\sum_{i=1}^N \psi_h(Z_i, \theta_{0h}, f_d^0, \eta_0) &(3.21)\\ &+ \frac{1}{N}\sum_{i=1}^N S_f^0(K_h(D_i - d) - E[K_h(D_i - d)]) &(3.22)\\ &+ o_p((Nh)^{-1/2}) &(3.23)\\ &+ \theta_0 - \theta_{0h} &(3.24)\end{aligned}$$

where (3.21) and (3.22) are averages of i.i.d. zero-mean terms with the variance growing with kernel bandwidth h, and recall that $S_f^0 = E[\partial_f\psi_h(Z, \theta_{0h}, f_d^0, \eta_0)]$; (3.23) are the terms that vanish when scaled by the (square root of) asymptotic variance; (3.24) is the bias term which is shown to be of order $O(h^2)$ in Lemma 3.4.1.

Since h grows with sample size N, we use the Lyapunov Central Limit Theorem for triangular arrays to establish the asymptotic results. Note that the only term in ψ_h that grows with N is the kernel term, therefore, it suffices to show that the Lyapunov conditions are satisfied for the kernel term. Then, we have

$$\begin{aligned} E[|K_h(D_i - d) - E[K_h(D_i - d)]|^2] &\leq E[(K_h(D_i - d))^2] \\ &= \int \frac{1}{h^2} \left[K\left(\frac{t-d}{h} \right) \right]^2 f_D(t) dt \\ &= \frac{f_D(d)}{h} \int K^2(u) du + o(h^{-1}) \end{aligned}$$

where $f_D(d)$ denotes the density of D at d, and the last line follows from change of variables. Moreover, by the same change of variables argument, we have

$$\begin{aligned} E[|K_h(D_i - d) - E[K_h(D_i - d)]|^3] &\leq 8E[|K_h(D_i - d)|^3] \\ &= 8 \int \frac{1}{h^3} \left| K\left(\frac{t-d}{h} \right) \right|^3 f_D(t) dt \\ &= \frac{f_D(d)}{h^2} \int |K(u)|^3 \, du + o(h^{-2}). \end{aligned}$$

Therefore, we have

$$\begin{aligned} \sigma_{i,N}^2 &:= Var(\psi_h(Z_i, \theta_{0h}, f_d^0, \eta_0)) = O(h^{-1}) \\ r_{i,N} &:= E[|\psi_h(Z_i, \theta_{0h}, f_d^0, \eta_0)|^3] = O(h^{-2}) \end{aligned}$$

Then, the Lyapunov condition is satisfied provided that $Nh \to \infty$ (which is assumed):

$$\frac{(\sum_{i=1}^N r_{i,N})^{1/3}}{(\sum_{i=1}^N \sigma_{i,N}^2)^{1/2}} = O((Nh)^{-1/6}) = o(1).$$

The same argument holds for (3.22). Therefore, by Lyapunov Central Limit Theorem, to-

gether with assumptions 3.4.2 and 3.4.3, we have

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

with σ_N defined by

$$\sigma_N^2 := E\left[\left(\psi_h - \frac{\theta_h}{f_d^0}\left(K_h(D-d) - E[K_h(D-d)]\right)\right)^2\right]$$

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

3.7.5 Proof of Theorem 3.4.1 (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 the parametric rate.

Let T_N be the set of square integrable $\eta := (f_h(d|X), g(X), \mathcal{E}_{\lambda Y}(X))$ such that assumption 3.4.4 holds. Let P_N be the set of $\lambda > 0$ such that $|\lambda - \lambda_0| \leq N^{-1/2}$. Let F_N be the set of $f > 0$ such that $|f - f_d^0| \leq (Nh)^{-1/2}$. Then assumption 3.4.5 implies that, with probability tending to 1, $\hat{\eta}_k \in T_N$, $\hat{f}_{d,k} \in F_N$, and $\hat{\lambda}_k \in P_N$ for all $k = 1, \cdots, K$. Throughout the proof, we use N to denote the sample size and $n := N/K$ to denote the size of any of the subsamples. In particular, since K is fixed, $n \asymp N$.

To simplify notation, let θ_0 denote the true $ATT(d)$, θ_{0h} denote the true $ATT_h(d)$, and $\hat{\theta}_h$ denote our cross-fitted estimator. In particular, recall that our estimator is

$$\hat{\theta}_h := \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i\in I_k}\frac{K_h(D_i-d)\hat{g}_k(X_i) - \mathbf{1}\{D_i=0\}\hat{f}_{h,k}(d|X_i)}{\hat{f}_{d,k}\hat{g}_k(X_i)} \times\left(\frac{T_i-\hat{\lambda}_k}{\hat{\lambda}_k(1-\hat{\lambda}_k)}Y_i - \hat{\mathcal{E}}_{\Delta Y,k}(X_i)\right)$$

Then we can decompose the following difference as

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

where (†) will be our main focus while the bias term (††) is shown in Lemma 3.4.1 to be $O(h^2)$ and asymptotically negligible by the assumption of the under-smoothing bandwidth h.

By definition,

$$\sqrt{N}(\hat{\theta}_h - \theta_{0h}) = \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi_h(Z_i, \theta_{0,h}, \hat{f}_{d,k}, \hat{\eta}_k)] \tag{3.25}$$

where ψ_h is defined as in (3.8), and $E_{n,k}(f) = \frac{1}{n}\sum_{i\in I_k} f(Z_i)$ denotes the empirical average of a generic function f over the set I_k. Then we have the following decomposition, using a multivariate version of Taylor's theorem,

$$\begin{aligned}
\sqrt{N}(\hat{\theta}_h - \theta_{0h}) &= \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \hat{\eta}_k)] && (3.26)\\
&+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \hat{\eta}_k)](\hat{\lambda}_k - \lambda_0) && (3.27)\\
&+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\partial_f \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \hat{\eta}_k)](\hat{f}_{d,k} - f_d^0) && (3.28)\\
&+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\partial_\lambda^2 \psi_h(Z, \theta_{0h}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)](\hat{\lambda}_k - \lambda_0)^2 && (3.29)\\
&+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\partial_f^2 \psi_h(Z, \theta_{0h}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)](\hat{f}_{d,k} - f_d^0)^2 && (3.30)\\
&+ \sqrt{N}\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\partial_\lambda \partial_f \psi_h(Z, \theta_{0h}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)](\hat{f}_{d,k} - f_d^0)(\hat{\lambda}_k - \lambda_0) && (3.31)
\end{aligned}$$

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 (3.29)-(3.31) can be

shown to be $o_p(1)$. The first-order term (3.28) can be analyzed in the same way as the repeat outcomes case. Moreover, since $\hat{\lambda}_k = E_{n,k}T_i$ converges at the parametric rate while the kernel estimator $\hat{f}_{d,k}$ converges at a slower rate, the influence of (3.27) on the asymptotic variance is negligible. The main term (3.26) can be analyzed in the same way as in the repeated outcomes case.

Step 1: Second Order Terms

First, we consider (3.29). By triangle inequality, we have

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

For J_{2k}, since $0 < c < \lambda_0 < 1 - c$, by the boundedness assumption, the score ψ_h satisfies

$$\partial_\lambda^2\psi_h(Z,\theta_{0h},\lambda_0,f_d^0,\eta_0) \lesssim K_h(D-d).$$

Therefore, by the assumption of the kernel function, we have

$$E[J_{2k}^2] \leq \frac{1}{N}E[(\partial_\lambda^2\psi_h(Z,\theta_{0h},\lambda_0,f_d^0,\eta_0))^2] \lesssim E[K_h^2(D-d)]/N \lesssim (hN)^{-1}.$$

Then by Markov's inequality, we have $J_{2k} \leq O_p((hN)^{-1/2})$.

For J_{1k}, note that

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

$$\leq \sup_{\substack{\lambda \in P_N, \ f \in F_N \\ \eta \in T_N}} E[|\partial_\lambda^2 \psi_h(Z, \theta_{0h}, \lambda, f, \eta) - \partial_\lambda^2 \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)|^2]$$

$$\lesssim h^{-1} \varepsilon_N^2 \qquad \text{(a)}$$

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 and hence can be treated as fixed in the conditional expectation. Then by conditional Markov's inequality, the assumption that $(\hat{\lambda}_k - \lambda)^2 \leq O_p(N^{-1})$, and assumption 3.4.4, we conclude that (3.29) $= o_p(1)$. We will show (a) at the end of this section.

Term (3.30) is bounded in the same way as the repeated outcomes case. By triangle inequality, we have

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

To bound J_{4k}, note that since f_d^0 is bounded away from zero,

$$\partial_f^2 \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0) = \frac{2}{(f_d^0)^2}(\psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0) + \theta_{0h}) \lesssim K_h(D - d)$$

which implies that

$$E[J_{4k}^2] \leq \frac{1}{N} E[(\partial_f^2 \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0))^2] \lesssim E[K_h^2(D - d)]/N \lesssim (hN)^{-1}.$$

and by Markov's inequality, we have $J_{4k} \leq O_p((hN)^{-1/2})$. For J_{3k}, we have

$$E[J_{3k}^2 | I_k^c] = E[|E_{n,k}[\partial_f^2 \psi_h(Z, \theta_{0h}, \bar{\lambda}_k, \bar{f}_k, \hat{\eta}_k)] - E_{n,k}[\partial_f^2 \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)]|^2 | I_k^c]$$

$$
\begin{aligned}
&\leq \sup_{\substack{\lambda\in P_N,\ f\in F_N \\ \eta\in T_N}} E[|\partial_f^2\psi_h(Z,\theta_{0h},\lambda,f,\eta) - \partial_f^2\psi_h(Z,\theta_{0h},\lambda_0,f_d^0,\eta_0)|^2|I_k^c] \\
&\leq \sup_{\substack{\lambda\in P_N,\ f\in F_N \\ \eta\in T_N}} E[|\partial_f^2\psi_h(Z,\theta_{0h},\lambda,f,\eta) - \partial_f^2\psi_h(Z,\theta_{0h},\lambda_0,f_d^0,\eta_0)|^2] \\
&\lesssim h^{-1}\varepsilon_N^2 \qquad \text{(b)}
\end{aligned}
$$

Then by conditional Markov's inequality, $(\hat{f}_{d,k} - f_d^0)^2 \leq O_p((Nh)^{-1})$, and assumption 3.4.4, we conclude that (3.30) $= o_p(1)$. We verify (b) at the end of this section.

Finally, we can bound (3.31) using similar arguments as those for (3.29) and (3.30). To avoid repetitiveness, we only highlight the difference. In particular, we need

$$
\sup_{\substack{\lambda\in P_N,\ f\in F_N \\ \eta\in T_N}} E[|\partial_\lambda\partial_f\psi_J(Z,\theta_{0J},\bar{\lambda}_k,\bar{f}_k,\hat{\eta}_k) - \partial_\lambda\partial_f\psi_J(Z,\theta_{0J},\lambda_0,f_d^0,\eta_0)|^2] \lesssim h^{-1}\varepsilon_N^2 \qquad \text{(c)}
$$

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 3.4.4, we conclude that (3.31) $= o_p(1)$. Claim (c) will be shown later. Therefore, we have shown that all the second-order terms are asymptotically negligible.

Step 2: First-Order Terms

We first consider (3.27). By triangle inequality, we have

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

To bound J_{6k}, since λ_0 is bounded away from zero, the score ψ satisfies,

$$
\partial_\lambda\psi_h(Z,\theta_{0h},\lambda_0,f_d^0,\eta_0) \lesssim K_h(D-d).
$$

This implies that

$$E[J_{6k}^2] \leq \frac{1}{N} E[(\partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0))^2] \lesssim E[K_h^2(D-d)]/N \lesssim (Nh)^{-1}.$$

and by Markov's inequality, we have $J_{6k} \leq O_p((Nh)^{-1/2})$. With the assumption that $Nh \to \infty$, we have $J_{6k} = o_p(1)$.

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

$$\begin{aligned}
E[J_{5k}^2 | I_k^c] &= E[|E_{n,k}[\partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \hat{\eta}_k)] - E_{n,k}[\partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)]|^2 | I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta) - \partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)|^2 | I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta) - \partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim h^{-1} \varepsilon_N^2 \qquad \text{(d)}
\end{aligned}$$

where the first equation holds by definition, the second line holds by the Cauchy-Schwarz inequality, and the third line holds by the construction that all the parameters are estimated using auxiliary sample I_k^c and hence can be treated as fixed. Then we conclude with conditional Markov's inequality that $J_{5k} = o_p(1)$. As before, we will show (d) at the end of this section.

Therefore,

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

Note that $(\hat{\lambda}_k - \lambda_0) = O_p(N^{-1/2})$, we can rewrite (3.27) as

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

$$= \sqrt{N}\frac{1}{N}\sum_{i=1}^{N} S_\lambda^0(T_i - \lambda_0) + o_p(1)$$

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 $S_\lambda^0 = E[\partial_\lambda \psi_h^0]$ is bounded by a constant and $\hat{\lambda}$ converges at parametric rate, (3.27) vanishes when scaled by the square-root of the asymptotic variance that grows with sample size.

Term (3.28) will be bounded using the same argument as in the repeated outcomes setting. First, by the triangle inequality

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

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

$$\partial_f \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0) = -\frac{1}{f_d^0}(\psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0) + \theta_{0h}) \lesssim K_h(D-d),$$

which implies that

$$E[J_{8k}^2] \leq \frac{1}{N}E[(\partial_f \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0))^2] \lesssim E[K_h^2(D-d)]/N \lesssim (Nh)^{-1}.$$

Then by Markov's inequality, we have $J_{8k} \leq O_p((Nh)^{-1/2})$. With the assumption that $Nh \to \infty$, we have $J_{8k} = o_p(1)$.

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

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

$$
\begin{aligned}
&\leq \sup_{\eta \in T_N} E[|\partial_f \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta) - \partial_f \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)|^2 | I_k^c] \\
&\leq \sup_{\eta \in T_N} E[|\partial_f \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta) - \partial_f \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim h^{-1} \varepsilon_N^2 \qquad \text{(e)}
\end{aligned}
$$

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,

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

Note that under the assumption, $(\hat{f}_{d,k} - f_d^0) = O_p((Nh)^{-1/2})$, we can rewrite (3.28) as

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

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 (3.26), which can be shown using the same argument as the repeated

outcomes case.

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

Since K is fixed, $n = O(N)$, it suffices to show that $R_{nk} = o_p(N^{-1/2}h^{-1})$, 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

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

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 the conditional version of the empirical process conditioning on the auxiliary sample I_k^c. Moreover,

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

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

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

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

and using the conditional Markov's inequality, we conclude that $R_{1k} = O_p(h^{-1/2}\varepsilon_N)$.

Now we bound R_{2k}. Note that by definition of the score, $E[\psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)] = 0$, so it suffices to bound $E[\psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \hat{\eta}_k)|I_k^c]$. Suppressing other arguments in the score, define

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

where by definition $h_k(0) = E[\psi_h(\eta_0)|I_k^c] = 0$ and $h_k(1) = E[\psi_h(\hat{\eta}_k)|I_k^c]$. 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,

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

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

$$R_{2k} = \sqrt{n}|h_k(1)| = \sqrt{n}|h_k''(\bar{r})|$$

$$\leq \sup_{r\in(0,1),\eta\in T_N} \sqrt{n}|\partial_r^2 E[\psi_h(\eta_0 + r(\hat{\eta}_k - \eta_0))]|$$
$$\lesssim \sqrt{n}h^{-1}\varepsilon_N^2 \qquad (g)$$

Combining the above results, we conclude that

$$\sqrt{N}R_{n,k} \lesssim h^{-1/2}\varepsilon_N + \sqrt{N}h^{-1/2}\varepsilon_N^2,$$

and for $\varepsilon_N = o(N^{-1/4})$, we have $\sqrt{N}R_{n,k} = o_p(h^{-1/2})$.

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. Therefore, we focus on (a), (c), and (d) in the repeated cross-sections setting.

First, recall that

$$(a): \quad \sup_{\substack{\lambda\in P_N,\ f\in F_N \\ \eta\in T_N}} E[|\partial_\lambda^2\psi_h(Z,\theta_{0h},\lambda,f,\eta) - \partial_\lambda^2\psi_h(Z,\theta_{0h},\lambda_0,f_d^0,\eta_0)|^2] \lesssim h^{-1}\varepsilon_N^2.$$

In particular,

$$\partial_\lambda^2\psi_h(\lambda,f_d,\eta) = \frac{\partial^2}{\partial\lambda^2}\frac{K_h(D-d)g(X) - \mathbf{1}\{D=0\}f_h(d|X)}{f_d\cdot g(X)}\frac{T-\lambda}{\lambda(1-\lambda)}Y$$

where we suppressed the common terms (Z,θ_{0h}) in ψ_h for simplicity. Then by Taylor's theorem,

$$\begin{aligned}\partial_\lambda^2\psi_h(\lambda,f_d,\eta) - \partial_\lambda^2\psi_h(\lambda_0,f_d^0,\eta_0) &= \partial_\lambda^2\psi_h(\lambda_0,f_d^0,\eta) - \partial_\lambda^2\psi_h(\lambda_0,f_d^0,\eta_0) \quad (\star)\\ &\quad + \partial_\lambda^2\partial_f\psi_h(\bar{\lambda},\bar{f}_d,\eta)(f_d - f_d^0) \quad (\star\star)\\ &\quad + \partial_\lambda^3\psi_h(\bar{\lambda},\bar{f}_d,\eta)(\lambda-\lambda_0) \quad (\star\star\star)\end{aligned}$$

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

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

Moreover, by assumption 3.4.4, 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

$$
\begin{aligned}
&\sup_{\substack{\lambda \in P_N, \ f \in F_N \\ \eta \in T_N}} E[|\partial_\lambda^2 \psi_h(Z, \theta_{0h}, \lambda, f, \eta) - \partial_\lambda^2 \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim \sup_{\eta \in T_N} E[|\partial_\lambda^2 \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta) - \partial_\lambda^2 \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim \sup_{\eta \in T_N} \|f_h(d|X) - f_h^0(d|X)\|_{P,2}^2 + \|g(X) - g_0(X)\|_{P,2}^2 \\
&\lesssim h^{-1} \epsilon_N^2
\end{aligned}
$$

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

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

and (c) holds by similar arguments as (a).

Finally, we show (d):

$$
\sup_{\eta \in T_N} E[|\partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta) - \partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)|^2] \lesssim h^{-1} \varepsilon_N^2.
$$

By the same argument as (a),

$$\partial_\lambda \psi_h(\lambda, f_d, \eta) = \frac{K_h(D-d)g(X) - \mathbf{1}\{D=0\}f_h(d|X)}{f_d \cdot g(X)} \frac{T-\lambda}{\lambda(1-\lambda)} Y,$$

which implies

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

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

$$\begin{aligned}
&\sup_{\eta \in T_N} E[|\partial_\lambda \psi_h(Z, \theta_{0h}, \lambda, f, \eta) - \partial_\lambda \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)|^2] \\
&\lesssim \sup_{\eta \in T_N} \|f_h(d|X) - f_h^0(d|X)\|_{P,2}^2 + \|g(X) - g_0(X)\|_{P,2}^2 \\
&+ \|f_h(d|X) - f_h^0(d|X)\|_{P,2} \|g(X) - g_0(X)\|_{P,2} \\
&\lesssim h^{-1}\epsilon_N^2.
\end{aligned}$$

This completes the proofs for the auxiliary results.

Combining previous results, we have

$$\begin{aligned}
&\widehat{ATT}(d) - ATT(d) \\
&= \frac{1}{N}\sum_{i=1}^N \psi_h(Z_i, \theta_{0h}, \lambda_0, f_d^0, \eta_0) && (3.32) \\
&+ \frac{1}{N}\sum_{i=1}^N S_f^0(K_h(D_i - d) - E[K_h(D_i - d)]) && (3.33) \\
&+ o_p((Nh)^{-1/2}) && (3.34)
\end{aligned}$$

$$+ \theta_0 - \theta_{0h} \tag{3.35}$$

where (3.32) and (3.33) are averages of i.i.d. zero-mean terms with the variance growing with kernel bandwidth h, and recall that $S_f^0 = E[\partial_f \psi_h(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)]$; (3.34) are the terms that vanish when scaled by the (square root of) asymptotic variance; (3.35) is the bias term which is shown to be of order $O(h^2)$ in Lemma 3.4.1.

Note that we have arrived at the identical decomposition as in the repeated outcomes case, and by the same argument, we have

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

with σ_N defined by

$$\sigma_N^2 := E\left[\left(\psi_h - \frac{\theta_h}{f_d^0}\left(K_h(D-d) - E[K_h(D-d)]\right)\right)^2\right]$$

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

3.7.6 Proof of Theorem 3.4.2 (Repeated Outcomes)

The proof uses the same idea as in CCDDHNR (2018) and Chang (2020). However, we need to adapt the proof to accommodate the presence of the kernel term. First, recall that the variance estimator is defined as

$$\begin{aligned}\hat{\sigma}_N^2 :=& \frac{1}{K}\sum_{k=1}^{K} E_{n,k}\left[\left(\psi_h(Z, \hat{\theta}_h, \hat{f}_{d,k}, \hat{\eta}_k) - \frac{\hat{\theta}_h}{\hat{f}_{d,k}}\left(K_h(D-d) - \hat{f}_{d,k}\right)\right)^2\right] \\ :=& \frac{1}{K}\sum_{k=1}^{K} E_{n,k}\left[\left(\tilde{\psi}_h(Z, \hat{\theta}_h, \hat{f}_{d,k}, \hat{\eta}_k)\right)^2\right]\end{aligned}$$

where we define

$$\tilde{\psi}_h(Z,\theta,f_d,\eta) := \frac{K_h(D-d)g(X) - \mathbf{1}\{D=0\}f_h(d|X)}{f_d g(X)}\left(\Delta Y - \mathcal{E}_{\Delta Y}(X)\right) - \frac{\theta_h}{f_d}K_h(D-d).$$

In particular, note that $\sigma_N^2 = E\left[\tilde{\psi}_h^2(Z,\theta_{0h},f_d^0,\eta_0)\right]$. Therefore, we need to show that

$$J_k := \left|E_{n,k}\left[\tilde{\psi}_h^2(Z,\hat{\theta}_h,\hat{f}_{d,k},\hat{\eta}_k)\right] - E\left[\tilde{\psi}_h^2(Z,\theta_{0h},f_d^0,\eta_0)\right]\right| = o_p(1).$$

By the triangle inequality, we have

$$\begin{aligned} J_k \leq & \underbrace{\left|E_{n,k}\left[\tilde{\psi}_h^2(Z,\hat{\theta}_h,\hat{f}_{d,k},\hat{\eta}_k)\right] - E_{n,k}\left[\tilde{\psi}_h^2(Z,\theta_{0h},f_d^0,\eta_0)\right]\right|}_{:=J_{1k}} \\ & + \underbrace{\left|E_{n,k}\left[\tilde{\psi}_h^2(Z,\theta_{0h},f_d^0,\eta_0)\right] - E\left[\tilde{\psi}_h^2(Z,\theta_{0h},f_d^0,\eta_0)\right]\right|}_{:=J_{2k}}. \end{aligned}$$

We bound each term separately.

First, we consider J_{2k}.

$$\begin{aligned} E[J_{2k}^2] &= E\left[\left(E_{n,k}\left[\tilde{\psi}_h^2(Z,\theta_{0h},f_d^0,\eta_0)\right] - E\left[\tilde{\psi}_h^2(Z,\theta_{0h},f_d^0,\eta_0)\right]\right)^2\right] \\ &\leq E\left[\left(\frac{1}{n}\sum_{i=1}^n \tilde{\psi}_h^2(Z_i,\theta_{0h},f_d^0,\eta_0)\right)^2\right] \\ &\leq \frac{1}{n}E\left[\tilde{\psi}_h^4(Z,\theta_{0h},f_d^0,\eta_0)\right] \\ &\lesssim \frac{1}{n}E\left[K_h^4(D-d)\right] \\ &\lesssim (nh^3)^{-1}, \end{aligned}$$

where the third line holds by Cauchy-Schwarz inequality, the fourth line holds by boundedness assumption, and the last line holds by change of variables using the assumptions on the kernel. Therefore, by Chebyshev's inequality, we have $J_{2k} = o_p(1)$ if $nh^3 \to \infty$.

Next, we consider J_{1k}. First, we state a convenient fact that will be used in the proof, see CCDDHNR (2018); Chang (2020) for example: for any constants a and δ,

$$|(a+\delta a)^2 - a^2| \leq 2|\delta a|(|a| + |\delta a|).$$

In our context, we define (for notation simplicity)

$$\begin{aligned} a &= \tilde{\psi}_h(Z_i, \theta_{0h}, f_d^0, \eta_0) := \psi_i \\ a + \delta a &= \tilde{\psi}_h(Z_i, \hat{\theta}_h, \hat{f}_{d,k}, \hat{\eta}_k) := \hat{\psi}_i \end{aligned}$$

Then, we have

$$\begin{aligned} J_{1k} &= \left| \frac{1}{n} \sum_{i \in I_k} \hat{\psi}_i^2 - \psi_i^2 \right| \leq \frac{1}{n} \sum_{i \in I_k} \left| \hat{\psi}_i^2 - \psi_i^2 \right| \\ &\leq \frac{2}{n} \sum_{i \in I_k} |\hat{\psi}_i - \psi_i|(|\psi_i| + |\hat{\psi}_i - \psi_i|) \\ &\leq 2 \left(\frac{1}{n} \sum_{i \in I_k} |\hat{\psi}_i - \psi_i|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i \in I_k} (|\psi_i| + |\hat{\psi}_i - \psi_i|)^2 \right)^{1/2} \\ &\leq 2 \left(\frac{1}{n} \sum_{i \in I_k} |\hat{\psi}_i - \psi_i|^2 \right)^{1/2} \left[\left(\frac{1}{n} \sum_{i \in I_k} |\psi_i|^2 \right)^{1/2} + \left(\frac{1}{n} \sum_{i \in I_k} |\hat{\psi}_i - \psi_i|^2 \right)^{1/2} \right]. \end{aligned}$$

where the third line holds by Cauchy-Schwarz inequality, and the last line holds by the triangle inequality. Then, we have

$$J_{1k}^2 \lesssim S_N \left(S_N + \frac{1}{n} \sum_{i \in I_k} \psi_i^2 \right)$$

where $S_N := \frac{1}{n} \sum_{i \in I_k} |\hat{\psi}_i - \psi_i|^2$.

We now bound S_N. By the definition of $\tilde{\psi}_h$, we have

$$\begin{aligned}
S_N &= \frac{1}{n}\sum_{i\in I_k}\left(\tilde{\psi}_h(Z_i,\hat{\theta}_h,\hat{f}_{d,k},\hat{\eta}_k)-\tilde{\psi}_h(Z_i,\theta_{0h},f_d^0,\eta_0)\right)^2 \\
&= \frac{1}{n}\sum_{i\in I_k}\left(\tilde{\psi}_h(Z_i,\theta_{0h},\hat{f}_{d,k},\hat{\eta}_k)-\tilde{\psi}_h(Z_i,\theta_{0h},f_d^0,\eta_0)+\frac{\partial}{\partial\theta}\tilde{\psi}_h(Z_i,\bar{\theta},\hat{f}_{d,k},\hat{\eta}_k)\right)^2 \\
&\lesssim \underbrace{\frac{1}{n}\sum_{i\in I_k}\left(\tilde{\psi}_h(Z_i,\theta_{0h},\hat{f}_{d,k},\hat{\eta}_k)-\tilde{\psi}_h(Z_i,\theta_{0h},f_d^0,\eta_0)\right)^2}_{:=S_{1N}} + \underbrace{\frac{1}{n}\sum_{i\in I_k}\left(\frac{K_h(D_i-d)}{\hat{f}_{d,k}}(\hat{\theta}_h-\theta_{0h})\right)^2}_{:=S_{2N}}
\end{aligned}$$

where the second line holds by Taylor's theorem with $\bar{\theta}$ between θ_{0h} and $\hat{\theta}_h$, and the last line holds by the fact that $\frac{\partial}{\partial\theta}\tilde{\psi}_h(Z_i,\bar{\theta},\hat{f}_{d,k},\hat{\eta}_k) = K_h(D_i-d)/\hat{f}_{d,k}$. We bound S_{1N} and S_{2N} separately.

To bound S_{2N}, note that

$$S_{2N} = \frac{(\hat{\theta}_h-\theta_{0h})^2}{\hat{f}_{d,k}^2}\frac{1}{n}\sum_{i\in I_k}K_h^2(D_i-d).$$

Since $E[K_h^2(D-d)] = O(h^{-1})$, by Markov's inequality, we have $\frac{1}{n}\sum_{i\in I_k}K_h^2(D_i-d) = O_p(h^{-1})$. Moreover, by Theorem 3.4.1, we have $(\hat{\theta}_h-\theta_{0h})^2 = O_p((Nh)^{-1})$. Therefore, we conclude

$$S_{2N} \le O_p((Nh^2)^{-1}).$$

Next, we bound S_{1N}. By Taylor's theorem, for $\bar{f}$ between f_d^0 and $\hat{f}_{d,k}$, we have

$$\begin{aligned}
&\tilde{\psi}_h(Z_i,\theta_{0h},\hat{f}_{d,k},\hat{\eta}_k)-\tilde{\psi}_h(Z_i,\theta_{0h},f_d^0,\eta_0) \\
&= \tilde{\psi}_h(Z_i,\theta_{0h},f_d^0,\hat{\eta}_k)-\tilde{\psi}_h(Z_i,\theta_{0h},f_d^0,\eta_0)+\frac{\partial}{\partial f}\tilde{\psi}_h(Z_i,\theta_{0h},\bar{f},\hat{\eta}_k)(\hat{f}_{d,k}-f_d^0).
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{\partial}{\partial_f}\tilde{\psi}_h(Z,\theta,f,\eta) \\
&= \frac{\partial}{\partial f}\left(\frac{K_h(D-d)g(X) - \mathbf{1}\{D=0\}f_h(d|X)}{fg(X)}\left(\Delta Y - \mathcal{E}_{\Delta Y}(X)\right) - \frac{\theta_h}{f}K_h(D-d)\right) \\
&\lesssim K_h(D-d)
\end{aligned}$$

where the last line holds by the boundedness assumption. Therefore, by the assumption on the kernel, we have

$$\|\partial_f\tilde{\psi}_h(Z,\theta_{0h},\bar{f},\hat{\eta}_k)\|_{P,2} \lesssim \|K_h(D-d)\|_{P,2} = O(h^{-1/2}).$$

Moreover, by definition

$$\tilde{\psi}_h(Z_i,\theta_{0h},f_d^0,\hat{\eta}_k) - \tilde{\psi}_h(Z_i,\theta_{0h},f_d^0,\eta_0) = \psi_h(Z_i,\theta_{0h},f_d^0,\hat{\eta}_k) - \psi_h(Z_i,\theta_{0h},f_d^0,\eta_0).$$

Then, by triangle inequality and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\|\tilde{\psi}_h(Z_i,\theta_{0h},\hat{f}_{d,k},\hat{\eta}_k) - \tilde{\psi}_h(Z_i,\theta_{0h},f_d^0,\eta_0)\|_{P,2}^2 \\
&\lesssim \|\psi_h(Z_i,\theta_{0h},f_d^0,\hat{\eta}_k) - \psi_h(Z_i,\theta_{0h},f_d^0,\eta_0)\|_{P,2}^2 + \|\partial_f\tilde{\psi}_h(Z_i,\theta_{0h},\bar{f},\hat{\eta}_k)\|_{P,2}^2\|\hat{f}_{d,k} - f_d^0\|_{P,2}^2 \\
&\lesssim h^{-1}\varepsilon_N^2 + h^{-2}N^{-1}.
\end{aligned}$$

where the last line holds by the assumptions on the rate of convergence of the $\hat{f}_{d,k}$ and $\|\psi_h(Z_i,\theta_{0h},f_d^0,\hat{\eta}_k) - \psi_h(Z_i,\theta_{0h},f_d^0,\eta_0)\|_{P,2}^2 \lesssim h^{-1}\varepsilon_N^2$ by the same arguments as in the proof of Theorem 3.4.1. Then by Markov's inequality, we have

$$S_{1N} = O_p\left(h^{-1}\varepsilon_N^2 + h^{-2}N^{-1}\right).$$

Combining the results, we have

$$S_N = O_p\left(h^{-1}\varepsilon_N^2 + h^{-2}N^{-1}\right).$$

Note that since $\psi_i \lesssim K_h(D-d)$, $\frac{1}{n}\sum_{i\in I_k}\psi_i^2 = O_p(h^{-1})$ by Markov's inequality. This implies that

$$J_{1k}^2 \lesssim S_N\left(S_N + \frac{1}{n}\sum_{i\in I_k}\psi_i^2\right) = O_p\left(h^{-2}\varepsilon_N^2 + h^{-3}N^{-1}\right).$$

Then $J_{1k} = o_p(1)$ if $h^{-2}\varepsilon_N^2 + h^{-3}N^{-1} \to 0$.

Therefore, we conclude that $\hat{\sigma}_N^2 = \sigma_N^2 + o_p(1)$. □

3.7.7 Proof of Theorem 3.4.2 (Repeated Cross-Sections)

The proof is nearly identical to the repeated outcomes case, and we only highlight the key differences. Again, the main idea follows from CCDDHNR (2018); Chang (2020), and our proof requires modifications to take into account the kernel function present in the score function.

First, recall that the variance estimator is defined as

$$\begin{aligned}\hat{\sigma}_N^2 :=& \frac{1}{K}\sum_{k=1}^{K} E_{n,k}\left[\left(\psi_h(Z,\hat{\theta}_h,\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k) - \frac{\hat{\theta}_h}{\hat{f}_{d,k}}\left(K_h(D-d) - \hat{f}_{d,k}\right)\right)^2\right] \\ :=& \frac{1}{K}\sum_{k=1}^{K} E_{n,k}\left[\left(\tilde{\psi}_h(Z,\hat{\theta}_h,\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k)\right)^2\right]\end{aligned}$$

where we define

$$\tilde{\psi}_h(Z,\theta,\lambda,f_d,\eta) := \frac{K_h(D-d)g(X) - \mathbf{1}\{D=0\}f_h(d|X)}{f_d g(X)}\left(\frac{T-\lambda}{\lambda(1-\lambda)}Y - \mathcal{E}_{\lambda Y}(X)\right)$$

$$- \frac{\theta_h}{f_d} K_h(D-d).$$

In particular, note that $\sigma_N^2 = E\left[\tilde{\psi}_h^2(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)\right]$. Therefore, we need to show that

$$J_k := \left| E_{n,k}\left[\tilde{\psi}_h^2(Z, \hat{\theta}_h, \hat{\lambda}_k, \hat{f}_{d,k}, \hat{\eta}_k)\right] - E\left[\tilde{\psi}_h^2(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)\right]\right| = o_p(1).$$

By the triangle inequality, we have

$$\begin{aligned} J_k \leq & \underbrace{\left| E_{n,k}\left[\tilde{\psi}_h^2(Z, \hat{\theta}_h, \hat{\lambda}_k, \hat{f}_{d,k}, \hat{\eta}_k)\right] - E_{n,k}\left[\tilde{\psi}_h^2(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)\right]\right|}_{:=J_{1k}} \\ & + \underbrace{\left| E_{n,k}\left[\tilde{\psi}_h^2(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)\right] - E\left[\tilde{\psi}_h^2(Z, \theta_{0h}, \lambda_0, f_d^0, \eta_0)\right]\right|}_{:=J_{2k}}. \end{aligned}$$

We bound each term separately.

Similar to the repeated outcomes case, by boundedness and assumptions on the kernel function, J_{2k} satisfies

$$E[J_{2k}^2] \lesssim \frac{1}{n} E\left[K_h^4(D-d)\right] \lesssim (nh^3)^{-1}.$$

Therefore, by Markov's inequality, we have $J_{2k} = o_p(1)$ if $nh^3 \to \infty$.

Next, we bound J_{1k}. For notation simplicity, we define

$$\begin{aligned} \psi_i :=& \tilde{\psi}_h(Z_i, \theta_{0h}, \lambda_0, f_d^0, \eta_0) \\ \hat{\psi}_i :=& \tilde{\psi}_h(Z_i, \hat{\theta}_h, \hat{\lambda}_k, \hat{f}_{d,k}, \hat{\eta}_k). \end{aligned}$$

Then, using the same argument as in the repeated outcomes case, we have

$$J_{1k}^2 \lesssim S_N \left(S_N + \frac{1}{n} \sum_{i \in I_k} \psi_i^2 \right)$$

where $S_N := \frac{1}{n}\sum_{i\in I_k}|\hat{\psi}_i - \psi_i|^2$.

By triangle inequality, we have

$$\begin{aligned}
S_N &= \frac{1}{n}\sum_{i\in I_k}\left(\tilde{\psi}_h(Z_i,\hat{\theta}_h,\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k) - \tilde{\psi}_h(Z_i,\theta_{0h},\lambda_0,f_d^0,\eta_0)\right)^2 \\
&= \frac{1}{n}\sum_{i\in I_k}\left(\tilde{\psi}_h(Z_i,\theta_{0h},\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k) - \tilde{\psi}_h(Z_i,\theta_{0h},\lambda_0,f_d^0,\eta_0) + \frac{\partial}{\partial\theta}\tilde{\psi}_h(Z_i,\bar{\theta},\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k)\right)^2 \\
&\lesssim \underbrace{\frac{1}{n}\sum_{i\in I_k}\left(\tilde{\psi}_h(Z_i,\theta_{0h},\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k) - \tilde{\psi}_h(Z_i,\theta_{0h},\lambda_0,f_d^0,\eta_0)\right)^2}_{:=S_{1N}} \\
&+ \underbrace{\frac{1}{n}\sum_{i\in I_k}\left(\frac{K_h(D_i-d)}{\hat{f}_{d,k}}(\hat{\theta}_h - \theta_{0h})\right)^2}_{:=S_{2N}}
\end{aligned}$$

where the second line holds by Taylor's theorem with $\bar{\theta}$ between θ_{0h} and $\hat{\theta}_h$, and the last line holds by the fact that $\frac{\partial}{\partial\theta}\tilde{\psi}_h(Z_i,\bar{\theta},\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k) = K_h(D_i-d)/\hat{f}_{d,k}$.

Note that, using the identical argument as in the repeated outcomes case,

$$S_{2N} = O_p(h^{-1}) \times O_p((Nh)^{-1}).$$

Moreover, by Taylor's theorem, for $\bar{f}$ between f_d^0 and $\hat{f}_{d,k}$, and for $\bar{\lambda}$ between λ_0 and $\hat{\lambda}_k$, we have

$$\begin{aligned}
&\|\tilde{\psi}_h(Z_i,\theta_{0h},\hat{\lambda}_k,\hat{f}_{d,k},\hat{\eta}_k) - \tilde{\psi}_h(Z_i,\theta_{0h},\lambda_0,f_d^0,\eta_0)\|_{P,2}^2 \\
&\lesssim \|\psi_h(Z_i,\theta_{0h},\lambda_0,f_d^0,\hat{\eta}_k) - \psi_h(Z_i,\theta_{0h},f_d^0,\eta_0)\|_{P,2}^2 \\
&+ \|\partial_\lambda\tilde{\psi}_h(Z_i,\theta_{0h},\bar{\lambda},\bar{f},\hat{\eta}_k)\|_{P,2}^2\|\hat{\lambda}_k - \lambda_0\|_{P,2}^2 \\
&+ \|\partial_f\tilde{\psi}_h(Z_i,\theta_{0h},\bar{\lambda},\bar{f},\hat{\eta}_k)\|_{P,2}^2\|\hat{f}_{d,k} - f_d^0\|_{P,2}^2
\end{aligned}$$

By boundedness assumption, we have

$$\frac{\partial}{\partial_f}\tilde{\psi}_h(Z,\theta,\lambda,f,\eta) \lesssim K_h(D-d)$$
$$\frac{\partial}{\partial_\lambda}\tilde{\psi}_h(Z,\theta,\lambda,f,\eta) \lesssim K_h(D-d)$$

and by the same argument as in the repeated outcomes case, we have

$$\|\partial_\lambda \tilde{\psi}_h(Z_i,\theta_{0h},\bar{\lambda},\bar{f},\hat{\eta}_k)\|_{P,2}^2 = O\left(h^{-1}\right)$$
$$\|\partial_f \tilde{\psi}_h(Z_i,\theta_{0h},\bar{\lambda},\bar{f},\hat{\eta}_k)\|_{P,2}^2 = O\left(h^{-1}\right).$$

Note that $\|\hat{\lambda}_k - \lambda_0\|_{P,2}^2 = O(N^{-1})$, $\|\hat{f}_{d,k} - f_d^0\|_{P,2}^2 = O((Nh)^{-1})$, and $\|\psi_h(Z_i,\theta_{0h},\lambda_0,f_d^0,\hat{\eta}_k) - \psi_h(Z_i,\theta_{0h},\lambda_0,f_d^0,\eta_0)\|_{P,2}^2 \lesssim h^{-1}\varepsilon_N^2$ by the same arguments as in the proof of Theorem 3.4.1. Therefore, by Markov's inequality, we have

$$S_{1N} = O_p\left(h^{-1}\varepsilon_N^2 + h^{-2}N^{-1}\right).$$

Combining the results, we have

$$S_N = O_p\left(h^{-1}\varepsilon_N^2 + h^{-2}N^{-1}\right).$$

Since $\psi_i \lesssim K_h(D-d)$, $\frac{1}{n}\sum_{i\in I_k}\psi_i^2 = O_p(h^{-1})$ by Markov's inequality. This implies that

$$J_{1k}^2 \lesssim S_N\left(S_N + \frac{1}{n}\sum_{i\in I_k}\psi_i^2\right) = O_p\left(h^{-2}\varepsilon_N^2 + h^{-3}N^{-1}\right).$$

Then $J_{1k} = o_p(1)$ if $h^{-2}\varepsilon_N^2 + h^{-3}N^{-1} \to 0$.

Therefore, we conclude that $\hat{\sigma}_N^2 = \sigma_N^2 + o_p(1)$. □

Bibliography

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.

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.

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.

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.

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. (2024). *Difference-in-differences with a continuous treatment* (No. w32117). National Bureau of Economic Research, 2024.

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.

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

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.

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

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*.

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.

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 Yao, Q. (2003). *Nonlinear time series: nonparametric and parametric methods* (Vol. 20). New York: Springer.

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.

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

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.

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*.

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.

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

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.