UNIVERSITY OF CALIFORNIA

Los Angeles

Three Essays on Causal Inference with High-dimensional Data and Machine Learning Methods

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Economics

by

Neng-Chieh Chang

2020

ABSTRACT OF THE DISSERTATION

Three Essays on Causal Inference with High-dimensional Data and Machine Learning Methods

by

Neng-Chieh Chang
Doctor of Philosophy in Economics
University of California, Los Angeles, 2020
Professor Denis Nikolaye Chetverikov, Chair

This dissertation consists of three chapters that study causal inference when applying machine learning methods. In Chapter 1, I propose an orthogonal extension of the semiparametric difference-in-differences estimator proposed in Abadie (2005). The proposed estimator enjoys the so-called Neyman-orthogonality (Chernozhukov et al. 2018) and thus it allows researchers to flexibly use a rich set of machine learning (ML) methods in the first-step estimation. It is particularly useful when researchers confront a high-dimensional data set when the number of potential control variables is larger than the sample size and the conventional nonparametric estimation methods, such as kernel and sieve estimators, do not apply. I apply this orthogonal difference-in-differences estimator to evaluate the effect of tariff reduction on corruption. The empirical results show that tariff reduction decreases corruption in large magnitude.

In Chapter 2, I study the estimation and inference of the mode treatment effect. Mean, median, and mode are three essential measures of the centrality of probability distributions. In program evaluation, the average treatment effect (mean) and the quantile treatment effect (median) have been intensively studied in the past decades. The mode treatment effect, however, has long been neglected in program evaluation. This paper fills the gap by

discussing both the estimation and inference of the mode treatment effect. I propose both traditional kernel and machine learning methods to estimate the mode treatment effect. I also derive the asymptotic properties of the proposed estimators and find that both estimators follow the asymptotic normality but with the rate of convergence slower than the regular rate $\sqrt{N}$, which is different from the rates of the classical average and quantile treatment effect estimators.

In Chapter 3 (joint with Liqiang Shi), we study the estimation and inference of the doubly robust extension of the semiparametric quantile treatment effect estimation discussed in Firpo (2007). This proposed estimator allows researchers to use a rich set of machine learning methods in the first-step estimation, while still obtaining valid inferences. Researchers can include as many control variables as they consider necessary, without worrying about the over-fitting problem which frequently happens in the traditional estimation methods. This paper complements Belloni et al. (2017), which provided a very general framework to discuss the estimation and inference of many different treatment effects when researchers apply machine learning methods.

The dissertation of Neng-Chieh Chang is approved.

Rosa Liliana Matzkin

Andres Santos

Arash Ali Amini

Denis Nikolaye Chetverikov, Committee Chair

University of California, Los Angeles

2020

DEDICATIONS

To my lovely family

Contents

List of Figures

List of Tables

ACKNOWLEDGMENT

I am grateful to Denis Chetverikov, Rosa Matzkin, Andres Santos, and Arash A. Amini for invaluable guidance and continuous support. I thank Jinyong Hahn, Zhipen Liao, Shuyang Sheng, and participants at UCLA econometrics proseminars for helpful comments and discussions. All errors, of course, are mine.

VITA

Education

University of California, Los Angeles	Los Angeles, USA
C.Phil., Economics, Department of Economics	2017
M.A., Economics, Department of Economics	2016
National Taiwan University	Taipei, Taiwan
B.S., Physics, Department of Physics	2013

Fellowships and Awards

UCLA Economic Departmental Teaching Assistantship, 2016-2020

Teaching Experience

Teaching Assistant

Econometrics Laboratory, UCLA, Summer 2018, Winter 2019, Spring 2019, Fall 2019

Microeconomic Theory I, UCLA, Fall 2016, Winter 2018

Statistics for Economists, UCLA, Spring 2020

Microeconomic Theory II, UCLA, Spring 2018, Fall 2018

Principles of Economics I, UCLA, Winter 2017, Spring 2017

Chapter 1

Double/Debiased Machine Learning for Difference-in-Differences Models

1.1 Introduction

The difference-in-differences (DiD) estimator has been widely used in empirical economics to evaluate causal effects when there exists a natural experiment with a treated group and an untreated group. By comparing the variation over time in an outcome variable between the treated group and the untreated group, the DiD estimator can be used to calculate the effect of treatment on the outcome variable. Applications of DiD include but are not limited to studies of the effects of immigration on labor markets (Card 1990), the effects of minimum wage law on wages (Card & Krueger 1994), the effect of tariffs liberalization on corruption (Sequeira 2016), the effect of household income on children's personalities (Akee et al. 2018), and the effect of corporate tax on wages (Fuest et al. 2018).

The traditional linear DiD estimator depends on a parallel trend assumption that in the absence of treatment, the difference of outcomes between treated and untreated groups remains constant over time. In many situations, however, this assumption may not hold because there are other individual characteristics that may be associated with the variations of the outcomes. The treatment may be taken as exogenous only after controlling for these characteristics. However, as noted by Meyer et al. (1995), including control variables in the linear specification of the traditional DiD estimator imposes strong constraints on the heterogeneous effect of treatment. To address this problem, Abadie (2005) proposed the semiparametric DiD estimator. Compared to the traditional linear DiD estimator, the advantages of Abadie's estimator are threefold. First, the characteristics are treated nonparametrically so that any estimation error caused by functional specification is avoided. Second, the effect of treatment is allowed to vary among individuals, while the traditional linear DiD estimator does not allow this heterogeneity. Third, the estimation framework proposed in Abadie (2005) will enable researchers to estimate how the effect of treatment varies with changes in the characteristics.

This paper provides an orthogonal extension of Abadie's semiparametric DiD estimator (DMLDiD hereafter)[1]. Abadie's semiparametric DiD estimator behaves well when researchers use conventional nonparametric methods, such as kernel and sieve estimators, to estimate propensity score in the first-step estimation. As shown in the classical semiparametric estimation literature, Abadie's DiD estimator is $\sqrt{N}$-consistent and asymptotically normal when using kernel or sieve in the first-step estimation. However, according to the general theory of inference developed in Chernozhukov et al. (2018), these desirable properties may fail if researchers use a rich set of newly developed nonparametric estimation methods, the so-called machine learning (ML) methods, such as Lasso, Logit Lasso, random forests, boosting, neural network, and their hybrids in the first-step estimation. This is especially a problem when researchers confront a high-dimensional data set where the number of potential control variables is more than the sample size, and thus the conventional nonparametric estimation methods do not apply.

In this paper, I propose DMLDiD for three different data structures: repeated outcomes, repeated cross-sections, and multilevel treatment, which are all based on the original paper by Abadie (2005) as well as the papers on the general inference theory of ML methods by Chernozhukov et al. (2018) and Chernozhukov et al. (2016). DMLDiD will allow researchers to apply a broad set of ML methods and still obtain valid inferences. The key difference is that DMLDiD, in contrast to Abadie's original DiD estimator, is constructed based on a score function that satisfies the so-called Neyman-orthogonality (Chernozhukov et al. 2018), which is an important property for obtaining valid inference when applying ML methods. With this property, DMLDiD can overcome the bias caused by the first-step ML estimation and achieve $\sqrt{N}$-consistency and asymptotic normality as long as the ML estimator converges to its true value at a rate faster than $N^{-1/4}$. Figure 1.1 shows the Monte Carlo simulation that illustrates the negative effect of directly combining ML methods on Abadie's estimator and the benefit of using DMLDiD. The histogram in the left panel shows that the simulated distribution of Abadie's estimator is biased, while the simulated distribution of DMLDiD in

[1] The R codes can be found on my Github: https://github.com/NengChiehChang/Diff-in-Diff

the right panel is centred at the true value.

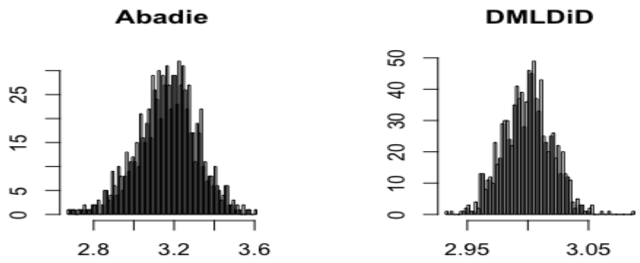


Figure 1.1: Comparison of Abadie's DiD and DMLDiD with the first-step ML estimation.

As an empirical example, I study the effect of tariff reduction on corruption using the trade data between South Africa and Mozambique during 2006 and 2014. The source of exogenous variation is the large tariff reduction on certain commodities occurring in 2008. This natural experiment was previously studied by Sequeira (2016) using the traditional linear DiD estimator. Based on Sequeira's linear specification, I include the interaction terms between the treatment and a vector of control variables. After controlling for the interaction terms, I find that the traditional linear DiD estimate becomes insignificantly different from zero. This suggests the existence of heterogeneous treatment effects, and Sequeira's result can be interpreted as a weighted average of these heterogeneous effects. As pointed out by Abadie (2005), it is ideal to treat the control variables nonparametrically when there exists heterogeneity in treatment effects, in order to avoid any inconsistency caused by functional form misspecification. I apply both Abadie's semiparametric DiD and DMLDiD on the same data set (Table 9 of Sequeira (2016)). In comparison to Sequeira's result, though with the same sign, Abadie's estimator is at least twice as large as previously reported by Sequeira (2016). This large effect, however, may be due to the lack of robustness of this estimation method and the finite sample bias in the first-step nonparametric estimation. DMLDiD removes the first-order bias and suggests a smaller effect that is closer to Sequeira's estimate. The value becomes only 60% higher than Sequeira's result. This extra effect can be explained by the misspecification of the traditional linear DiD estimator. Therefore, I obtain the same conclusion as Sequeira (2016) that tariff reduction decreases corruption, but my estimate suggests an even larger magnitude.

The DMLDiD proposed in this paper relies heavily on the recent high-dimensional and ML literature: Belloni et al. (2012), Belloni et al. (2014), Chernozhukov et al. (2015), Belloni et al. (2017), and Chernozhukov et al. (2018). This paper is also closely related to the robustness of average treatment effect estimation discussed in (Robins & Rotnitzky, 1995) and the general discussion in (Chernozhukov, Escanciano, Ichimura, & Newey, 2016). The asymptotic properties of the robust estimators discussed in these papers remain unaffected if only one of the first-step estimation with classical nonparametric method is inconsistent. In independent and contemporaneous works, Zimmert (2019), Sant'Anna & Zhao (2019), Li (2019), and Lu, Nie, & Wager (2019) also consider the orthogonal property of Abadie's DiD estimator. Zimmert (2019) further discusses its efficiency while Sant'Anna & Zhao (2019) and Li (2019) focuses on classical first-step estimation. Lu, Nie, & Wager (2019) discusses the situation where control variables are correlated with time.

Plan of the paper. Section 2 reviews both the traditional linear DiD estimator and Abadie's semiparametric DiD estimator and discusses their limitations. Section 3 presents DMLDiD and discusses its theoretical properties. Section 4 conducts the Monte Carlo simulation to shed some light on the finite sample performance of the proposed DiD estimator. Section 5 provides the empirical application, and Section 6 concludes the paper.

1.2 The Semiparametric DiD Estimator

In this section, I review the traditional linear DID estimator and Abadie's semiparametric DID estimator. Let $Y_i\left(t\right)$ be the outcome of interest for individual i at time t and $D_i\left(t\right) \in \{0,1\}$ the treatment status. The population is observed in a pre-treatment period $t = 0$, and in a post-treatment period $t = 1$. With potential outcome notations (Rubin 1974), we have $Y_i\left(t\right) = Y_i^0\left(t\right) + \left(Y_i^1\left(t\right) - Y_i^0\left(t\right)\right) D_i\left(t\right)$, where $Y_i^0\left(t\right)$ is the outcome that individual i

would attain at time t in the absence of the treatment, and $Y_i^1(t)$ represents the outcome that individual i would attain at time t if exposed to the treatment. Since individuals are only exposed to treatment at $t = 1$, we have $D_i(0) = 0$ for all i. To reduce notation, I define $D_i \equiv D_i(1)$. Then the specification for the traditional linear DiD without control variables is

$$Y_i(t) = \mu + \tau \cdot D_i + \delta \cdot t + \alpha \cdot D_i(t) + \varepsilon_i(t),$$

where α is the parameter of interest, $\varepsilon_i(t)$ is an exogenous shock that has mean zero, and (μ, τ, δ) are constant parameters. If the common trend assumption holds unconditionally, then the parameter α captures the effect of treatment. When the treated and untreated groups are thought to be unbalanced with some characteristics, researchers often include a vector of control variables, $X_i \in \mathbb{R}^d$, into the above linear specification:

$$Y_i(t) = \mu + X_i'\pi(t) + \tau \cdot D_i + \delta \cdot t + \alpha \cdot D_i(t) + \varepsilon_i(t).$$

As noted by Meyer, Viscusi, & Durbin (1995), including control variables in this linear specification may not be appropriate if the treatment has different effects for different groups in the population. One may also need to include the interaction terms between X_i and $D_i(t)$ to capture the heterogeneous effect of treatment. Hence, it is ideal to treat the control variables nonparametrically as suggested by Abadie (2005). In the following, I review Abadie's semiparametric DiD estimator.

Let the parameter of interest be the average treatment effect on the treated (ATT)

$$\theta_0 \equiv E\left[Y_i^1(1) - Y_i^0(1) \mid D_i = 1\right].$$

Abadie (2005) discussed three data types: repeated outcomes, repeated cross sections, and multi-level treatment. To avoid repetition, I only focus on the first two cases. The discussion for multilevel treatments is provided in appendix.

Case 1 (Repeated outcomes): Suppose that researchers observe both pre-treatment and post-treatment outcomes for individual of interest. That is, researchers observe $\{Y_i(0), Y_i(1), D_i, X_i\}_{i=1}^N$. In this case, we can identify the ATT under the following assumptions (Abadie 2005):

Assumption 1.1. $E\left[Y_i^0(1) - Y_i^0(0) \mid X_i, D_i = 1\right] = E\left[Y_i^0(1) - Y_i^0(0) \mid X_i, D_i = 0\right]$.

Assumption 1.2. $P(D_i = 1) > 0$ *and* $P(D_i = 1 \mid X_i) < 1$ *with probability one.*

Assumption 1.1 is the conditional parallel trend assumption. It states that conditional on individual's characteristics X_i, the average outcomes for treated and untreated groups would have followed parallel paths in the absence of treatment. Assumption 1.2 states that the support of the propensity score of the treated group is a subset of the support for the untreated. With these two assumptions, Abadie (2005) identified the ATT:

$$\theta_0 = E\left[\frac{Y_i(1) - Y_i(0)}{P(D_i = 1)} \frac{D_i - P(D_i = 1 \mid X_i)}{1 - P(D_i = 1 \mid X_i)}\right]. \tag{1.1}$$

Case 2 (Repeated cross sections): Suppose what researchers observe is repeated cross-section data. That is, researchers observe $\{Y_i, D_i, T_i, X_i\}_{i=1}^N$, where $Y_i = Y_i(0) + T_i(Y_i(1) - Y_i(0))$ and T_i is a time indicator that takes value one if the observation belongs to the post-treatment sample.

Assumption 1.3. *Conditional on* $T = 0$, *the data are i.i.d. from the distribution of* $(Y(0), D, X)$, *and conditional on* $T = 1$, *the data are i.i.d. from the distribution of* $(Y(1), D, X)$.

Supposing Assumptions 1.1-1.3 hold, the ATT is identified (Abadie 2005) as

$$\theta_0 = E\left[\frac{T_i - \lambda_0}{\lambda_0(1 - \lambda_0)} \frac{Y_i}{P(D_i = 1)} \frac{D_i - P(D_i = 1 \mid X_i)}{1 - P(D_i = 1 \mid X_i)}\right], \tag{1.2}$$

where $\lambda_0 \equiv P(T_i = 1)$.

Then the semiparametric DiD estimator would be the sample analog of 1.1 and 1.2. For example, in Case 1 in which researchers confront repeated outcomes data, the sample analog of 1.1 is

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i\left(1\right)-Y_i\left(0\right)}{\hat{p}}\frac{D_i-\hat{g}\left(X_i\right)}{1-\hat{g}\left(X_i\right)}.$$

where $\hat{p}$ is the estimator of $p_0 \equiv P\left(D=1\right)$ and $\hat{g}(X_i)$ is the estimator of the propensity score $g_0(X) \equiv P\left(D=1 \mid X\right)$. When $\hat{g}$ is estimated using classical nonparametric methods such as the kernel or series estimators, the estimator $\hat{\theta}$ is able to achieve $\sqrt{N}$-consistent and asymptotically normal under certain conditions, as shown in the semiparametric estimation literature (Newey 1994; Newey & McFadden 1994).

When $\hat{g}$ is an ML estimator, however, the estimator $\hat{\theta}$ is not necessarily to be $\sqrt{N}$-consistent in general. According to the general theory of inference of ML methods developed in Chernozhukov et al. (2018), the reason is twofold. First, the score function based on 1.1, $\varphi\left(W,\theta_0,p_0,g_0\right) \equiv \frac{Y(1)-Y(0)}{P(D=1)}\frac{D-g_0(X)}{1-g_0(X)} - \theta_0$, has a non-zero directional (Gateaux) derivative with respect to the propensity score g_0:

$$\partial_g E\left[\varphi\left(W,\theta_0,p_0,g_0\right)\right]\left[g-g_0\right] \neq 0,$$

where the directional (Gateaux) derivative is defined in Section 1.3. Second, ML estimators usually have a convergence rate slower than $N^{-1/2}$ due to regularization bias. Similarly, the estimators obtained by directly plugging ML estimators into 1.2 will not be $\sqrt{N}$-consistent in general. The Monte Carlo simulation in Section 1.4 supports this theoretical insight and reveals significant bias on the estimators based on 1.1 and 1.2 when using ML estimators in the first-step nonparametric estimation.

The next section proposes DMLDiD based on 1.1 and 1.2. A distinctive feature of DMLDiD is that the Gateaux derivatives of the score functions are zero with respect to their infinite-dimensional nuisance parameters. This property helps us remove the first-order bias of the first-step ML estimation.

1.3 The DMLDiD Estimator

In this section, I propose DMLDiD based on Abadie's results 1.1 and 1.2. In section 1.2, I present the new score functions derived from 1.1 and 1.2 and propose an algorithm to construct DMLDiD. In section 1.3.1, I show the theoretical properties of the proposed estimator.

Suppose Assumptions 1.1-1.3 hold and consider the following new score functions.

Case 1 (Repeated outcomes): The new score function for repeated outcomes is

$$\begin{aligned}\psi_1\left(W,\theta_0,p_0,\eta_{10}\right) &= \frac{Y\left(1\right)-Y\left(0\right)}{P\left(D=1\right)}\frac{D-P\left(D=1\mid X\right)}{1-P\left(D=1\mid X\right)}-\theta_0 \\ &- \underbrace{\frac{D-P\left(D=1\mid X\right)}{P\left(D=1\right)\Big(1-P\left(D=1\mid X\right)\Big)}E\left[Y\left(1\right)-Y\left(0\right)\mid X,D=0\right]}_{c_1} \quad (1.3)\end{aligned}$$

with the unknown constant $p_0 = P(D=1)$ and the infinite-dimensional nuisance parameter

$$\eta_{10}=\Big(P\left(D=1\mid X\right),E\left[Y\left(1\right)-Y\left(0\right)\mid X,D=0\right]\Big)\equiv\left(g_0,\ell_{10}\right).$$

Case 2 (Repeated cross sections): The new score function for repeated cross sections is

$$\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}\right)=\frac{T-\lambda_0}{\lambda_0\left(1-\lambda_0\right)}\frac{Y}{P\left(D=1\right)}\frac{D-P\left(D=1\mid X\right)}{1-P\left(D=1\mid X\right)}-\theta_0-c_2, \qquad (1.4)$$

where the adjustment term c_2 is

$$c_2=\frac{D-P\left(D=1\mid X\right)}{\lambda_0\left(1-\lambda_0\right)\cdot P\left(D=1\right)\cdot\Big(1-P\left(D=1\mid X\right)\Big)}\times E\left[\left(T-\lambda_0\right)Y\mid X,D=0\right].$$

The nuisance parameters are the unknown constants $p_0 = P(D=1)$ and $\lambda_0 = P(T=1)$, and the unknown function

$$\eta_{20} = \Big(P\left(D = 1 \mid X\right), E\left[\left(T - \lambda\right)Y \mid X, D = 0\right]\Big) \equiv \left(g_0, \ell_{20}\right).$$

Notice that the above the new functions are equal to the original score functions 1.1 and 1.2 plus the adjustment terms, (c_1, c_2), which have zero expectations. Thus, the new score functions 1.3 and 1.4 still identify the ATT in each case. The purpose of the adjustment terms is to make the Gateaux derivative of the new score functions zero with respect to infinite-dimensional nuisance parameters, which is the so-called Neyman-orthogonal property in Chernozhukov et al. (2018). I combine the new scores 1.3 and 1.4 with the cross-fitting algorithm of Chernozhukov et al. (2018) to propose DMLDiD.

Definition. *(a) Take a K-fold random partition $(I_k)_{k=1}^{K}$ of observation indices $[N] = \{1, ..., N\}$. For simplicity, assume that each fold I_k has the same size $n = N/K$. For each $k \in [K] = \{1, ..., K\}$, define the auxiliary sample $I_k^c \equiv \{1, ..., N\} \setminus I_k$. (b) For each k, construct the intermediate ATT estimators*

$$\tilde{\theta}_k = \frac{1}{n}\sum_{i \in I_k} \frac{D_i - \hat{g}_k\left(X_i\right)}{\hat{p}_k\left(1 - \hat{g}_k\left(X_i\right)\right)} \times \Big(Y_i\left(1\right) - Y_i\left(0\right) - \hat{\ell}_{1k}\left(X_i\right)\Big) \quad \textit{(rep-outcomes)}$$

$$\tilde{\theta}_k = \frac{1}{n}\sum_{i \in I_k} \frac{D_i - \hat{g}_k\left(X_i\right)}{\hat{p}_k\hat{\lambda}_k\left(1 - \hat{\lambda}_k\right)\left(1 - \hat{g}_k\left(X_i\right)\right)} \times \Big(\left(T_i - \hat{\lambda}_k\right)Y_i - \hat{\ell}_{2k}\left(X_i\right)\Big) \quad \textit{(rep-cross-sections)}$$

where $\hat{p}_k = \frac{1}{n}\sum_{i \in I_k^c} D_i$, $\hat{\lambda}_k = \frac{1}{n}\sum_{i \in I_k^c} T_i$, and $\left(\hat{g}_k, \hat{\ell}_{1k}, \hat{\ell}_{2k}\right)$ are the estimators of $\left(g_0, \ell_{10}, \ell_{20}\right)$ constructed using the auxiliary sample I_k^c. (c) Construct the final ATT estimator $\tilde{\theta} = \frac{1}{K}\sum_{k=1}^{K}\tilde{\theta}_k$.

The estimators $\left(\hat{g}_k, \hat{\ell}_{1k}, \hat{\ell}_{2k}\right)$ can be constructed using any ML methods or classical estimators such as kernel or series estimators. For completeness, I present the Logit Lasso and Lasso estimators here.

Consider a class of approximating functions of X_i,

$$q_i \equiv \left(q_1\left(X_i\right), ..., q_p\left(X_i\right)\right)'.$$

For example, q_i can be polynomials or B-splines. Let $\Lambda\left(u\right) \equiv 1/\left(1+\exp\left(-u\right)\right)$ be the cumulative distribution function of the standard Logistic distribution, construct the estimator of the propensity score g_0 by

$$\hat{g}_k\left(x_i\right) \equiv \Lambda\left(q_i'\hat{\beta}_k\right), \tag{3.4}$$

where

$$\hat{\beta}_k \equiv \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{M}\sum_{i\in I_k^c}\left\{-D_i(q_i'\beta)+\log\left(1+\exp\left(q_i'\beta\right)\right)\right\}+\lambda_k \parallel \beta \parallel_1$$

is the Logit Lasso estimator and $M = N - n$ is the sample size of the auxiliary sample I_k^c. Next, define M_k the sample size of $I_k^c \cap \{i : D_i = 0\}$. Construct the estimators of ℓ_{10} and ℓ_{20} by

$$\hat{\ell}_{1k}\left(x_i\right) \equiv q_i'\hat{\beta}_{1k},$$

$$\hat{\ell}_{2k}\left(x_i\right) \equiv q_i'\hat{\beta}_{2k},$$

where

$$\hat{\beta}_{1k} \in \arg\min_{\beta\in\mathbb{R}^p}\left[\frac{1}{M_k}\sum_{i\in I_k^c}\left(1-D_i\right)\left(Y_i\left(1\right)-Y_i\left(0\right)-q_i'\beta\right)^2\right]+\frac{\lambda_{1k}}{M_k}\parallel \hat{\Upsilon}_{1k}\beta \parallel_1$$

and

$$\hat{\beta}_{2k} \in \arg\min_{\beta\in\mathbb{R}^p}\left[\frac{1}{M_k}\sum_{i\in I_k^c}\left(1-D_i\right)\left(\left(T_i-\hat{\lambda}_k\right)Y_i-q_i'\beta\right)^2\right]+\frac{\lambda_{2k}}{M_k}\parallel \hat{\Upsilon}_{2k}\beta \parallel_1$$

are the modified Lasso estimators proposed in Belloni et al. (2012). The choices of the penalty levels and loadings $\left(\lambda_{1k}, \lambda_{2k}, \hat{\Upsilon}_{1k}, \hat{\Upsilon}_{2k}\right)$ suggested by Belloni et al. (2012) are provided in appendix.

1.3.1 Asymptotic Properties

In this section, I show the theoretical properties of the DMLDiD estimator $\tilde{\theta}$. In particular, I will show that the estimator $\tilde{\theta}$ can achieve $\sqrt{N}$-consistency and asymptotic normality as long as the first-step estimators converge at rates faster than $N^{-1/4}$. This rate of convergence can be achieved by many ML methods, including Lasso and Logit Lasso.

The critical difference between DMLDiD and Abadie's DiD estimator is the score functions on which they are based. The new score functions 1.3 and 1.4 have the directional (or the Gateaux) derivatives equal to zero with respect to their infinite-dimensional nuisance parameters, while the scores based on 1.1 and 1.2 do not have this property. This property is the so-called Neyman orthogonality in Chernozhukov et al. (2018).

The definition of the Neyman-orthogonal score provided here is slightly different from the definition in Chernozhukov et al. (2018). Instead of being orthogonal against all nuisance parameters, the Neyman-orthogonal score defined here is orthogonal against only those infinite-dimensional nuisance parameters. Formally, let $\theta_0 \in \Theta$ be the low-dimensional parameter of interest, ρ_0 be the true value of the finite-dimensional nuisance parameter ρ, and η_0 the true value of the infinite-dimensional nuisance parameter $\eta \in \mathcal{T}$. Suppose that W is a random element taking values in a measurable space $(\mathcal{W}, \mathcal{A}_{\mathcal{W}})$ with probability measure P. Define the directional (or the Gateaux) derivative against the infinite-dimensional nuisance parameter $D_r : \tilde{\mathcal{T}} \to \mathbb{R}$, where $\tilde{\mathcal{T}} = \{\eta - \eta_0 : \eta \in \mathcal{T}\}$,

$$D_r \left[\eta - \eta_0\right] \equiv \partial_r \left\{ E_P \left[\psi \left(W, \theta_0, \rho_0, \eta_0 + r\left(\eta - \eta_0\right)\right) \right] \right\}, \eta \in \mathcal{T},$$

for all $r \in [0, 1)$. For convenience, denote

$$\partial_\eta E_P \psi \left(W, \theta_0, \rho_0, \eta_0\right) \left[\eta - \eta_0\right] \equiv D_0 \left[\eta - \eta_0\right], \eta \in \mathcal{T}.$$

In addition, let $\mathcal{T}_N \subset \mathcal{T}$ be a nuisance realization set such that the estimator of η_0 take values in this set with high probability.

Definition. *The score ψ obeys the Neyman orthogonality condition at $(\theta_0, \rho_0, \eta_0)$ with respect to the nuisance parameter realization set $\mathcal{T}_N \subset \mathcal{T}$ if the directional derivative map $D_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in \mathcal{T}_N$ and vanishes at $r = 0$:*

$$\partial_\eta E_P \psi (W, \theta_0, \rho_0, \eta_0) [\eta - \eta_0] = 0, \textit{for all } \eta \in \mathcal{T}_N.$$

Lemma 1.1. *The new score functions 1.3 and 1.4 obey the Neyman orthogonality.*

The proof of this lemma can be found in the online appendix. In fact, it is also possible to derive the Neyman-orthogonal scores with respect to both finite- and infinite-dimensional nuisance parameters. However, the functional forms are much more complicated than the score functions 1.3 and 1.4, and this will make the corresponding estimator not as neat as the estimators proposed here. Since they will enjoy the same asymptotic properties, here I only focus on the estimators based on 1.3 and 1.4.

In the following, I will discuss the theoretical properties of the new estimator $\tilde{\theta}$ when data belongs to repeated outcomes and repeated cross sections. The results of multilevel treatment can be proven using the same arguments. Let κ and C be strictly positive constants, $K \geq 2$ be a fixed integer, and ε_N be a sequence of positive constants approaching zero. Denote by $\| \cdot \|_{P,q}$ the L^q norm of some probability measure P: $\| f \|_{P,q} \equiv (\int \mid f(w) \mid^q dP(w))^{1/q}$ and $\| f \|_{P,\infty} \equiv \sup_w \mid f(w) \mid$.

Assumption 1.4. *(Regularity Conditions for Repeated Outcomes) Let P be the probability law for $(Y(0), Y(1), D, X)$. Let $D = g_0(X) + U$ and $Y(1) - Y(0) = \ell_{10}(X) + V_1$ with $E_P[U \mid X] = 0$ and $E_P[V_1 \mid X, D = 0] = 0$. Define $G_{1p0} \equiv E_P\left[\partial_p \psi_1 (W, \theta_0, p_0, \eta_{10})\right]$ and $\Sigma_{10} \equiv E_P\left[\left(\psi_1(W, \theta_0, p_0, \eta_{10}) + G_{1p0}(D - p_0)\right)^2\right]$. For the above definition, the following conditions hold: (a) $Pr(\kappa \leq g_0(X) \leq 1 - \kappa) = 1$; (b) $\| UV_1 \|_{P,4} \leq C$; (c) $E\left[U^2 \mid X\right] \leq C$; (d) $E\left[V_1^2 \mid X\right] \leq C$; (e) $\Sigma_{10} > 0$; and (f) given the auxiliary sample I_k^c, the estimator*

$\hat{\eta}_{1k} = \left(\hat{g}_k, \hat{\ell}_{1k}\right)$ *obeys the following conditions. With probability* $1-o\left(1\right)$, $\parallel \hat{\eta}_{1k}-\eta_{10} \parallel_{P,2}\leq \varepsilon_N$, $\parallel \hat{g}_k - 1/2 \parallel_{P,\infty}\leq 1/2-\kappa$, *and* $\parallel \hat{g}_k - g_0 \parallel_{P,2}^2 + \parallel \hat{g}_k - g_0 \parallel_{P,2} \times \parallel \hat{\ell}_{1k} - \ell_{10} \parallel_{P,2}\leq \left(\varepsilon_N\right)^2$.

Assumption 1.5. *(Regularity Conditions for Repeated Cross Sections) Let* P *be the probability law for* (Y, T, D, X). *Let* $D = g_0\left(X\right) + U$ *and* $\left(T-\lambda_0\right)Y = \ell_{20}\left(X\right) + V_2$ *with* $E_p\left[U \mid X\right] = 0$ *and* $E_p\left[V_2 \mid X, D=0\right] = 0$. *Define* $G_{2p0} \equiv E_P\left[\partial_p\psi_2\left(W, \theta_0, p_0, \lambda_0, \eta_{20}\right)\right]$, $G_{2\lambda 0} \equiv E_P\left[\partial_\lambda\psi_2\left(W, \theta_0, p_0, \lambda_0, \eta_{20}\right)\right]$, *and* $\Sigma_{20} \equiv E_P[(\psi_1\left(W, \theta_0, p_0, \eta_{10}\right) + G_{2p0}\left(D-p_0\right) + G_{2\lambda 0}\left(T-\lambda_0\right))^2]$. *For the above definition, the following conditions hold: (a)* $Pr(\kappa \leq g_0\left(X\right) \leq 1-\kappa) = 1$; *(b)* $\parallel UV_2 \parallel_{P,4}\leq C$; *(c)* $E\left[U^2 \mid X\right] \leq C$; *(d)* $E\left[V_2^2 \mid X\right] \leq C$; *(e)* $E_P\left[Y^2 \mid X\right] \leq C$; *(f)* $\mid E_P\left[YU\right] \mid\leq C$; *(g)* $\Sigma_{20} > 0$; *and (h) given the auxiliary sample* I_k^c, *the estimators* $\hat{\eta}_{2k} = \left(\hat{g}_k, \hat{\ell}_{2k}\right)$ *obeys the following conditions. With probability* $1-o\left(1\right)$, $\parallel \hat{\eta}_{2k}-\eta_{20} \parallel_{P,2}\leq \varepsilon_N$, $\parallel \hat{g}_k - 1/2 \parallel_{P,\infty}\leq 1/2-\kappa$, *and* $\parallel \hat{g}_k - g_0 \parallel_{P,2}^2 + \parallel \hat{g}_k - g_0 \parallel_{P,2} \times \parallel \hat{\ell}_{2k} - \ell_{20} \parallel_{P,2}\leq \left(\varepsilon_N\right)^2$.

Theorem 1.1. *For repeated outcomes, suppose Assumptions 1.1, 1.2 and 1.4 hold. For repeated cross sections, suppose Assumptions 1.1-1.3 and 1.5 hold. If* $\varepsilon_N = o\left(N^{-1/4}\right)$, *the new ATT estimator* $\tilde{\theta}$ *obeys*

$$\sqrt{N}\left(\tilde{\theta}-\theta_0\right) \to N\left(0, \Sigma\right)$$

with $\Sigma = \Sigma_{10}$ *for repeated outcomes and* $\Sigma = \Sigma_{20}$ *for repeated cross sections.*

Theorem 1.2. *Construct the estimators of the asymptotic variances as*

$$\hat{\Sigma}_1 = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\left(\psi_1\left(W, \tilde{\theta}, \hat{p}_k, \hat{\eta}_{1k}\right) + \hat{G}_{1p}\left(D-\hat{p}_k\right)\right)^2\right] \qquad \text{(repeated outcomes)}$$

$$\hat{\Sigma}_2 = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\left(\psi_2\left(W, \tilde{\theta}, \hat{p}_k, \hat{\lambda}_k, \hat{\eta}_{2k}\right) + \hat{G}_{2p}\left(D-\hat{p}_k\right) + \hat{G}_{2\lambda}\left(T-\hat{\lambda}_k\right)\right)^2\right]$$

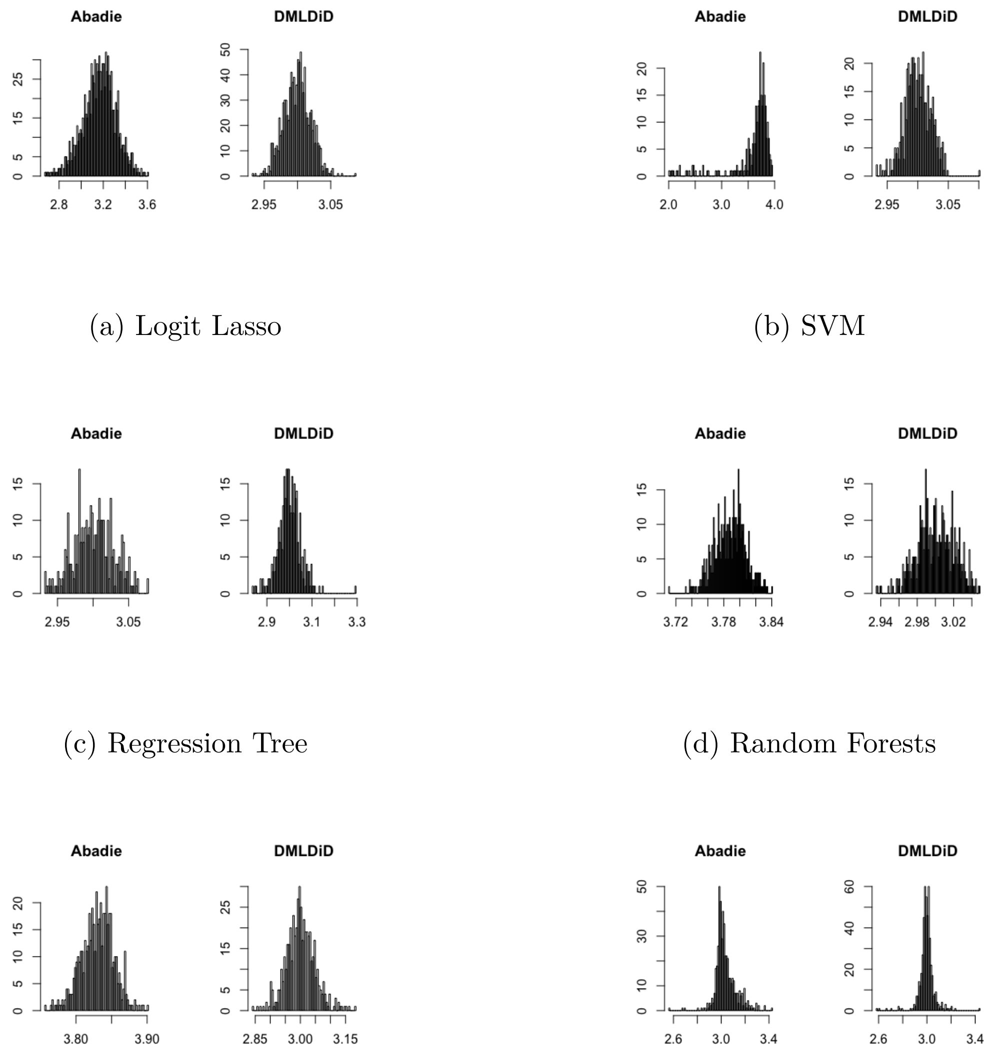
(repeated cross sections)

where $\mathbb{E}_{n,k}\left[f\left(W\right)\right] = n^{-1}\sum_{i\in I_k} f\left(W_i\right)$, $\hat{G}_{1p} = \hat{G}_{2p} = -\tilde{\theta}/\hat{p}_k$, *and* $\hat{G}_{2\lambda}$ *is a consistent estimator of* $G_{2\lambda 0}$. *If the assumptions of Theorem 1.1 hold,* $\hat{\Sigma}_1 = \Sigma_{10}+o_P\left(1\right)$ *and* $\hat{\Sigma}_2 = \Sigma_{20}+o_P\left(1\right)$.

Theorem 1.1 shows that DMLDiD $\tilde{\theta}$ can achieve $\sqrt{N}$-consistency and asymptotic normality if the first-step estimators of the infinite dimensional nuisance parameters converge at a rate faster than $N^{-1/4}$. This rate of convergence can be achieved by many ML methods. In particular, Van de Geer (2008) and Belloni et al. (2012) provided detail conditions for Logit Lasso and the modified Lasso estimators to satisfy this rate of convergence. Theorem 3.2 provides consistent estimators for the asymptotic variance of $\tilde{\theta}$. The proofs of Theorem 1.1 and Theorem 1.2 can be found in the appendix.

1.4 Simulation

In the online appendix, I conduct Monte Carlo simulations to shed some light on the finite sample properties of Abadie (2005)'s DiD estimator and the DMLDiD estimator $\tilde{\theta}$ in all three data structures: repeated outcomes, repeated cross sections, and multilevel treatment. For the first-step ML estimation, I generate high-dimensional (HD) data and estimate the propensity score by Logit Lasso, SVM, regression tree, random forests, boosting, and neural nets. I use random forests with 500 regression trees to estimate the remaining infinite-dimensional nuisance parameters. I find that while Abadie's DiD estimator suffers from the bias of a variety of ML methods, the DMLDiD estimator $\tilde{\theta}$ can successfully correct the bias and is centred at the true value. Figure 1.2 shows the Monte Carlo simulation for *repeated outcomes*. Other cases and details are provided in the online appendix.



(a) Logit Lasso

(b) SVM

(c) Regression Tree

(d) Random Forests

(e) Boosting

(f) Neural Nets

Figure 1.2: The simulation for repeated outcomes with the true value $\theta_0 = 3$.

The DGP for Repeated outcomes: Let $N = 200$ be the sample size and $p = 100$ the dimension of control variables, $X_i \sim N\left(0, I_{p\times p}\right)$. Let $\gamma_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0) \in \mathbb{R}^p$ and D_i is generated by the propensity score $P(D = 1 \mid X) = \frac{1}{1+\exp(-X'\gamma_0)}$. Also, let the potential outcomes be $Y_i^0(0) = X_i'\beta_0 + \varepsilon_1$, $Y_i^0(1) = Y_i^0(0) + 1 + \varepsilon_2$, and $Y_i^1(1) = \theta_0 + Y_i^0(1) + \varepsilon_3$, where $\beta_0 = \gamma_0 + 0.5$ and $\theta_0 = 3$, and all error terms follow $N(0, 0.1)$. Researchers observe $\{Y_i(0), Y_i(1), D_i, X_i\}$ for $i = 1, ..., N$, where $Y_i(0) = Y_i^0(0)$

and $Y_i(1) = Y_i^0(1)(1 - D_i) + Y_i^1(1) D_i$.

1.5 Empirical Example

In this example, I analyze the effect of tariffs reduction on corruption behaviors using the bribe payment data collected by Sequeira (2016) between South Africa and Mozambique. There have been theoretical and empirical debates on whether higher tariff rates increase incentives for corruption to (Clotfelter 1983; Sequeira & Djankov 2014) or lower tariffs encourage agents to pay higher bribes through an income effect (Feinstein 1991; Slemrod & Yitzhaki 2002). The former argues that an increase in the tariff rate makes it more profitable to evade taxes on the margin, while the latter asserts that an increased tariff rate makes the taxpayers less wealthy and this, under the decreasing risk aversion of being penalized, tend to reduce evasion (Allingham & Sandmo 1972).

Sequeira (2016) collected primary data on the bribed payments between the ports in Mozambique and South Africa from 2007 to 2013. The cargo owners bribed the border officials who were in charge of validating clearance documentation and collecting all tariff payments in exchange for tariff evasion. The exogenous variation used in Sequeira (2016) to study the effect of tariff reduction on corruption was the significant reduction in the average nominal tariff rate (of 5 percent) on certain products occurring in 2008. Since not all products were on the tariff reduction list, a credible control group of products is available. This credible control group allows for a DiD estimation. Sequeira (2016) pooled together the cross section data between 2007 and 2013 and estimated the effect of treatment through the traditional linear DiD with many control variables. Table 9 of Sequeira (2016) presented the result of the following specification:

$$\begin{aligned} y_{it} &= \gamma_1 TariffChangeCategory_i \times POST \\ &+ \mu POST + \beta_1 TariffChangeCategory_i \\ &+ \beta_2 BaselineTariffi + \Gamma_i + p_i + w_t + \delta_i + \epsilon_{it}, \end{aligned} \tag{1.5}$$

where y_{it} is the natural log of the amount of bribe paid for shipment i in period t, conditional on paying a bribe. $TariffChangeCategory \in \{0,1\}$ denotes the treatment status of commodities, $POST \in \{0,1\}$ is an indicator for the years following 2008, and $BaselineTariff$ is the tariff rate before the tariff reduction. The specification also includes a vector of characteristics Γ_i, and time and individual fixed effects p_i, w_t, and δ_i. The parameter γ_1 is the parameter of interest in Eq. 1.5. Sequeira (2016) found that the amount of bribe paid dropped after the tariff reduction ($\hat{\gamma}_1 = -2.928^{**}$). However, as noted by Meyer et al. (1995), this result of Equation 1.5 excludes the heterogeneous treatment effects. The estimate might be different if we take into account the heterogeneity. To shed some light on the heterogeneous treatment effect, I include the interaction terms between $TariffChangeCategory \times POST$ (TP) and the characteristics Γ_i into 1.5. The specification becomes

$$\begin{aligned} y_{it} &= \gamma_1 TariffChangeCategory_i \times POST + \gamma_2 TP_i \times \Gamma_i \\ &+ \mu POST + \beta_1 TariffChangeCategory_i \\ &+ \beta_2 BaselineTariffi + \Gamma_i + p_i + w_t + \delta_i + \epsilon_{it}, \end{aligned} \tag{1.6}$$

where γ_2 is a 10×1 vector. Table 1 shows the comparison of the estimates of 1.5 and 1.6.

Column (2) of Table 1.1 shows that (a) after controlling for the interaction terms, the estimate for γ_1 becomes insignificantly different from zero and (b) most of the coefficients of the interaction terms are negative. This suggests that there exists a large set of negative heterogeneous treatment effects and that Sequeira's estimate may be a weighted average of these heterogeneous treatment effects. The negative coefficients of the interaction terms justify the sign of Sequeira's estimate. However, it is ideal to treat the covariates nonparametrically when there exists heterogeneity in treatment effects, in order to avoid any potential inconsistency created by functional form misspecification (Abadie 2005).

Table 1.1: Estimation results of interaction.

	Eq. (5.1)	Eq. (5.2)
$\hat{\gamma}_1$	-2.928**	0.934
	(0.944)	(2.690)
$TP \times diff$		-0.986
		(0.959)
$TP \times agri$		-1.170**
		(0.580)
$TP \times lvalue$		-0.098
		(0.129)
$TP \times perishable$		0.859
		(1.213)
$TP \times largefirm$		-0.576
		(0.988)
$TP \times day\_arri$		-0.002
		(0.106)
$TP \times inspection$		-0.525
		(0.911)
$TP \times monitor$		-0.482
		(0.713)
$TP \times 2007tariff$		0.009
		(0.048)
$TP \times SouthAfrica$		-2.706***
		(0.912)

I estimate the average treatment effect on the treated using both Abadie's DiD estimator and DMLDiD. Since the data is repeated cross sections, I construct the estimators based on 1.2 and 1.4, respectively. The estimators with first-step kernel estimation contain one individual characteristic (the natural log of shipment value per ton), which is the only significant and continuous control variable in Table 9 of Sequeira (2016). The estimators with first-step Lasso estimation contain a list of the covariates included in Table 9 of Sequeira (2016), which consists of the characteristics of product, shipment, firm, and border officials. I choose both the bandwidth kernel and penalty level of Lasso by 10-fold cross-validations. Table 1.2 shows the estimation result. First, we can observe that the estimates with first-step kernel are much larger than the estimates with first-step Lasso. The reason may be that more control variables are included in the latter estimates. Second, though with the same sign, Abadie's estimator (-8.168 or -6.432) is at least twice as large as previously reported by Sequeira (2016). This large effect, however, may be due to not only the robustness of semiparametric estimation on the functional form but also the finite-sample bias in the first-step nonparametric estimation. The DMLDiD estimator (-5.222) removes the first-order bias and suggests a smaller effect that is closer to Sequeira's estimate. Its value is only 60% higher than Sequeira's result. This extra effect can be explained by the misspecification of the traditional linear DiD estimator. Therefore, I obtain the same conclusion as Sequeira (2016) that tariff reduction decreases corruption, but my estimate suggests an even larger magnitude.

Table 1.2: The results of DMLDiD estimation.

	Sequeira (2016)	Abadie (kernel)	DMLDiD (kernel)	Abadie (Lasso)	DMLDiD (Lasso)
ATT	-2.928**	-8.168**	-6.998*	-6.432**	-5.222*
	(0.944)	(3.072)	(3.752)	(2.737)	(2.647)

1.6 Conclusion

The DiD estimator survive as one of the most popular methods in the causal inference literature. A practical problem that empirical researchers face is the selection of important control variables when they confront a large number of candidate variables. Researchers may want to use ML methods to handle a rich set of control variables while taking the strength of the DiD estimator. I improve its original versions by proposing DMLDiD to allow researchers to use ML methods while still obtains valid inferences. This additional benefits will make DiD more flexible for empirical researchers to explore a broader set of popular estimation methods and analyze more types of data sets.

1.A Appendix

1.A.1 More on Estimation

Multilevel treatments: Individuals can also be exposed to different levels of treatment. Let $W \in \{0, w_1, ..., w_J\}$ be the level of treatment, where $W = 0$ denotes the untreated individuals. Researchers observe $\{Y_i(0), Y_i(1), W_i, X_i\}_{i=1}^{N}$. For $w \in \{0, w_1, ..., w_J\}$ and $t \in \{0, 1\}$, let $Y^w(t)$ be the potential outcome for treatment level w at period t. Denote the ATT for each level of treatment w by

$$\theta_0^w \equiv E\left[Y^w(1) - Y^0(1) \mid W = w\right].$$

Suppose that Assumptions (2.1) and (2.2) hold for each $w \in \{w_1, ..., w_J\}$:

$$E\left[Y_i^0(1) - Y_i^0(0) \mid X_i, W_i = w\right] = E\left[Y_i^0(1) - Y_i^0(0) \mid X_i, W_i = 0\right],$$

$P(W_i = w) > 0$ and with probability one $P(W_i = w \mid X_i) < 1$. Then we have (Abadie 2005)

$$\theta_0^w = E\left[\frac{Y(1) - Y(0)}{P(W = w)} \frac{I(W = w) \cdot P(W = 0 \mid X) - I(W = 0) \cdot P(W = w \mid X)}{P(W = 0 \mid X)}\right],$$

where $I(\cdot)$ is an indicator function. The Neyman-orthogonal score function for multilevel treatments is

$$\begin{aligned} \psi_w(W, \theta_{w0}, p_{w0}, \eta_{w0}) &= \frac{Y(1) - Y(0)}{P(W = w)} \frac{I(W = w)P(W = 0 \mid X) - I(W = 0)P(W = w \mid X)}{P(W = 0 \mid X)} \\ &- \theta_{w0} - c_w. \end{aligned} \tag{1.7}$$

The adjustment term c_w is

$$c_w = \left(\frac{I(W = w) \cdot P(W = 0 \mid X) - I(W = 0) \cdot P(W = w \mid X)}{P(W = w) \cdot P(W = 0 \mid X)} \right) \times$$
$$E[Y(1) - Y(0) \mid X, I(W = 0) = 1].$$

The nuisance parameters are the unknown constant $p_{w0} \equiv P(W = w)$ and the infinite-dimensional parameter $\eta_{w0} = (g_{w0}, g_{z0}, \ell_{30})$, where $g_{w0} = P(W = w \mid X)$, $g_{z0} = P(W = 0 \mid X)$, and $\ell_{30} = E[Y(1) - Y(0) \mid X, I(W = 0) = 1]$.

Multilevel treatments algorithm:

(i) *Take a K-fold random partition $(I_k)_{k=1}^{K}$ of observation indices $[N] = \{1, ..., N\}$ such that the size of each fold I_k is $n = N/K$. For each $k \in [K] = \{1, ..., K\}$, define the auxiliary sample $I_k^c \equiv \{1, ..., N\} \setminus I_k$.*

(ii) *For each $k \in [K]$, construct the estimator of p_0 and λ_0 by $\hat{p}_w = \frac{1}{n} \sum_{i \in I_k^c} D_i$. Also, construct the estimators of g_w, g_z, and ℓ_{30} using the auxiliary sample I_k^c: $\hat{g}_{wk} = \hat{g}_w \left((W_i)_{i \in I_k^c} \right)$, $\hat{g}_{zk} = \hat{g}_z \left((W_i)_{i \in I_k^c} \right)$, and $\hat{\ell}_{3k} = \hat{\ell}_3 \left((W_i)_{i \in I_k^c} \right)$.*

(iii) *For each k, construct the intermediate ATT estimators*

$$\tilde{\theta}_{wk} = \frac{1}{n} \sum_{i \in I_k} \frac{I(W_i = w) \cdot \hat{g}_{zk}(X_i) - I(W_i = 0) \cdot \hat{g}_{wk}(X_i)}{\hat{p}_w \hat{g}_{zk}(X_i)}$$
$$\times \left(Y(1) - Y(0) - \hat{\ell}_{3k}(X_i) \right),$$

(iv) *Construct the final ATT estimators $\tilde{\theta} = \frac{1}{K} \sum_{k=1}^{K} \tilde{\theta}_k$.*

Lasso penalty. The following is suggested by Belloni, Chen, Chernozhukov, & Hansen (2012). Let y_i denote $Y_i(1) - Y_i(0)$ or $\left(T_i - \hat{\lambda}_k \right)$, λ_k denote λ_{1k} or λ_{2k}, and $\hat{\Upsilon}_k$ denote $\hat{\Upsilon}_{1k}$

or $\hat{\Upsilon}_{2k}$. For $k \in [K]$, the loading $\hat{\Upsilon}_k$ is a diagonal matrix with entries $\hat{\gamma}_{kj}$, $j = 1, ..., p$, constructed by the following steps:

$$\text{Initial } \hat{\gamma}_{kj} = \sqrt{\frac{1}{M_k} \sum_{i \in I_k^c} (1 - D_i) q_{ij}^2 (y_i - \bar{y}_k)^2}, \lambda_k = 2c\sqrt{M_k}\Phi^{-1}(1 - \gamma/2p),$$

$$\text{Refined } \hat{\gamma}_{kj} = \sqrt{\frac{1}{M_k} \sum_{i \in I_k^c} (1 - D_i) q_{ij}^2 \hat{\varepsilon}_i^2}, \lambda_k = 2c\sqrt{M_k}\Phi^{-1}(1 - \gamma/2p),$$

where $\bar{y}_k = M^{-1} \sum_{i \in I_k^c} y_i$, $c > 1$ and $\gamma \to 0$. The empirical residual $\hat{\varepsilon}_i$ is calculated by the modified Lasso estimator β_k^* in the previous step: $\hat{\varepsilon}_i = y_i - q_i' \beta_k^*$. Repeat the second step $B > 0$ times to obtain the final loading.

1.A.2 Monte Carlo

ML Estimation (Repeated cross sections): Let N be the sample size and p the dimension of control variables, $X_i \sim N\left(0, I_{p \times p}\right)$. Also, let $\gamma_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0) \in \mathbb{R}^p$ and D is generated by the propensity score

$$P(D = 1 \mid X) = \frac{1}{1 + \exp(-X'\gamma_0)} (\text{Logistic}).$$

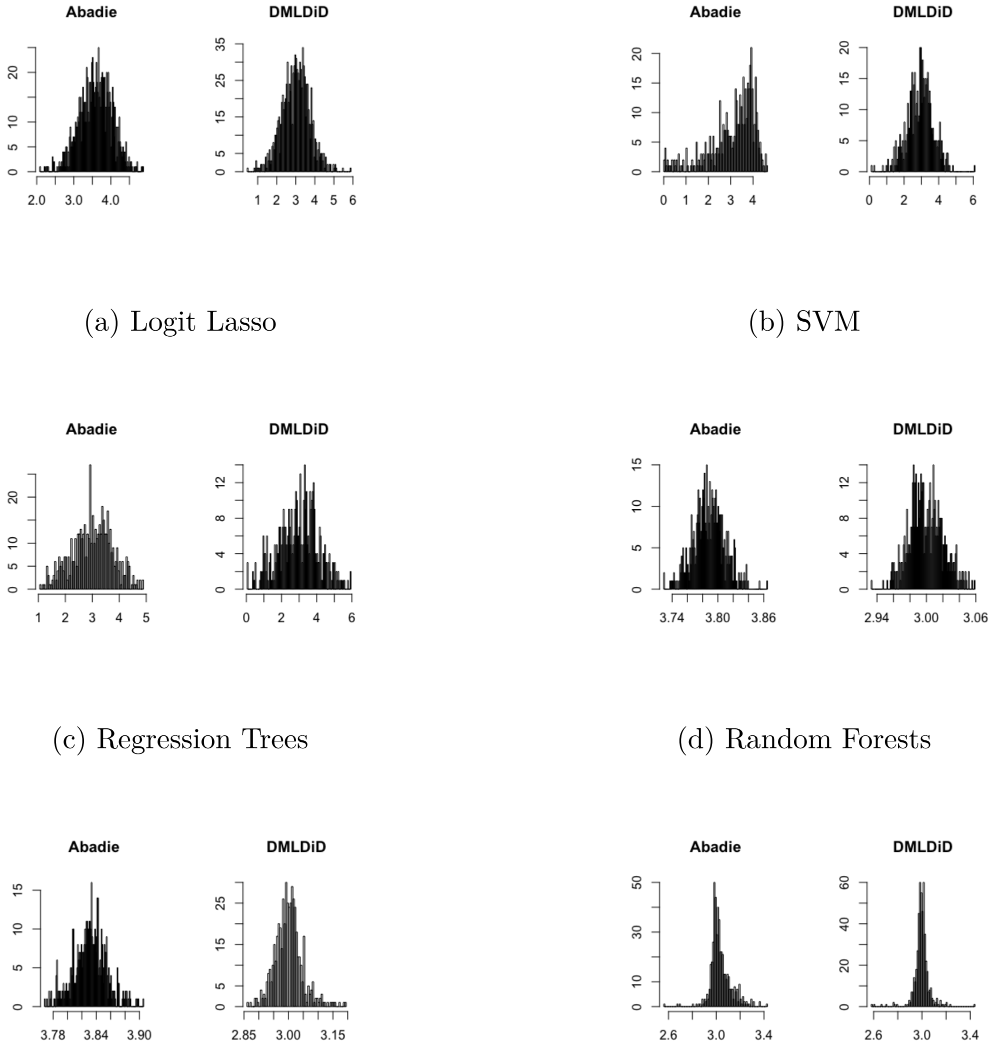
The potential outcomes are generated by $Y_i^0(0) = 1 + \varepsilon_1$, $Y_i^0(1) = Y_i^0(0) + 1 + \varepsilon_2$, $Y_i^1(1) = \theta_0 + Y_i^0(1) + \varepsilon_3$, where $\beta_0 = \gamma_0 + 0.5$ and $\theta_0 = 3$, and all error terms follow $N(0, 0.1)$. Define $Y_i(0) = Y_i^0(0)$ and $Y_i(1) = Y_i^0(1)(1 - D_i) + Y_i^1(1) D_i$. Let T_i follow a Bernoulli distribution with parameter 0.5. Researchers observe $\{Y_i, T_i, D_i, X_i\}$ for $i = 1, ..., N$, where $Y_i = Y_i(0) + T_i(Y_i(1) - Y_i(0))$.

ML Estimation (Multilevel Treatments): Suppose there are two levels of treatment so that $W \in \{0, 1, 2\}$. Let N be the sample size and p the dimension of control variables,

$X_i \sim N\left(0, I_{p\times p}\right)$. Also, let $\gamma_0 \in\in \mathbb{R}^p$ such that $\gamma_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0)$ and

$$(P(W=0), P(W=1), P(W=2)) = (0.3, 0.3, 0.4)$$

The potential outcome are generated by $Y_i^0(0) = X'\beta_0 + \varepsilon_1$, $Y_i^0(1) = Y_i^0(0) + 1 + \varepsilon_2$, $Y_i^1(1) = \theta_{10} + Y_i^0(1) + \varepsilon_3$, $Y_i^2(1) = \theta_{20} + Y_i^0(1) + \varepsilon_4$, where $\beta_0 = \gamma_0 + 0.5$ and $\theta_{10} = 3$ and $\theta_{20} = 6$, and all error terms follow $N(0, 0.1)$. Researchers observe $\{Y_i(0), Y_i(1), W_i, X_i\}$ for $i = 1, ..., N$, where $Y_i(0) = Y_i^0(0)$ and $Y_i(1) = Y_i^0(1) I(W_i = 0) + Y_i^1(1) I(W_i = 1) + Y_i^2(1) I(W_i = 2)$. I focus on the estimation of the second level ATT θ_{20}.



(a) Logit Lasso

(b) SVM

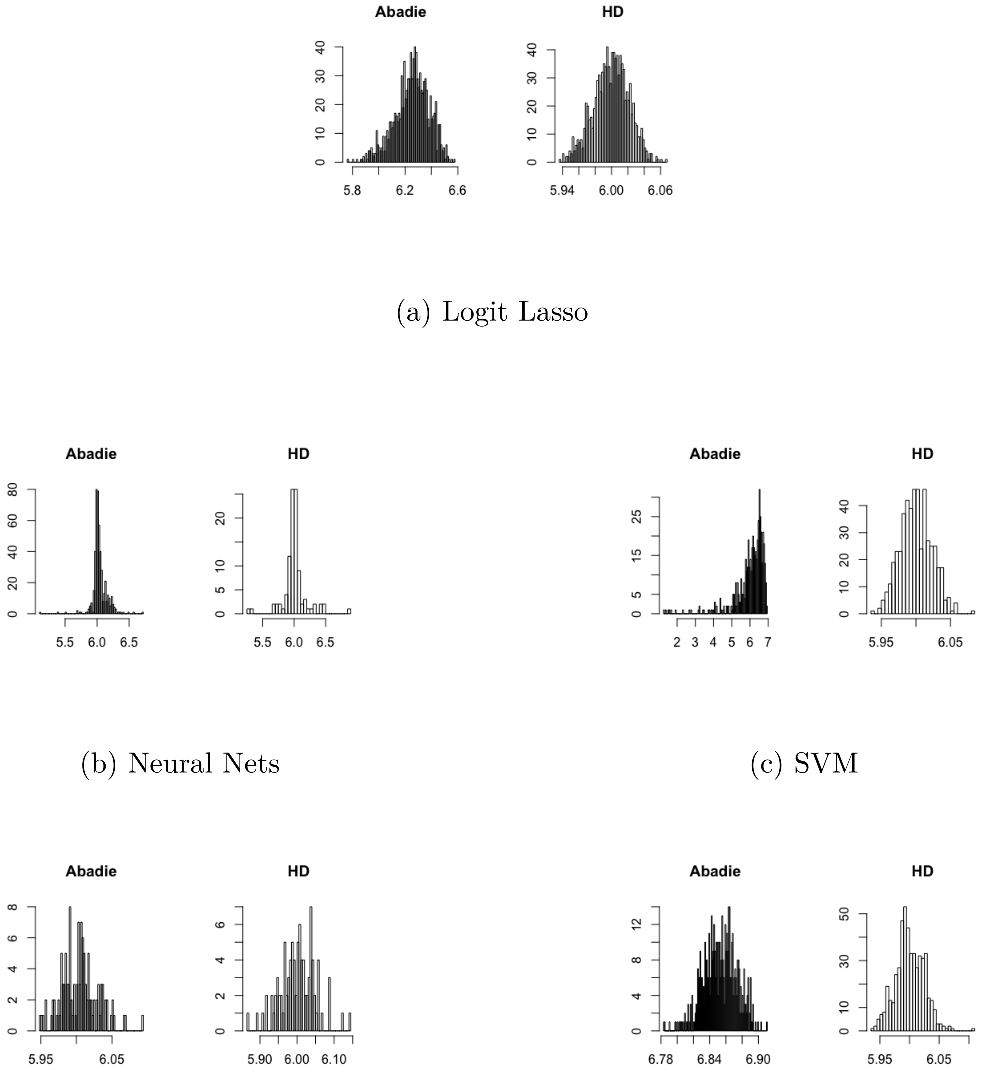
(c) Regression Trees

(d) Random Forests

(e) Boosting

(f) Neural Nets

Figure 1.3: Repeated Cross Sections



(a) Logit Lasso

(b) Neural Nets

(c) SVM

(d) Regression Trees

(e) Random Forests

Figure 1.4: Multi-level Treatment

1.A.3 Proofs of the Neyman-orthogonal Scores

Proof of Lemma 1.1

Repeated outcomes:

The Gateaux derivative of 1.3 in the direction $\eta_1 - \eta_{10} = (g - g_0, \ell_1 - \ell_{10})$ is

$$\begin{aligned}\partial_{\eta_1} E_P\left[\psi_1\left(W, \theta_0, p_0, \eta_{10}\right)\right] =& E_P\left[\frac{(D-1)\left(Y(1)-Y(0)-\ell_{10}(X)\right)}{p_0\left(1-g_0(X)\right)^2}\left(g(X)-g_0(X)\right)\right] \\ & - E_P\left[\frac{D-g_0(X)}{p_0\left(1-g_0(X)\right)}\left(\ell_1(X)-\ell_{10}(X)\right)\right] \\ =& - E_P\left[\frac{g(X)-g_0(X)}{p_0\left(1-g_0(X)\right)}E[Y(1)-Y(0)-\ell_{10}(X) \mid X, D=0]\right] \\ & - E_P\left[\frac{\left(\ell_1(X)-\ell_{10}(X)\right)}{p_0\left(1-g_0(X)\right)}E_P\left[D-g_0(X) \mid X\right]\right] \\ =& - E_P\left[\frac{g(X)-g_0(X)}{p_0\left(1-g_0(X)\right)}\left(\ell_{10}(X)-\ell_{10}(X)\right)\right]-0 \\ =& 0,\end{aligned}$$

where the second inequality follows from the law of iterated expectations, the third from the definition of $\ell_{10}(X)$ and $E_P\left[D-g_0(X) \mid X\right]=0$.

Repeated cross sections:

Similar to the proof of repeated outcomes, the Gateaux derivative of 1.4 in the direction $\eta_2 - \eta_{20} = (g - g_0, \ell_2 - \ell_{20})$ is

$$\begin{aligned}\partial_{\eta_2} E_P\left[\psi_2\left(W, \theta_0, p_0, \lambda_0, \eta_{20}\right)\right] =& E_P\left[\frac{(D-1)\left(\left(T-\lambda_0\right)Y-\ell_{20}(X)\right)}{p_0'\left(1-g_0(X)\right)^2}\left(g(X)-g_0(X)\right)\right] \\ & - E_P\left[\frac{D-g_0(X)}{p_0'\left(1-g_0(X)\right)}\left(\ell_2(X)-\ell_{20}(X)\right)\right] \\ =& - E_P\left[\frac{g(X)-g_0(X)}{p_0'\left(1-g_0(X)\right)}\left(\ell_{20}(X)-\ell_{20}(X)\right)\right] \\ & - E_P\left[\frac{\ell_2(X)-\ell_{20}(X)}{p\lambda(1-\lambda)(1-g(X))}E_P\left[D-g_0(X) \mid X\right]\right] \\ =& 0,\end{aligned}$$

where $p_0' \equiv p_0\lambda_0\left(1-\lambda_0\right)$.

Multilevel treatment:

Let $\Delta_w = g_w - g_{w0}$, $\Delta_z = g_z - g_{z0}$, and $\Delta_{\ell 3} = \ell_3 - \ell_{30}$. The Gateaux derivative of 1.7 in the direction $\eta_w - \eta_{w0} = (g_w - g_{w0}, g_z - g_{z0}, \ell_3 - \ell_{30})$ is

$$\begin{aligned}\partial_{\eta_w} E_P \left[\psi_w \left(W, \theta_0, p_{w0}, \eta_{w0}\right)\right] =& E_P \left[\frac{I\left(W=0\right) g_{w0}\left(X\right)}{p_{w0} g_{z0}\left(X\right)^2} \left(Y\left(1\right) - Y\left(0\right) - \ell_{30}\right) \Delta_w\right] \\ & - E_P \left[\frac{I\left(W=0\right)}{p_{w0} g_{z0}\left(X\right)} \left(Y\left(1\right) - Y\left(0\right) - \ell_{30}\right) \Delta_z\right] \\ & + E_P \left[\frac{I\left(W=0\right) g_{w0}\left(X\right) - I\left(W=w\right) g_{z0}\left(X\right)}{p_{w0} g_{z0}\left(X\right)} \Delta_{\ell 3}\right] \\ =& 0\end{aligned}$$

by the law of iterated expectation on each terms. □

1.A.4 Additional proofs

Proof of Theorem 1.1:

The proof proceeds in five steps. In Step 1, I show the main result using the auxiliary results (A.1)-(A.4). In Step 2-5, I prove the auxiliary results.

$$\sup_{\eta_1 \in \mathcal{T}_N} \left(E\left[\| \psi_1\left(W, \theta_0, p_0, \eta_1\right) - \psi_1\left(W, \theta_0, p_0, \eta_{10}\right) \|^2\right]\right)^{1/2} \leq \varepsilon_N, \tag{A.1}$$

$$\sup_{r \in (0,1), \eta_1 \in \mathcal{T}_N} \| \partial_r^2 E\left[\psi_1\left(W, \theta_0, p_0, \eta_{10} + r\left(\eta_1 - \eta_{10}\right)\right)\right] \| \leq \left(\varepsilon_N\right)^2, \tag{A.2}$$

$$\sup_{\eta_1 \in \mathcal{T}_N} \left(E_P\left[\| \partial_p \psi_1\left(W, \theta_0, p_0, \eta_1\right) - \partial_p \psi_1\left(W, \theta_0, p_0, \eta_{10}\right) \|^2\right]\right)^{1/2} \leq \varepsilon_N, \tag{A.3}$$

$$\sup_{p \in \mathcal{P}_N, \eta_1 \in \mathcal{T}_N} \left(E_P\left[\| \partial_p^2 \psi_1\left(W, \theta_0, p, \eta_1\right) - \partial_p^2 \psi_1\left(W, \theta_0, p_0, \eta_{10}\right) \|^2\right]\right)^{1/2} \leq \varepsilon_N, \tag{A.4}$$

where $\mathcal{T}_N$ is the set of all $\eta_1 = (g, \ell_1)$ consisting of P-square-integrable functions g and ℓ_1

such that

$$\parallel \eta_1 - \eta_{10} \parallel_{P,2} \leq \varepsilon_N,$$

$$\parallel g - 1/2 \parallel_{P,\infty} \leq 1/2 - \kappa,$$

$$\parallel g - g_0 \parallel_{P,2}^2 + \parallel g - g_0 \parallel_{P,2} \times \parallel \ell_1 - \ell_{10} \parallel_{P,2} \leq (\varepsilon_N)^2,$$

and $\mathcal{P}_N$ is the set of all $p > 0$ such that $\mid p - p_0 \mid \leq N^{-1/2}$. Then by Assumption 1.4 and $\mid \hat{p}_k - p_0 \mid = O_P\left(N^{-1/2}\right)$, we have $\hat{\eta}_{1k} \in \mathcal{T}_N$ and $\hat{p}_k \in \mathcal{P}_N$ with probability $1 - o(1)$.

Step 1. Observe that we have the decomposition

$$\begin{aligned}
\sqrt{N}\left(\tilde{\theta} - \theta_0\right) =& \sqrt{N}\left(\frac{1}{K}\sum_{k=1}^{K}\tilde{\theta}_k - \theta_0\right) \\
=& \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi_1\left(W, \theta_0, \hat{p}_k, \hat{\eta}_{1k}\right)\right] \\
=& \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi_1\left(W, \theta_0, p_0, \hat{\eta}_{1k}\right)\right] \\
& + \underbrace{\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\partial_p\psi_1\left(W, \theta_0, p_0, \hat{\eta}_{1k}\right)\right]\left(\hat{p}_k - p_0\right)}_{a} \\
& + \underbrace{\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\partial_p^2\psi_1\left(W, \theta_0, \bar{p}_k, \hat{\eta}_{1k}\right)\right]\left(\hat{p}_k - p_0\right)^2,}_{b}
\end{aligned}$$

where $\bar{p}_k \in (\hat{p}_k, p_0)$. By the triangle inequality, the expectation in term (a) satisfies

$$\left|\mathbb{E}_{n,k}\left[\partial_p\psi_1\left(W, \theta_0, p_0, \hat{\eta}_{1k}\right)\right] - E_P\left[\partial_p\psi_1\left(W, \theta_0, p_0, \eta_{10}\right)\right]\right| \leq J_{1,k} + J_{2,k},$$

where

$$J_{1,k} = \left|\mathbb{E}_{n,k}\left[\partial_p\psi_1\left(W, \theta_0, p_0, \hat{\eta}_{1k}\right)\right] - \mathbb{E}_{n,k}\left[\partial_p\psi_1\left(W, \theta_0, p_0, \eta_{10}\right)\right]\right|,$$

$$J_{2,k} = \left|\mathbb{E}_{n,k}\left[\partial_p\psi_1\left(W, \theta_0, p_0, \eta_{10}\right)\right] - E_p\left[\partial_p\psi_1\left(W, \theta_0, p_0, \eta_{10}\right)\right]\right|.$$

The goal is to show that $J_{1,k} = o_p(1)$ and $J_{2,k} = o_p(1)$. To bound $J_{2,k}$, we have $E_P\left[J_{2,k}\right] = 0$ and

$$\begin{aligned}
E_P\left[J_{2,k}^2\right] \leq & n^{-1} E_P\left[\left(\partial_p \psi_1\left(W, \theta_0, p_0, \eta_{10}\right)^2\right)\right] \\
= & n^{-1} E_P\left[\frac{1}{p_0^4} \frac{U^2 V_1^2}{\left(1-g_0\right)^2}\right] \\
\leq & n^{-1}\left(\frac{C^2}{p_0^4 \kappa^2}\right),
\end{aligned}$$

where the last inequality follows from Assumption (3.1). By Chebyshev's inequality, $J_{2,k} = O_P\left(n^{-1/2}\right) = o_P(1)$. Next, we bound $J_{1,k}$. Conditional on the auxiliary sample I_k^c, $\hat{\eta}_{1k}$ can be treated as fixed. Under the event that $\hat{\eta}_{1k} \in \mathcal{T}_N$, we have

$$\begin{aligned}
E_P\left[J_{1,k}^2 \mid \left(W_i\right)_{i \in I_k^c}\right] = & E_P\left[\| \partial_p \psi_1\left(W, \theta_0, p_0, \hat{\eta}_{1k}\right) - \partial_p \psi_1\left(W, \theta_0, p_0, \eta_{10}\right) \|^2 \mid \left(W_i\right)_{i \in I_k^c}\right] \\
\leq & \sup_{\eta_1 \in \mathcal{T}_N} E_P\left[\| \partial_p \psi_1\left(W, \theta_0, p_0, \eta_1\right) - \partial_p \psi_1\left(W, \theta_0, p_0, \eta_{10}\right) \|^2\right] \\
= & \varepsilon_N^2
\end{aligned}$$

by (A.3). Since conditional convergence implies unconditional convergence (Lemma A.1), $J_{1,k} = O_P\left(\varepsilon_N\right) = o_P(1)$. Together, we have

$$\mathbb{E}_{n,k}\left[\partial_p \psi_1\left(W, \theta_0, p_0, \hat{\eta}_{1k}\right)\right] \xrightarrow{p} E_p\left[\partial_p \psi_1\left(W, \theta_0, p_0, \eta_{10}\right)\right] = G_{1p0}.$$

By the triangle inequality again, the expectation in term (b) satisfies

$$\left|\mathbb{E}_{n,k}\left[\partial_p^2 \psi_1\left(W, \theta_0, \bar{p}_k, \hat{\eta}_{1k}\right)\right] - E_p\left[\partial_p^2 \psi_1\left(W, \theta_0, p_0, \eta_{10}\right)\right]\right| \leq J_{3,k} + J_{4,k},$$

where

$$J_{3,k} = \left|\mathbb{E}_{n,k}\left[\partial_p^2 \psi_1\left(W, \theta_0, \bar{p}_k, \hat{\eta}_{1k}\right)\right] - \mathbb{E}_{n,k}\left[\partial_p^2 \psi_1\left(W, \theta_0, p_0, \eta_{10}\right)\right]\right|,$$

$$J_{4,k} = \left| \mathbb{E}_{n,k} \left[\partial_p^2 \psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \right] - E_p \left[\partial_p^2 \psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \right] \right| .$$

To bound $J_{4,k}$, we have

$$\begin{aligned} E_P \left[J_{4,k}^2 \right] \leq & n^{-1} E_P \left[\left(\partial_p^2 \psi_1 \left(W, \theta_0, p_0, \eta_{10} \right)^2 \right) \right] \\ = & n^{-1} E_P \left[\frac{4}{p_0^6} \frac{U^2 V_1^2}{\left(1 - g_0 \right)^2} \right] \\ \leq & n^{-1} \left(\frac{4C^2}{p_0^6 \kappa^2} \right), \end{aligned}$$

where the last inequality follows from the regularity conditions. By Chebyshev's inequality, $J_{4,k} = O_P \left(n^{-1/2} \right) = o_P \left(1 \right)$. Conditional on I_k^c, both $\bar{p}_k$ and $\hat{\eta}_{1k}$ can be treated as fixed. Under the event that $\hat{p}_k \in \mathcal{P}_N$ (thus $\bar{p}_k \in \mathcal{P}_N$) and $\hat{\eta}_{1k} \in \mathcal{T}_N$, we have

$$\begin{aligned} E_P \left[J_{3,k}^2 \mid (W_i)_{i \in I_k^c} \right] = & E_P \left[\| \, \partial_p^2 \psi_1 \left(W, \theta_0, \bar{p}_k, \hat{\eta}_{1k} \right) - \partial_p^2 \psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \|^2 | \, (W_i)_{i \in I_k^c} \right] \\ \leq & \sup_{p \in \mathcal{P}_N, \eta_1 \in \mathcal{T}_N} E_P \left[\| \, \partial_p \psi_1 \left(W, \theta_0, p, \eta_1 \right) - \partial_p \psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \|^2 \right] \\ \leq & \varepsilon_N^2 \end{aligned}$$

by (A.4). By Lemma A.1 again, $J_{3,k} = O_P \left(\varepsilon_N \right) = o_P \left(1 \right)$. Hence,

$$\mathbb{E}_{n,k} \left[\partial_p^2 \psi_1 \left(W, \theta_0, \bar{p}_k, \hat{\eta}_{1k} \right) \right] \xrightarrow{p} E_P \left[\partial_p^2 \psi_1 \left(W, \theta_0, \bar{p}_k, \hat{\eta}_{1k} \right) \right] .$$

Combine the above results with that $\hat{p}_k - p_0 = \mathbb{E}_{n,k}[D - p_0]$ and $(\hat{p}_k - p_0)^2 = O_P(N^{-1})$, the decomposition of $\tilde{\theta}$ becomes

$$\begin{aligned}\sqrt{N}\left(\tilde{\theta}-\theta_0\right) &= \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi_1\left(W,\theta_0,p_0,\hat{\eta}_{1k}\right)\right] \\ &+ \left[\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}G_{1p0}\mathbb{E}_{n,k}\left[\left(D-p_0\right)\right]+o_p\left(1\right)\right]+O_p\left(N^{-1/2}\right) \\ &= \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi_1\left(W,\theta_0,p_0,\hat{\eta}_{1k}\right)+G_{1p0}\left(D-p_0\right)\right]+o_P\left(1\right) \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\psi_1\left(W_i,\theta_0,p_0,\eta_{10}\right)+G_{1p0}\left(D_i-p_0\right)\right]+\sqrt{N}R_N+o_P\left(1\right),\end{aligned}$$

where

$$\begin{aligned}R_N =&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi_1\left(W,\theta_0,p_0,\hat{\eta}_{1k}\right)+G_{1p0}\left(D-p_0\right)\right] \\ &-\frac{1}{N}\sum_{i=1}^{N}\left[\psi_1\left(W_i,\theta_0,p_0,\eta_{10}\right)+G_{1p0}\left(D_i-p_0\right)\right] \\ =&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi_1\left(W,\theta_0,p_0,\hat{\eta}_{1k}\right)\right]-\frac{1}{N}\sum_{i=1}^{N}\psi_1\left(W_i,\theta_0,p_0,\eta_{10}\right).\end{aligned}$$

It remains to show that $\sqrt{N}R_N = o_P\left(1\right)$.

This part is essentially identical to Step 3 in the proof of Theorem 3.1 (DML2) in Chernozhukov et al. (2018). I reproduce it here for reader's convenience. Since K is a fixed integer, which is independent of N, it suffices to show that for any $k \in [K]$,

$$\mathbb{E}_{n,k}\left[\psi_1\left(W,\theta_0,p_0,\hat{\eta}_{1k}\right)\right]-\frac{1}{n}\sum_{i\in I_k}\psi_1\left(W_i,\theta_0,p_0,\eta_{10}\right)=o_P\left(N^{-1/2}\right).$$

Define the empirical process notation:

$$\mathbb{G}_{n,k}\left[\phi\left(W\right)\right]=\frac{1}{\sqrt{n}}\sum_{i\in I_k}\left(\phi\left(W_i\right)-\int\phi\left(w\right)dP\right),$$

where ϕ is any P-integrable function on $\mathcal{W}$. By the triangle inequality, we have

$$\| \mathbb{E}_{n,k} \left[\psi_1 \left(W, \theta_0, p_0, \hat{\eta}_{1k} \right) \right] - \frac{1}{n} \sum_{i \in I_k} \psi_1 \left(W_i, \theta_0, p_0, \eta_{10} \right) \| \leq \frac{I_{1,k} + I_{2,k}}{\sqrt{n}},$$

where

$$I_{1,k} \equiv \| \mathbb{G}_{n,k} \left[\psi_1 \left(W, \theta_0, p_0, \hat{\eta}_{1k} \right) \right] - \mathbb{G}_{n,k} \left[\psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \right] \|,$$

$$I_{2,k} \equiv \sqrt{n} \parallel E_P \left[\psi_1 \left(W, \theta_0, p_0, \hat{\eta}_{1k} \right) \mid \left(W_i \right)_{i \in I_k^c} \right] - E_P \left[\psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \right] \parallel .$$

To bound $I_{1,k}$, note that conditional on $\left(W_i \right)_{i \in I_k^c}$ the estimator $\hat{\eta}_{1k}$ is nonstochastic. Under the event that $\hat{\eta}_{1k} \in \mathcal{T}_N$, we have

$$\begin{aligned} E_P \left[I_{1,k}^2 \mid \left(W_i \right)_{i \in I_k^c} \right] =& E_P \left[\| \psi_1 \left(W, \theta_0, p_0, \hat{\eta}_{1k} \right) - \psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \|^2 \mid \left(W_i \right)_{i \in I_k^c} \right] \\ & \leq \sup_{\eta_1 \in \mathcal{T}_N} E_P \left[\| \psi_1 \left(W, \theta_0, p_0, \eta_1 \right) - \psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \|^2 \mid \left(W_i \right)_{i \in I_k^c} \right] \\ & = \sup_{\eta_1 \in \mathcal{T}_N} E_P \left[\| \psi_1 \left(W, \theta_0, p_0, \eta_1 \right) - \psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \|^2 \right] \\ & = \left(\varepsilon_N \right)^2 \end{aligned}$$

by (A.1). Hence, $I_{1,k} = O_P \left(\varepsilon_N \right)$ by Lemma A.1. To bound $I_{2,k}$, define the following function

$$f_k \left(r \right) = E_P \left[\psi_1 \left(W, \theta_0, p_0, \eta_{10} + r \left(\hat{\eta}_{1k} - \eta_{10} \right) \right) \mid \left(W_i \right)_{i \in I_k^c} \right] - E \left[\psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \right]$$

for $r \in [0, 1)$. By Taylor series expansion, we have

$$f_k \left(1 \right) = f_k \left(0 \right) + f_k' \left(0 \right) + f_k'' \left(\tilde{r} \right) / 2, \text{for some } \tilde{r} \in \left(0, 1 \right).$$

Note that $f_k \left(0 \right) = 0$ since $E \left[\psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \mid \left(W_i \right)_{i \in I_k^c} \right] = E \left[\psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \right]$. Further, on the event $\hat{\eta}_{1k} \in \mathcal{T}_N$,

$$\| f_k' \left(0 \right) \| = \| \partial_{\eta_1} E_P \psi_1 \left(W, \theta_0, p_0, \eta_{10} \right) \left[\hat{\eta}_{1k} - \eta_{10} \right] \| = 0$$

by the orthogonality of ψ_1. Also, on the event $\hat{\eta}_{1k} \in \mathcal{T}_N$,

$$\parallel f_k''(\tilde{r}) \parallel \leq \sup_{r\in(0,1)} \parallel f_k''(r) \parallel \leq (\varepsilon_N)^2$$

by (A.2). Thus,

$$I_{2,k} = \sqrt{n} \parallel f_k(1) \parallel = O_P\left(\sqrt{n}\,(\varepsilon_N)^2\right).$$

Together with the result on $I_{1,k}$, we have

$$\begin{aligned}\mathbb{E}_{n,k}\left[\psi_1\left(W,\theta_0,p_0,\hat{\eta}_{1k}\right)\right] - \frac{1}{n}\sum_{i\in I_k}\psi_1\left(W_i,\theta_0,p_0,\eta_{10}\right) \leq & \frac{I_{1,k}+I_{2,k}}{\sqrt{n}} \\ = & O_P\left(n^{-1/2}\varepsilon_N + (\varepsilon_N)^2\right) \\ = & o_P\left(N^{-1/2}\right)\end{aligned}$$

by $n = O(N)$ and $\varepsilon_N = o\left(N^{-1/4}\right)$. Hence, $\sqrt{N}R_N = o_P(1)$.

Step 2. In this step, I present the proof of (A.1). We have the following decomposition:

$$\begin{aligned}\psi_1\left(W,\theta_0,p_0,\eta_1\right) - \psi_1\left(W,\theta_0,p_0,\eta_{10}\right) = & \frac{D-g(X)}{p_0\left(1-g(X)\right)}\left(Y(1)-Y(0)-\ell_1(X)\right) \\ & - \frac{D-g_0(X)}{p_0\left(1-g_0(X)\right)}\left(Y(1)-Y(0)-\ell_{10}(X)\right) \\ = & \frac{U+g_0(X)-g(X)}{p_0\left(1-g(X)\right)}\left(V_1+\ell_{10}(X)-\ell_1(X)\right) \\ & - \frac{UV_1}{p_0\left(1-g_0(X)\right)}.\end{aligned}$$

Thus, we have

$$\begin{aligned}\psi_1\left(W,\theta_0,p_0,\eta_1\right) - \psi_1\left(W,\theta_0,p_0,\eta_{10}\right) = & \frac{UV_1}{p_0\left(1-g(X)\right)} + \frac{U\left(\ell_{10}(X)-\ell_1(X)\right)}{p_0\left(1-g(X)\right)} \\ & + \frac{\left(g_0(X)-g(X)\right)V_1}{p_0\left(1-g(X)\right)} - \frac{UV_1}{p_0\left(1-g_0(X)\right)} \\ & + \frac{\left(g_0(X)-g(X)\right)\left(\ell_{10}(X)-\ell_1(X)\right)}{p_0\left(1-g(X)\right)}.\end{aligned}$$

Given $\kappa \leq g_0(X) \leq 1-\kappa$ and $\kappa \leq g(X) \leq 1-\kappa$,

$$\begin{aligned}\parallel \psi_1(W,\theta_0,p_0,\eta_1) - \psi_1(W,\theta_0,p_0,\eta_{10}) \parallel_{P,2} \leq & \frac{1}{p_0\kappa^2} \parallel UV_1(1-g_0(X)) \\ & + U(\ell_{10}(X) - \ell_1(X))(1-g_0(X)) \\ & + V_1(g_0(X) - g(X))(1-g_0(X)) \\ & + (g_0 - g)(\ell_{10} - \ell_1)(1-g_0(X)) \\ & - UV_1(1-g(X)) \parallel_{P,2} .\end{aligned}$$

By $\kappa \leq g_0(X) \leq 1-\kappa$ and $\kappa \leq g(X) \leq 1-\kappa$ again, we can obtain

$$\begin{aligned}\parallel \psi_1(W,\theta_0,p_0,\eta_1) - \psi_1(W,\theta_0,p_0,\eta_{10}) \parallel_{P,2} \leq & \frac{1-\kappa}{p_0\kappa^2} \parallel UV_1 + U(\ell_{10}(X) - \ell_1(X)) \\ & + V_1(g_0(X) - g(X)) \\ & + (g_0(X) - g(X))(\ell_{10}(X) - \ell_1(X)) \\ & - UV_1 \parallel_{P,2} .\end{aligned}$$

Thus, by $E_P\left[U^2 \mid X\right] \leq C$ and $E_P\left[V_1^2 \mid X\right] \leq C$,

$$\begin{aligned}\parallel \psi_1(W,\theta_0,p_0,\eta_1) - \psi_1(W,\theta_0,p_0,\eta_{10}) \parallel_{P,2} \leq & \frac{(1-\kappa)\sqrt{C}}{p_0\kappa^2} \parallel \ell_{10} - \ell_1 \parallel_{P,2} \\ & + \frac{(1-\kappa)\sqrt{C}}{p_0\kappa^2} \parallel g_0 - g \parallel_{P,2} \\ & + \frac{(1-\kappa)}{p_0\kappa^2} \parallel g_0 - g \parallel_{P,2} \parallel \ell_{10} - \ell_1 \parallel_{P,2} \\ \leq & O\left(\varepsilon_N + \varepsilon_N + (\varepsilon_N)^2\right) \\ = & O(\varepsilon_N) .\end{aligned}$$

Step 3. In this step, I present the proof of (A.2). Define

$$f(r) = E_P\left[\psi_1\left(W, \theta_0, p_0, \eta_{10} + r\left(\eta_1 - \eta_{10}\right)\right)\right].$$

Then its second-order derivative is

$$\begin{aligned}\partial_r^2 f(r) =& \frac{2}{p_0} E_P\left[\frac{(D-1)\left(g-g_0\right)^2}{\left(1-g_0-r\left(g-g_0\right)\right)^3}\left(Y(1)-Y(0)-\ell_{10}-r\left(\ell_1-\ell_{10}\right)\right)\right] \\ & -\frac{2}{p_0} E_P\left[\frac{D-1}{\left(1-g_0-r\left(g-g_0\right)\right)^2}\left(\ell_1-\ell_{10}\right)\left(g-g_0\right)\right].\end{aligned}$$

It follows that

$$\mid \partial_r^2 f(r) \mid \leq O\left(\| \left(g-g_0\right) \|_{P,2}^2 + \| \left(g-g_0\right) \|_{P,2} \times \| \left(\ell_1-\ell_{10}\right) \|_{P,2}\right) \leq \left(\varepsilon_N\right)^2.$$

Step 4. Notice that

$$\begin{aligned}\partial_p \psi_1\left(W, \theta, p, \eta_1\right) &= -\frac{1}{p} \frac{D-g(X)}{1-g(X)}\left(Y(1)-Y(0)-\ell_1(X)\right) \\ &= -\frac{1}{p}\left(\psi_1\left(W, \theta, p, \eta_1\right)+\theta\right),\end{aligned}$$

then we have

$$\begin{aligned}& \| \partial_p \psi_1\left(W, \theta_0, p_0, \eta_1\right) - \partial_p \psi_1\left(W, \theta_0, p_0, \eta_{10}\right) \|_{P,2} \\ &= \frac{1}{p_0} \| \psi_1\left(W, \theta_0, p_0, \eta_1\right) - \psi_1\left(W, \theta_0, p_0, \eta_{10}\right) \|_{P,2} \\ &= O\left(\varepsilon_N\right)\end{aligned}$$

by Step 2.

Step 5. Notice that

$$\begin{aligned}\partial_p^2\psi_1\left(W,\theta,p,\eta_1\right) =& \frac{2}{p^3}\frac{D-g\left(X\right)}{1-g\left(X\right)}\left(Y\left(1\right)-Y\left(0\right)-\ell_1\left(X\right)\right)\\ =& \frac{2}{p^2}\left(\psi_1\left(W,\theta,p,\eta_1\right)+\theta\right),\end{aligned}$$

then we have

$$\begin{aligned}\partial_p^2\psi_1\left(W,\theta_0,p,\eta_1\right)-\partial_p^2\psi_1\left(W,\theta_0,p_0,\eta_{10}\right) =& \partial_p^2\psi_1\left(W,\theta_0,p_0,\eta_1\right)-\partial_p^2\psi_1\left(W,\theta_0,p_0,\eta_{10}\right)\\ &+\partial_{p^3}^3\psi_1\left(W,\theta_0,\bar{p},\eta_1\right)\left(p-p_0\right)\\ =& \frac{2}{p_0^2}\left(\psi_1\left(W,\theta_0,p_0,\eta_1\right)-\psi_1\left(W,\theta_0,p_0,\eta_{10}\right)\right)\\ &-\frac{6}{\bar{p}^4}\frac{\left(D-g\left(X\right)\right)\left(Y\left(1\right)-Y\left(0\right)-\ell_1\left(X\right)\right)}{1-g\left(X\right)}\\ &\times\left(p-p_0\right),\end{aligned}$$

where $\bar{p}\in\left(p,p_0\right)$. Thus, $\parallel\partial_p^2\psi_1\left(W,\theta_0,p,\eta_1\right)-\partial_p^2\psi_1\left(W,\theta_0,p_0,\eta_{10}\right)\parallel_{P,2}$ is bounded by

$$\begin{aligned}&\frac{2}{p_0^2}\parallel\psi_1\left(W,\theta_0,p_0,\eta_1\right)-\psi_1\left(W,\theta_0,p_0,\eta_{10}\right)\parallel_{P,2}\\ &+\parallel\frac{6}{\bar{p}^4}\frac{D-g\left(X\right)}{1-g\left(X\right)}\left(Y\left(1\right)-Y\left(0\right)-\ell_1\left(X\right)\right)\parallel_{P,2}\times\mid p-p_0\mid.\end{aligned}$$

The term in the second line is bounded by

$$\begin{aligned}\frac{6}{\bar{p}^4\kappa}\parallel(U+g_0-g)(V_1+\ell_{10}-\ell_1)\parallel_{P,2}\leq&\frac{6}{\bar{p}^4\kappa}\parallel UV_1\parallel_{P,2}+\frac{6}{\bar{p}^4\kappa}\parallel U(\ell_{10}-\ell_1)\parallel_{P,2}\\&+\frac{6}{\bar{p}^4\kappa}\parallel V_1(g_0-g)\parallel_{P,2}\\&+\frac{6}{\bar{p}^4\kappa}\parallel g_0-g\parallel_{P,2}\parallel\ell_{10}-\ell_1\parallel_{P,2}\\\leq&\frac{6}{\bar{p}^4\kappa}\left(C+\sqrt{C}\parallel\ell_{10}-\ell_1\parallel_{P,2}\right)\\&+\frac{6}{\bar{p}^4\kappa}\sqrt{C}\parallel g_0-g\parallel_{P,2}\\&+\frac{6}{\bar{p}^4\kappa}\parallel g_0-g\parallel_{P,2}\parallel\ell_{10}-\ell_1\parallel_{P,2}\\=&O(1)\end{aligned}$$

by $\parallel UV_1\parallel_{P,2}\leq\parallel UV_1\parallel_{P,4}\leq C$, $E_P\left[U^2\mid X\right]\leq C$, $E_P\left[V_1^2\mid X\right]\leq C$, and the conditions on the rates of convergence. Together with Step 2, we obtain

$$\begin{aligned}\parallel\partial_p^2\psi_1(W,\theta_0,p,\eta_1)-\partial_p^2\psi_1(W,\theta_0,p_0,\eta_{10})\parallel_{P,2}\leq&O(\varepsilon_N)+O(1)\times O\left(N^{-1/2}\right)\\=&O(\varepsilon_N),\end{aligned}$$

where I assume that ε_N converges to zero no faster than $N^{-1/2}$.

Repeated cross sections:

In step 1, I show the main result with the following auxiliary results:

$$\sup_{\eta_2\in\mathcal{T}_N}\left(E\left[\parallel\psi_2(W,\theta_0,p_0,\lambda_0,\eta_2)-\psi_2(W,\theta_0,p_0,\lambda_0,\eta_{20})\parallel^2\right]\right)^{1/2}\leq\varepsilon_N,\qquad\text{(A.5)}$$

$$\sup_{r\in(0,1),\eta_2\in\mathcal{T}_N}\parallel\partial_r^2E\left[\psi_2(W,\theta_0,p_0,\lambda_0,\eta_{20}+r(\eta_2-\eta_{20}))\right]\parallel\leq(\varepsilon_N)^2.\qquad\text{(A.6)}$$

$$\sup_{\eta_2\in\mathcal{T}_N}\left(E_P\left[\parallel\partial_p\psi_2(W,\theta_0,p_0,\lambda_0,\eta_2)-\partial_p\psi_2(W,\theta_0,p_0,\lambda_0,\eta_{20})\parallel^2\right]\right)^{1/2}\leq\varepsilon_N,\qquad\text{(A.7)}$$

$$\sup_{\eta_2 \in \mathcal{T}_N} \left(E_P \left[\| \partial_\lambda \psi_2 \left(W, \theta_0, p_0, \lambda_0, \eta_2 \right) - \partial_\lambda \psi_2 \left(W, \theta_0, p_0, \lambda_0, \eta_{20} \right) \|^2 \right] \right)^{1/2} \leq \varepsilon_N, \tag{A.8}$$

$$\sup_{p \in \mathcal{P}_N, \eta_2 \in \mathcal{T}_N} \left(E_P \left[\| \partial_p^2 \psi_2 \left(W, \theta_0, p, \lambda_0, \eta_2 \right) - \partial_p^2 \psi_2 \left(W, \theta_0, p_0, \lambda_0, \eta_{20} \right) \|^2 \right] \right)^{1/2} \leq \varepsilon_N, \tag{A.9}$$

$$\sup_{p \in \mathcal{P}_N, \lambda \in \Lambda_N, \eta_2 \in \mathcal{T}_N} \left(E_P \left[\| \partial_\lambda^2 \psi_2 \left(W, \theta_0, p, \lambda, \eta_2 \right) - \partial_\lambda^2 \psi_2 \left(W, \theta_0, p_0, \lambda_0, \eta_{20} \right) \|^2 \right] \right)^{1/2} \leq \varepsilon_N, \tag{A.10}$$

$$\sup_{p \in \mathcal{P}_N, \eta_2 \in \mathcal{T}_N} \left(E_P \left[\| \partial_\lambda \partial_p \psi_2 \left(W, \theta_0, p, \lambda_0, \eta_2 \right) - \partial_\lambda \partial_p \psi_2 \left(W, \theta_0, p_0, \lambda_0, \eta_{20} \right) \|^2 \right] \right)^{1/2} \leq \varepsilon_N, \tag{A.11}$$

where $\mathcal{T}_N$ is the set of all $\eta_2 = (g, \ell_2)$ consisting of P-square-integrable functions g and ℓ_2 such that

$$\| \eta_2 - \eta_{20} \|_{P,2} \leq \varepsilon_N,$$

$$\| g - 1/2 \|_{P,\infty} \leq 1/2 - \kappa,$$

$$\| (g - g_0) \|_{P,2}^2 + \| (g - g_0) \|_{P,2} \times \| (\ell_2 - \ell_{20}) \|_{P,2} \leq (\varepsilon_N)^2,$$

$\mathcal{P}_N$ and Λ_N are the sets consisting all $p > 0$ and $\lambda > 0$ such that $| p - p_0 | \leq N^{-1/2}$ and $| \lambda - \lambda_0 | \leq N^{-1/2}$, respectively. By the regularity condition (3.2), $| \hat{p}_k - p_0 | = O_P \left(N^{-1/2} \right)$, and $| \hat{\lambda}_k - \lambda_0 | = O_P \left(N^{-1/2} \right)$, we have $\hat{\eta}_{2k} \in \mathcal{T}_N$, $\hat{p}_k \in \mathcal{P}_N$, and $\hat{\lambda}_k \in \Lambda_N$ with probability $1 - o(1)$.

In Step 2-4, I show the above auxiliary results.

Step 1. Notice that

$$
\begin{aligned}
\sqrt{N}\left(\tilde{\theta}-\theta_0\right) =&\sqrt{N}\left(\frac{1}{K}\sum_{k=1}^{K}\tilde{\theta}_k-\theta_0\right)\\
=&\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi_2\left(W,\theta_0,\hat{p}_k,\hat{\lambda}_k,\hat{\eta}_{2k}\right)\right]\\
=&\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi_2\left(W,\theta_0,p_0,\lambda_0,\hat{\eta}_{2k}\right)\right]\\
&+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\partial_p\psi_2\left(W,\theta_0,p_0,\lambda_0,\hat{\eta}_{2k}\right)\right]\left(\hat{p}_k-p_0\right)\\
&+\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\partial_\lambda\psi_2\left(W,\theta_0,p_0,\lambda_0,\hat{\eta}_{2k}\right)\right]\left(\hat{\lambda}_k-\lambda_0\right)+o_P\left(1\right),
\end{aligned}
$$

where the term $o_P\left(1\right)$, by the same arguments for the term b in repeated outcomes and the auxiliary results (A.9)-(A.11), contains the second-order terms

$$\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\partial_p^2\psi_2\left(W,\theta_0,\bar{p}_k,\lambda_0,\hat{\eta}_{2k}\right)\right]\left(\hat{p}_k-p_0\right)^2,$$

$$\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\partial_\lambda^2\psi_2\left(W,\theta_0,\hat{p}_k,\bar{\lambda}_k,\hat{\eta}_{2k}\right)\right]\left(\hat{\lambda}_k-\lambda_0\right)^2,$$

$$\sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\partial_\lambda\partial_p\psi_2\left(W,\theta_0,\bar{p}_k,\lambda_0,\hat{\eta}_{2k}\right)\right]\left(\hat{\lambda}_k-\lambda_0\right)\left(\hat{p}_k-p_0\right),$$

where $\bar{p}_k\in\left(\hat{p}_k,p_0\right)$ and $\bar{\lambda}_k\in\left(\hat{\lambda}_k,\lambda_0\right)$. On the other hand, by the same arguments for the term a in repeated outcomes and the auxiliary results (A.7)-(A.8), we have

$$\mathbb{E}_{n,k}\left[\partial_p\psi_2\left(W,\theta_0,p_0,\lambda_0,\hat{\eta}_{2k}\right)\right]\xrightarrow{p}E_p\left[\partial_p\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}\right)\right]=G_{2p0},$$

$$\mathbb{E}_{n,k}\left[\partial_\lambda\psi_2\left(W,\theta_0,p_0,\lambda_0,\hat{\eta}_{2k}\right)\right]\xrightarrow{p}E_p\left[\partial_\lambda\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}\right)\right]=G_{2\lambda0}.$$

Hence, since $\hat{p}_k - p_0 = \mathbb{E}_{n,k}\left[D - p_0\right]$ and $\hat{\lambda}_k - \lambda_0 = \mathbb{E}_{n,k}\left[T - \lambda_0\right]$, we have

$$\begin{aligned}\sqrt{N}\left(\tilde{\theta} - \theta_0\right) =& \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi_2\left(W, \theta_0, p_0, \lambda_0, \hat{\eta}_{2k}\right)\right] \\ =& \sqrt{N}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\Big[\psi_2\left(W, \theta_0, p_0, \lambda_0, \hat{\eta}_{1k}\right) + G_{2p0}\left(D - p_0\right) + G_{2\lambda 0}\left(T - \lambda_0\right)\Big] \\ & + o_P\left(1\right) \\ =& \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big[\psi_2\left(W_i, \theta_0, p_0, \lambda_0, \eta_{20}\right) + G_{2p0}\left(D_i - p_0\right) + G_{2\lambda 0}\left(T_i - \lambda_0\right)\Big] \\ & + \sqrt{N}R'_N + o_P\left(1\right),\end{aligned}$$

where

$$\begin{aligned}R'_N =& \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\Big[\psi_2\left(W, \theta_0, p_0, \lambda_0, \hat{\eta}_{2k}\right) + G_{2p0}\left(D - p_0\right) + G_{2\lambda 0}\left(T - \lambda_0\right)\Big] \\ & - \frac{1}{N}\sum_{i=1}^{N}\Big[\psi_2\left(W_i, \theta_0, p_0, \lambda_0, \eta_{20}\right) + G_{2p0}\left(D_i - p_0\right) + G_{2\lambda 0}\left(T_i - \lambda_0\right)\Big] \\ =& \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi_2\left(W, \theta_0, p_0, \lambda_0, \hat{\eta}_{2k}\right)\right] - \frac{1}{N}\sum_{i=1}^{N}\psi_2\left(W_i, \theta_0, p_0, \lambda_0, \eta_{10}\right).\end{aligned}$$

Using (A.5)-(A.6) and the same arguments as the step 1 in repeated outcomes, one can show that $\sqrt{N}R'_N = o_P\left(1\right)$. Hence, it remains to prove the auxiliary results (A.5)-(A.11).

Step 2. Recall that $p'_0 = p_0\lambda_0\left(1 - \lambda_0\right)$. For (A.5), notice that

$$\begin{aligned}\psi_2\left(W, \theta_0, p_0, \lambda_0, \eta_2\right) - \psi_2\left(W, \theta_0, p_0, \lambda_0, \eta_{20}\right) =& \frac{D - g\left(X\right)}{p'_0\left(1 - g\left(X\right)\right)}\left(\left(T - \lambda_0\right)Y - \ell_2\left(X\right)\right) \\ & - \frac{D - g_0\left(X\right)}{p'_0\left(1 - g_0\left(X\right)\right)}\left(\left(T - \lambda_0\right)Y - \ell_{20}\left(X\right)\right) \\ =& \frac{U + g_0\left(X\right) - g\left(X\right)}{p'_0\left(1 - g\left(X\right)\right)}\left(V_2 + \ell_{20}\left(X\right) - \ell_2(X)\right) \\ & - \frac{UV_2}{p'_0\left(1 - g_0\left(X\right)\right)}.\end{aligned}$$

The decomposition becomes

$$\begin{aligned}\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_2\right)-\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}\right)=&\frac{UV_2}{p_0'\left(1-g\left(X\right)\right)}+\frac{U\left(\ell_{20}\left(X\right)-\ell_2\left(X\right)\right)}{p_0'\left(1-g\left(X\right)\right)}\\&+\frac{\left(g_0\left(X\right)-g\left(X\right)\right)V_2}{p_0'\left(1-g\left(X\right)\right)}\\&+\frac{\left(g_0\left(X\right)-g\left(X\right)\right)\left(\ell_{20}\left(X\right)-\ell_2\left(X\right)\right)}{p_0'\left(1-g\left(X\right)\right)}\\&-\frac{UV_2}{p_0'\left(1-g_0\left(X\right)\right)}.\end{aligned}$$

Given that $\kappa\leq g_0\left(X\right)\leq1-\kappa$, $\kappa\leq g\left(X\right)\leq1-\kappa$, we have

$$\begin{aligned}\parallel\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_2\right)-\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}\right)\parallel_{P,2}\leq&\frac{1}{p_0'\kappa^2}\parallel UV_2\left(1-g_0\left(X\right)\right)\\&+U\left(\ell_{20}\left(X\right)-\ell_2\left(X\right)\right)\left(1-g_0\left(X\right)\right)\\&+V_2\left(g_0\left(X\right)-g\left(X\right)\right)\left(1-g_0\left(X\right)\right)\\&+\left(g_0-g\right)\left(\ell_{20}-\ell_2\right)\left(1-g_0\left(X\right)\right)\\&-UV_2\left(1-g\left(X\right)\right)\parallel_{P,2}.\end{aligned}$$

By $\kappa\leq g_0\left(X\right)\leq1-\kappa$, $\kappa\leq g\left(X\right)\leq1-\kappa$ again, we obtain

$$\begin{aligned}\parallel\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_2\right)-\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}\right)\parallel_{P,2}\leq&\frac{1-\kappa}{p_0'\kappa^2}\parallel UV_2\\&+U\left(\ell_{20}\left(X\right)-\ell_2\left(X\right)\right)\\&+V_2\left(g_0\left(X\right)-g\left(X\right)\right)\\&+\left(g_0-g\right)\left(\ell_{20}-\ell_2\right)\\&-UV_2\parallel_{P,2}.\end{aligned}$$

Given $E_P\left[U^2 \mid X\right] \leq C$, $E_P\left[V_2^2 \mid X\right] \leq C$, and the conditions on the rates of convergence,

$$\begin{aligned}\| \psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_2\right)-\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}\right)\|_{P,2} \leq & \frac{(1-\kappa)\sqrt{C}}{p_0'\kappa^2} \| \ell_{20}\left(X\right)-\ell_2\left(X\right)\|_{P,2} \\ & +\frac{(1-\kappa)\sqrt{C}}{p_0'\kappa^2} \| g_0\left(X\right)-g\left(X\right)\|_{P,2} \\ & +\frac{(1-\kappa)}{p_0'\kappa^2} \| g_0-g\|_{P,2} \| \ell_{20}-\ell_2\|_{P,2} \\ & \leq O\left(\varepsilon_N+\varepsilon_N+\left(\varepsilon_N\right)^2\right) \\ & =O\left(\varepsilon_N\right).\end{aligned}$$

For (A.6), let $f\left(r\right)=E_P\left[\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}+r\left(\eta_2-\eta_{20}\right)\right)\right]$. Then the second-order derivative is

$$\begin{aligned}\partial_r^2 f\left(r\right)= & \frac{2}{p_0'}E_P\left[\frac{\left(D-1\right)\left(g-g_0\right)^2}{\left(1-g_0-r\left(g-g_0\right)\right)^3}\left(\left(T-\lambda_0\right)Y-\ell_{20}-r\left(\ell_2-\ell_{20}\right)\right)\right] \\ & -\frac{2}{p_0'}E_P\left[\frac{D-1}{\left(1-g_0-r\left(g-g_0\right)\right)^2}\left(\ell_2-\ell_{20}\right)\left(g-g_0\right)\right]\end{aligned}$$

It follows that

$$\mid \partial_r^2 f\left(r\right) \mid \leq O\left(\| \left(g-g_0\right)\|_{P,2}^2+\| \left(g-g_0\right)\|_{P,2} \times \| \left(\ell_2-\ell_{20}\right)\|_{P,2}\right) \leq \left(\varepsilon_N\right)^2.$$

Step 3. For (A.7), notice that

$$\begin{aligned}\partial_p\psi_2\left(W,\theta,p,\lambda,\eta_2\right)= & -\frac{1}{p^2\lambda\left(1-\lambda\right)}\frac{D-g\left(X\right)}{1-g\left(X\right)}\left(\left(T-\lambda\right)Y-\ell_2\left(X\right)\right) \\ = & -\frac{1}{p}\left(\psi_2\left(W,\theta,p,\lambda,\eta_2\right)+\theta\right),\end{aligned}$$

then we have

$$\begin{aligned}\parallel \partial_p\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_2\right)-\partial_p\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}\right)\parallel_{P,2}=&\frac{1}{p_0}\parallel \psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_2\right)\\&-\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}\right)\parallel_{P,2}\\=&O\left(\varepsilon_N\right)\end{aligned}$$

by the proof of (A.5).

For (A.8), notice that

$$\begin{aligned}\partial_\lambda\psi_2\left(W,\theta,p,\lambda,\eta_2\right)=&-\frac{1-2\lambda}{\lambda^2\left(1-\lambda\right)^2}\frac{D-g\left(X\right)}{p\left(1-g\left(X\right)\right)}\left(\left(T-\lambda\right)Y-\ell_2\left(X\right)\right)\\&-\frac{Y}{p\lambda\left(1-\lambda\right)}\frac{D-g\left(X\right)}{1-g\left(X\right)}.\end{aligned}$$

Define $\partial_\lambda\psi_{20}\equiv\partial_\lambda\psi_2\left(W,\theta_0,p_0,\lambda_0,\eta_{20}\right)$, then

$$\begin{aligned}\parallel \partial_\lambda\psi_2(W,\theta_0,p_0,\lambda_0,\eta_2)-\partial_\lambda\psi_{20}\parallel_{P,2}=&\parallel \psi_2(W,\theta_0,p_0,\lambda_0,\eta_2)-\psi_2(W,\theta_0,p_0,\lambda_0,\eta_{20})\parallel_{P,2}\\&\times\frac{\mid 1-2\lambda_0\mid}{\lambda_0\left(1-\lambda_0\right)}\\&+\parallel \frac{Y}{p_0'}\left(\frac{D-g\left(X\right)}{1-g\left(X\right)}-\frac{D-g_0\left(X\right)}{1-g_0\left(X\right)}\right)\parallel_{P,2}\\=&O\left(\varepsilon_N\right)+\parallel \frac{Y}{p_0'}\left(\frac{D-g\left(X\right)}{1-g\left(X\right)}-\frac{D-g_0\left(X\right)}{1-g_0\left(X\right)}\right)\parallel_{P,2}\\\leq&O\left(\varepsilon_N\right)+\frac{1}{p_0'\kappa^2}\parallel Y\left(g-g_0\right)\left(D-1\right)\parallel_{P,2}\\\leq&O\left(\varepsilon_N\right)+\frac{\sqrt{C}}{p_0'\kappa^2}\parallel g-g_0\parallel_{P,2}\\=&O\left(\varepsilon_N\right),\end{aligned}$$

by (A.5) and $E_P\left[Y^2\mid X\right]\leq C$.

Step 4. For (A.9), notice that we have

$$\partial_p^2 \psi_2 (W, \theta, p, \lambda, \eta_2) = \frac{2}{p^3 \lambda (1-\lambda)} \frac{D - g(X)}{1 - g(X)} ((T - \lambda) Y - \ell_2 (X)).$$

Define $\partial_p^2 \psi_{20} \equiv \partial_p^2 \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_{20})$, then we have

$$\begin{aligned} \partial_p^2 \psi_2 (W, \theta_0, p, \lambda_0, \eta_2) - \partial_p^2 \psi_{20} =& \partial_p^2 \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_2) - \partial_p^2 \psi_{20} \\ &+ \partial_p^3 \psi_2 (W, \theta_0, \bar{p}, \lambda_0, \eta_2) (p - p_0) \\ =& \frac{2}{p^2} (\psi_2 (W, \theta_0, p_0, \lambda_0, \eta_2) - \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_{20})) \\ &+ \partial_p^3 \psi_2 (W, \theta_0, \bar{p}, \lambda_0, \eta_2) (p - p_0), \end{aligned}$$

where $\bar{p} \in (p, p_0)$. Hence, we have

$$\begin{aligned} \| \partial_p^2 \psi_2 (W, \theta_0, p, \lambda_0, \eta_2) - \partial_p^2 \psi_{20} \|_{P,2} \leq & \frac{2}{p^2} \| \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_2) - \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_{20}) \| \\ &+ \| \frac{D - g(X)}{1 - g(X)} ((T - \lambda_0) Y - \ell_2 (X)) \|_{P,2} \\ &\times \frac{6}{\bar{p}^4 \lambda_0 (1 - \lambda_0)} \mid p - p_0 \mid. \end{aligned}$$

By (A.5), we have $\| \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_2) - \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_{20}) \|_{P,2} = O(\varepsilon_N)$. The term in the second line is bounded by

$$\begin{aligned} \frac{1}{\kappa} \| (U + g_0 - g)(V_2 + \ell_{20} - \ell_2) \|_{P,2} \leq & \frac{1}{\kappa} \| U V_2 \|_{P,2} + \frac{1}{\kappa} \| U (\ell_{20} - \ell_2) \|_{P,2} \\ &+ \frac{1}{\kappa} \| V_2 (g_0 - g) \|_{P,2} \\ &+ \frac{1}{\kappa} \| g_0 - g \|_{P,2} \| \ell_{20} - \ell_2 \|_{P,2} \\ \leq & \frac{1}{\kappa} \left(C + \sqrt{C} \| \ell_{20} - \ell_2 \|_{P,2} + \sqrt{C} \| g_0 - g \|_{P,2} \right) \\ &+ \frac{1}{\kappa} \| g_0 - g \|_{P,2} \| \ell_{20} - \ell_2 \|_{P,2} \\ =& O(1) \end{aligned}$$

by $\| UV_2 \|_{P,2} \leq \| UV_2 \|_{P,4} \leq C$, $E_P \left[U^2 \mid X \right] \leq C$, and $E_P \left[V_2^2 \mid X \right] \leq C$. Thus, we obtain

$$\begin{aligned} \| \partial_p^2 \psi_2 \left(W, \theta_0, p, \lambda_0, \eta_2 \right) - \partial_p^2 \psi_{20} \|_{P,2} \leq & O \left(\varepsilon_N \right) + O \left(1 \right) \times O \left(N^{-1/2} \right) \\ = & O \left(\varepsilon_N \right), \end{aligned}$$

where I assume that ε_N converges to zero no faster than $N^{-1/2}$.

For (A.10), notice that we have

$$\begin{aligned} \partial_\lambda^2 \psi_2 \left(W, \theta, p, \lambda, \eta_2 \right) = & \frac{c_1}{p \lambda^3 \left(1 - \lambda \right)^3} \frac{D - g \left(X \right)}{1 - g \left(X \right)} \left(\left(T - \lambda \right) Y - \ell_2 \left(X \right) \right) \\ & + \frac{2 - 4\lambda}{p \lambda^2 \left(1 - \lambda \right)^2} \frac{D - g \left(X \right)}{1 - g \left(X \right)} Y, \end{aligned}$$

where c_1 is a constant depending on λ. Define $\partial_\lambda^2 \psi_{20} \equiv \partial_\lambda^2 \psi_2 \left(W, \theta_0, p_0, \lambda_0, \eta_{20} \right)$, we have

$$\begin{aligned} \partial_\lambda^2 \psi_2 \left(W, \theta_0, p, \lambda, \eta_2 \right) - \partial_\lambda^2 \psi_{20} = & \partial_\lambda^2 \psi_2 \left(W, \theta_0, p_0, \lambda_0, \eta_2 \right) - \partial_\lambda^2 \psi_{20} \\ & + \partial_\lambda^2 \partial_p \psi_2 \left(W, \theta_0, \bar{p}, \lambda, \eta_2 \right) \left(p - p_0 \right) \\ & + \partial_\lambda^3 \psi_2 \left(W, \theta_0, p_0, \bar{\lambda}, \eta_2 \right) \left(\lambda - \lambda_0 \right) \\ = & \frac{c_1}{\lambda_0^2 \left(1 - \lambda_0 \right)^2} \left(\psi_2 \left(W, \theta_0, p_0, \lambda_0, \eta_2 \right) - \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_{20}) \right) \\ & + \frac{2 - 4\lambda_0}{p_0 \lambda_0^2 \left(1 - \lambda_0 \right)^2} \left(\frac{D - g \left(X \right)}{1 - g \left(X \right)} - \frac{D - g_0 \left(X \right)}{1 - g_0 \left(X \right)} \right) Y \\ & + \partial_\lambda^2 \partial_p \psi_2 \left(W, \theta_0, \bar{p}, \lambda, \eta_2 \right) \left(p - p_0 \right) \\ & + \partial_\lambda^3 \psi_2 \left(W, \theta_0, p_0, \bar{\lambda}, \eta_2 \right) \left(\lambda - \lambda_0 \right), \end{aligned}$$

where $\bar{p} \in (p, p_0)$ and $\bar{\lambda} \in (\lambda, \lambda_0)$. By the triangle inequality, we have

$$\begin{aligned}\parallel \partial_\lambda^2 \psi_2 (W, \theta_0, p, \lambda, \eta_2) - \partial_\lambda^2 \psi_{20} \parallel_{P,2} \leq & \frac{\mid c_1 \mid}{\lambda^2 (1-\lambda)^2} \times \\ & \parallel \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_2) - \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_{20}) \parallel_{P,2} \\ & + \frac{\mid 2 - 4\lambda_0 \mid Y}{p_0 \lambda_0^2 (1-\lambda_0)^2} \parallel \left(\frac{D - g(X)}{1 - g(X)} - \frac{D - g_0(X)}{1 - g_0(X)} \right) \parallel_{P,2} \\ & + \parallel \partial_\lambda^2 \partial_p \psi_2 (W, \theta_0, \bar{p}, \lambda, \eta_2) \parallel_{P,2} \mid p - p_0 \mid \\ & + \parallel \partial_\lambda^3 \psi_2 \left(W, \theta_0, p_0, \bar{\lambda}, \eta_2 \right) \parallel_{P,2} \mid \lambda - \lambda_0 \mid .\end{aligned}$$

The norm term is the second line is bounded by

$$\begin{aligned}\frac{1}{\kappa^2} \parallel Y (D-1)(g - g_0) \parallel_{P,2} \leq & \frac{\sqrt{C}}{\kappa^2} \parallel g - g_0 \parallel_{P,2} \\ = & O(\varepsilon_N),\end{aligned}$$

by $E_P \left[Y^2 \mid X \right] \leq C$ and $D \in \{0, 1\}$. The two high-order terms are bounded by

$$\begin{aligned}\parallel \partial_\lambda^2 \partial_p \psi_2 (W, \theta_0, \bar{p}, \lambda, \eta_2) \parallel_{P,2} \leq & \frac{\mid c_1 \mid}{\bar{p}^2 \lambda^3 (1-\lambda)^3} \parallel \frac{D - g(X)}{1 - g(X)} ((T - \lambda) Y - \ell_2(X)) \parallel_{P,2} \\ & + \frac{\mid 2 - 4\lambda \mid}{\bar{p} \lambda^2 (1-\lambda)^2} \parallel \frac{D - g(X)}{1 - g(X)} Y \parallel_{P,2} .\end{aligned}$$

and

$$\begin{aligned}\parallel \partial_\lambda^3 \psi_2 \left(W, \theta_0, p_0, \bar{\lambda}, \eta_2 \right) \parallel_{P,2} \leq & \frac{\mid c_2 \mid}{p_0 \bar{\lambda}^4 \left(1 - \bar{\lambda}\right)^4} \parallel \frac{D - g(X)}{1 - g(X)} \left(\left(T - \bar{\lambda} \right) Y - \ell_2(X) \right) \parallel_{P,2} \\ & + \frac{\mid c_3 \mid}{p_0 \bar{\lambda}^3 \left(1 - \bar{\lambda}\right)^3} \parallel \frac{D - g(X)}{1 - g(X)} \times Y \parallel_{P,2},\end{aligned}$$

where c_2 and c_3 are constants depending on λ. Using the same arguments in (A.9), one can show that

$$\parallel \frac{D - g(X)}{1 - g(X)} ((T - \lambda) Y - \ell_2(X)) \parallel_{P,2} \leq O(1),$$

$$\parallel \frac{D-g\left(X\right)}{1-g\left(X\right)}\left(\left(T-\bar{\lambda}\right)Y-\ell_{2}\left(X\right)\right)\parallel_{P,2}\leq O\left(1\right).$$

Also, we have

$$\begin{aligned}\parallel \frac{D-g\left(X\right)}{1-g\left(X\right)}\times Y\parallel_{P,2}= & \parallel \frac{U+g_{0}\left(X\right)-g\left(X\right)}{1-g\left(X\right)}\times Y\parallel_{P,2}\\ \leq & \frac{1}{\kappa}\left(\parallel UY\parallel_{P,2}+\parallel\left(g_{0}-g\right)Y\parallel_{P,2}\right)\\ \leq & \frac{1}{\kappa}\left(C+\sqrt{C}\parallel g_{0}-g\parallel_{P,2}\right)\\ = & O\left(1\right)\end{aligned}$$

by $\parallel UY\parallel_{P,2}\leq C$ and $E_{P}\left[Y^{2}\mid X\right]\leq C$.

Finally, we obtain

$$\begin{aligned}\parallel \partial_{\lambda}^{2}\psi_{2}\left(W,\theta_{0},p,\lambda,\eta_{2}\right)-\partial_{\lambda}^{2}\psi_{20}\parallel_{P,2}\leq & O\left(\varepsilon_{N}\right)+O\left(\varepsilon_{N}\right)+O\left(1\right)O\left(N^{-1/2}\right)\\ & +O\left(1\right)O\left(N^{-1/2}\right)\\ = & O\left(\varepsilon_{N}\right),\end{aligned}$$

where I assume that ε_{N} converges to zero no faster than $N^{-1/2}$.

For (A.11), notice that the derivative is

$$\begin{aligned}\partial_{\lambda}\partial_{p}\psi_{2}\left(W,\theta,p,\lambda,\eta_{2}\right)= & \frac{1-2\lambda}{p^{2}\lambda^{2}\left(1-\lambda\right)^{2}}\frac{D-g\left(X\right)}{1-g\left(X\right)}\left(\left(T-\lambda\right)Y-\ell_{2}\left(X\right)\right)\\ & +\frac{Y}{p^{2}\lambda\left(1-\lambda\right)}\frac{D-g\left(X\right)}{1-g\left(X\right)}.\end{aligned}$$

Define $\partial_{\lambda}\partial_{p}\psi_{20}\equiv\partial_{\lambda}\partial_{p}\psi_{2}\left(W,\theta_{0},p_{0},\lambda_{0},\eta_{20}\right)$, then we have

$$\begin{aligned}\partial_{\lambda}\partial_{p}\psi_{2}\left(W,\theta_{0},p,\lambda_{0},\eta_{2}\right)-\partial_{\lambda}\partial_{p}\psi_{20}= & \partial_{\lambda}\partial_{p}\psi_{2}\left(W,\theta_{0},p_{0},\lambda_{0},\eta_{2}\right)-\partial_{\lambda}\partial_{p}\psi_{20}\\ & +\partial_{\lambda}\partial_{p}^{2}\psi_{2}\left(W,\theta_{0},\bar{p},\lambda_{0},\eta_{2}\right)\left(p-p_{0}\right),\end{aligned}$$

where $\bar{p} \in (p, p_0)$. By the triangle inequality, we obtain

$$\begin{aligned}\| \partial_\lambda \partial_p \psi_2 (W, \theta_0, p, \lambda_0, \eta_2) - \partial_\lambda \partial_p \psi_{20} \|_{P,2} \leq & \frac{1}{p} \| \partial_\lambda \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_2) - \partial_\lambda \psi_{20} \|_{P,2} \\ & + \| \partial_\lambda \partial_p^2 \psi_2 (W, \theta_0, \bar{p}, \lambda_0, \eta_2) \|_{P,2} | p - p_0 | \, .\end{aligned}$$

Using the same arguments in (A.9) and (A.10), one can show that the high-order term is bounded by

$$\begin{aligned}\| \partial_\lambda \partial_p^2 \psi_2 (W, \theta_0, \bar{p}, \lambda_0, \eta_2) \|_{P,2} \leq & \| \frac{2 - 4\lambda_0}{\bar{p}^3 \lambda_0^2 (1 - \lambda_0)^2} \frac{D - g(X)}{1 - g(X)} ((T - \lambda_0) Y - \ell_2 (X)) \|_{P,2} \\ & + \| \frac{2Y}{\bar{p}^3 \lambda_0 (1 - \lambda_0)} \frac{D - g(X)}{1 - g(X)} \|_{P,2} \\ \leq & O(1) \, .\end{aligned}$$

Together with (A.8), we obtain

$$\begin{aligned}\| \partial_\lambda \partial_p \psi_2 (W, \theta_0, p, \lambda_0, \eta_2) - \partial_\lambda \partial_p \psi_{20} \|_{P,2} \leq & O(\varepsilon_N) + O(1) O\left(N^{-1/2}\right) \\ = & O(\varepsilon_N) \, ,\end{aligned}$$

where I assume that ε_N converges to zero no faster than $N^{-1/2}$.

□

Proof of Theorem 1.2:

In Step 1, I show the main result using the auxiliary results

$$\sup_{p \in \mathcal{P}_N, \eta_1 \in \mathcal{T}_N} \left(E_P \left[\| \bar{\psi}_1 (W, \theta_0, p, \eta_1) - \bar{\psi}_1 (W, \theta_0, p_0, \eta_{10}) \|^2 \right] \right)^{1/2} \leq \varepsilon_N, \tag{A.12}$$

$$\left(E_P \left[\bar{\psi}_1 (W, \theta_0, p_0, \eta_{10})^4 \right] \right)^{1/4} \leq C_1, \tag{A.13}$$

where $\mathcal{P}_N$ and $\mathcal{T}_N$ are specified in the proof of Theorem 3.1, C_1 is a constant, and

$$\bar{\psi}_1 \left(W, \theta, p, \eta_1\right) \equiv \frac{1}{p} \frac{D - g\left(X\right)}{1 - g\left(X\right)} \left(Y\left(1\right) - Y\left(0\right) - \ell_1\left(X\right)\right) - \frac{D\theta}{p}.$$

In fact, we have $E_P\left[\left(\bar{\psi}_1\left(W, \theta_0, p_0, \eta_{10}\right)\right)^2\right] = \Sigma_{10}$. In Step 2, I show the auxiliary results (A.12) and (A.13).

Step 1. Notice that

$$\begin{aligned}
\hat{\Sigma}_1 =& \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{n,k}\left[\left(\psi_1\left(W, \tilde{\theta}, \hat{p}_k, \hat{\eta}_{1k}\right) + \hat{G}_{1p}\left(D - \hat{p}_k\right)\right)^2\right] \\
=& \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{n,k}\left[\left(\frac{1}{\hat{p}_k} \frac{D - \hat{g}_k\left(X\right)}{1 - \hat{g}_k\left(X\right)} \left(Y\left(1\right) - Y\left(0\right) - \hat{\ell}_{1k}\left(X\right)\right) - \frac{D\tilde{\theta}}{\hat{p}_k}\right)^2\right] \\
=& \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{n,k}\left[\bar{\psi}_1\left(W, \tilde{\theta}, \hat{p}_k, \hat{\eta}_{1k}\right)^2\right],
\end{aligned}$$

where the second equality follows from $\hat{G}_{1p} = -\tilde{\theta}/\hat{p}_k$.

Since K is fixed, which is independent of N, it suffices to show that for each $k \in [k]$,

$$I_k \equiv \left|\mathbb{E}_{n,k}\left[\bar{\psi}_1\left(W, \tilde{\theta}, \hat{p}_k, \hat{\eta}_{1k}\right)^2\right] - E_P\left[\bar{\psi}_1\left(W, \theta_0, p_0, \eta_{10}\right)^2\right]\right| = o_P\left(1\right).$$

By the triangle inequality, we have

$$I_k \leq I_{3,k} + I_{4,k},$$

where

$$\begin{aligned}
I_{3,k} \equiv& \left|\mathbb{E}_{n,k}\left[\bar{\psi}_1\left(W, \tilde{\theta}, \hat{p}_k, \hat{\eta}_{1k}\right)^2\right] - \mathbb{E}_{n,k}\left[\bar{\psi}_1\left(W, \theta_0, p_0, \eta_{10}\right)^2\right]\right|, \\
I_{4,k} \equiv& \left|\mathbb{E}_{n,k}\left[\bar{\psi}_1\left(W, \theta_0, p_0, \eta_{10}\right)^2\right] - E_P\left[\bar{\psi}_1\left(W, \theta_0, p_0, \eta_{10}\right)^2\right]\right|.
\end{aligned}$$

To bound $I_{4,k}$, we have

$$\begin{aligned}
E_P\left[I_{4,k}^2\right] \leq & n^{-1} E_P\left[\bar{\psi}_1\left(W, \theta_0, p_0, \eta_{10}\right)^4\right] \\
\leq & n^{-1} C_1^4,
\end{aligned}$$

where the last inequality follows from (A.13). Then we have $I_{4,k} = O_P\left(n^{1/2}\right)$.

Next, we bound $I_{3,k}$. This part is essentially identical to the proof of Theorem 3.2 in Chernozhukov et al. (2018), I reproduce it here for reader's convenience. Observe that for any number a and δa,

$$\mid \left(a + \delta a\right)^2 - a^2 \mid \leq 2\left(\delta a\right)\left(a + \delta a\right).$$

Denote $\psi_i = \bar{\psi}_1\left(W_i, \theta_0, p_0, \eta_{10}\right)$ and $\hat{\psi}_i = \bar{\psi}_1\left(W_i, \tilde{\theta}, \hat{p}_k, \hat{\eta}_{1k}\right)$, and $a \equiv \psi_i$, $a + \delta a \equiv \hat{\psi}_i$. Then

$$\begin{aligned}
I_{3,k} = \mid \frac{1}{n}\sum_{i \in I_k}\left(\hat{\psi}_i\right)^2 - \left(\psi_i\right)^2 \mid \leq & \frac{1}{n}\sum_{i \in I_k} \mid \left(\hat{\psi}_i\right)^2 - \left(\psi_i\right)^2 \mid \\
\leq & \frac{2}{n}\sum_{i \in I_k} \mid \hat{\psi}_i - \psi_i \mid \times \left(\mid \psi_i \mid + \mid \hat{\psi}_i - \psi_i \mid\right) \\
\leq & \left(\frac{2}{n}\sum_{i \in I_k} \mid \hat{\psi}_i - \psi_i \mid^2\right)^{1/2} \left(\frac{2}{n}\sum_{i \in I_k}\left(\mid \psi_i \mid + \mid \hat{\psi}_i - \psi_i \mid\right)^2\right)^{1/2} \\
\leq & \left(\frac{2}{n}\sum_{i \in I_k} \mid \hat{\psi}_i - \psi_i \mid^2\right)^{1/2} \left[\left(\frac{2}{n}\sum_{i \in I_k} \mid \psi_i \mid^2\right)^{1/2} + \left(\frac{2}{n}\sum_{i \in I_k} \mid \hat{\psi}_i - \psi_i \mid^2\right)^{1/2}\right].
\end{aligned}$$

Thus,

$$I_{3,k}^2 \lesssim S_N \times \left(\frac{1}{n}\sum_{i \in I_k} \parallel \bar{\psi}_1\left(W_i, \theta_0, p_0, \eta_{10}\right) \parallel^2 + S_N\right),$$

where

$$S_N \equiv \frac{1}{n}\sum_{i \in I_k} \parallel \bar{\psi}_1\left(W_i, \tilde{\theta}, \hat{p}_k, \hat{\eta}_{1k}\right) - \bar{\psi}_1\left(W_i, \theta_0, p_0, \eta_{10}\right) \parallel^2.$$

Since $\frac{1}{n}\sum_{i \in I_k} \parallel \bar{\psi}_1\left(W_i, \theta_0, p_0, \eta_0\right) \parallel^2 = O_P\left(1\right)$, it suffices to bound S_N. We have the decomposition

$$\begin{aligned}
S_N =& \frac{1}{n}\sum_{i\in I_k} \parallel \bar{\psi}_1\left(W_i,\theta_0,\hat{p}_k,\hat{\eta}_{1k}\right) + \partial_\theta\bar{\psi}_1\left(W_i,\bar{\theta},\hat{p}_k,\hat{\eta}_{1k}\right)\left(\tilde{\theta}-\theta_0\right) - \bar{\psi}_1\left(W_i,\theta_0,p_0,\eta_{10}\right) \parallel^2 \\
=& \frac{1}{n}\sum_{i\in I_k} \parallel \bar{\psi}_1\left(W_i,\theta_0,\hat{p}_k,\hat{\eta}_{1k}\right) + \frac{D_i}{\hat{p}_k}\left(\tilde{\theta}-\theta_0\right) - \bar{\psi}_1\left(W_i,\theta_0,p_0,\eta_{10}\right) \parallel^2 \\
\leq& \frac{1}{n}\sum_{i\in I_k} \parallel \frac{D_i}{\hat{p}_k}\left(\tilde{\theta}-\theta_0\right) \parallel^2 + \frac{1}{n}\sum_{i\in I_k} \parallel \bar{\psi}_1\left(W_i,\theta_0,\hat{p}_k,\hat{\eta}_{1k}\right) - \bar{\psi}_1\left(W_i,\theta_0,p_0,\eta_{10}\right) \parallel^2,
\end{aligned}$$

where $\bar{\theta} \in \left(\tilde{\theta}-\theta_0\right)$. The first term is bounded by

$$\begin{aligned}
\frac{1}{n}\sum_{i\in I_k} \parallel \frac{D_i}{\hat{p}_k}\left(\tilde{\theta}-\theta_0\right) \parallel^2 \leq& \left(\frac{1}{n}\sum_{i\in I_k}\left(\frac{D_i}{\hat{p}_k}\right)^2\right) \parallel \tilde{\theta}-\theta_0 \parallel^2 \\
=& \left(\frac{1}{n}\sum_{i\in I_k}\left(\frac{D_i}{p_0}\right)^2 + o_P(1)\right) \parallel \tilde{\theta}-\theta_0 \parallel^2 \\
=& O_P(1) \times O_P\left(N^{-1}\right).
\end{aligned}$$

Also, notice that conditional on $(W_i)_{i\in I_k^c}$, both $\hat{p}_k$ and $\hat{\eta}_{1k}$ can be treated as fixed. Under the event that $\hat{p}_k \in \mathcal{P}_N$ and $\hat{\eta}_{1k} \in \mathcal{T}_N$, we have

$$\begin{aligned}
& E_P\left[\parallel \bar{\psi}_1\left(W_i,\theta_0,\hat{p}_k,\hat{\eta}_{1k}\right) - \bar{\psi}_1\left(W_i,\theta_0,p_0,\eta_{10}\right) \parallel^2 \mid (W_i)_{i\in I_k^c}\right] \\
\leq & \sup_{p\in\mathcal{P}_N,\eta_1\in\mathcal{T}_N} E_P\left[\parallel \bar{\psi}_1\left(W_i,\theta_0,p,\eta_1\right) - \bar{\psi}_1\left(W_i,\theta_0,p_0,\eta_{10}\right) \parallel^2\right] = \left(\varepsilon_N\right)^2
\end{aligned}$$

by (A.12). It follows that $S_N = O_P\left(N^{-1} + \left(\varepsilon_N\right)^2\right)$. Therefore, we obtain

$$I_k = O_P\left(N^{-1/2}\right) + O_P\left(N^{-1/2} + \varepsilon_N\right) = o_P(1).$$

Hence, $\hat{\Sigma}_1 \xrightarrow{p} \Sigma_{10}$.

Step 2. It remains to prove (A.12) and (A.13). By Taylor series expansion,

$$\begin{aligned}\bar{\psi}_1\left(W,\theta_0,p,\eta_1\right)-\bar{\psi}_1\left(W,\theta_0,p_0,\eta_{10}\right)=&\bar{\psi}_1\left(W,\theta_0,p_0,\eta_1\right)-\bar{\psi}_1\left(W,\theta_0,p_0,\eta_{10}\right)\\&+\partial_p\psi_1\left(W,\theta_0,\bar{p},\eta_1\right)\left(p-p_0\right)\\=&\psi_1\left(W,\theta_0,p_0,\eta_1\right)-\psi_1\left(W,\theta_0,p_0,\eta_{10}\right)\\&+\partial_p\psi_1\left(W,\theta_0,\bar{p},\eta_1\right)\left(p-p_0\right),\end{aligned}$$

where $\bar{p}\in(p,p_0)$. Then we have

$$\begin{aligned}\parallel\bar{\psi}_1\left(W,\theta_0,p,\eta_1\right)-\bar{\psi}_1\left(W,\theta_0,p_0,\eta_{10}\right)\parallel_{P,2}\leq&\parallel\psi_1\left(W,\theta_0,p_0,\eta_1\right)-\psi_1\left(W,\theta_0,p_0,\eta_{10}\right)\parallel_{P,2}\\&+\parallel\frac{1}{\bar{p}^2}\frac{D-g\left(X\right)}{1-g\left(X\right)}\left(Y\left(1\right)-Y\left(0\right)-\ell_1\left(X\right)\right)\\&+\frac{D\theta_0}{\bar{p}^2}\parallel_{P,2}\times\mid p-p_0\mid.\end{aligned}$$

By (A.1), we have $\parallel\psi_1\left(W,\theta_0,p_0,\eta_1\right)-\psi_1\left(W,\theta_0,p_0,\eta_{10}\right)\parallel_{P,2}=O\left(\varepsilon_N\right)$. The term in the second line is bounded by

$$\begin{aligned}&\parallel\frac{1}{\bar{p}^2}\frac{U+g_0-g}{1-g}\left(U+\ell_{10}-\ell_1\right)\parallel_{P,2}+\parallel\frac{D\theta_0}{\bar{p}^2}\parallel_{P,2}\\&\leq\frac{1}{\bar{p}^2\kappa}\parallel UV_1\parallel_{P,2}+\frac{1}{\bar{p}^2\kappa}\parallel U\left(\ell_{10}-\ell_1\right)\parallel_{P,2}\\&+\frac{1}{\bar{p}^2\kappa}\parallel V_1\left(g_0-g\right)\parallel_{P,2}+\frac{1}{\bar{p}^2}\mid\theta_0\mid\\&+\frac{1}{\bar{p}^2\kappa}\parallel g_0-g_1\parallel_{P,2}\parallel\ell_{10}-\ell_1\parallel_{P,2}\\\leq&\frac{1}{\bar{p}^2\kappa}\left(C+\sqrt{C}\parallel\ell_{10}-\ell_1\parallel_{P,2}+\sqrt{C}\parallel g_0-g\parallel_{P,2}\right)\\&+\frac{C}{\bar{p}^2p_0\kappa}+\frac{1}{\bar{p}^2\kappa}\parallel g_0-g_1\parallel_{P,2}\parallel\ell_{10}-\ell_1\parallel_{P,2}\\=&O\left(1\right),\end{aligned}$$

where I use $\| UV_1 \|_{P,2} \leq \| UV_1 \|_{P,4} \leq C$, $E_P\left[U^2 \mid X\right] \leq C$, $E_P\left[V_1^2 \mid X\right] \leq C$, and

$$\begin{aligned}
\mid \theta_0 \mid = & \mid E_P\left[\frac{Y(1)-Y(0)}{p_0}\frac{D-g_0(X)}{1-g_0(X)}\right] \mid \\
\leq & \frac{1}{p_0\kappa} \mid E_P\left[(Y(1)-Y(0))\,U\right] \mid \\
= & \frac{1}{p_0\kappa} \mid E_P\left[(\ell_{10}(X)+V_1)\,U\right] \mid \\
= & \frac{1}{p_0\kappa} \mid E_P\left[UV_1\right] \mid \\
\leq & \frac{C}{p_0\kappa}
\end{aligned}$$

by $\mid E_P[UV_1] \mid \leq \| UV_1 \|_{P,4} \leq C$. Thus, we obtain

$$\begin{aligned}
\| \bar{\psi}_1(W,\theta_0,p,\eta_1) - \bar{\psi}_1(W,\theta_0,p_0,\eta_{10}) \|_{P,2} \leq & O(\varepsilon_N) + O(1)\,O\left(N^{-1/2}\right) \\
= & O(\varepsilon_N),
\end{aligned}$$

where I assume that ε_N converges to zero no faster than $N^{-1/2}$.

For (A.13),

$$\begin{aligned}
\| \bar{\psi}_1(W,\theta_0,p_0,\eta_{10}) \|_{P,4} = & \| \frac{1}{p_0}\frac{UV_1}{1-g_0} - \frac{D\theta_0}{p_0} \|_{P,4} \\
\leq & \| \frac{1}{p_0}\frac{UV_1}{1-g_0} \|_{P,4} + \| \frac{D\theta_0}{p_0} \|_{P,4} \\
\leq & \frac{1}{p_0\kappa} \| UV_1 \|_{P,4} + \frac{1}{p_0} \mid \theta_0 \mid \\
\leq & \frac{C}{p_0\kappa} + \frac{C}{p_0^2\kappa}
\end{aligned}$$

since $\| UV_1 \|_{P,4} \leq C$.

Repeated cross sections:

In Step 1, I show the main result with the auxiliary results:

$$\sup_{p\in\mathcal{P}_N,\lambda\in\Lambda_N,\eta_2\in\mathcal{T}_N}(E_P[\| \bar{\psi}_2\left(W,\theta_0,p,\lambda,G_{2\lambda 0},\eta_2\right)-\bar{\psi}_2\left(W,\theta_0,p_0,\lambda_0,G_{2\lambda 0},\eta_{20}\right)\|^2])^2\leq\varepsilon_N, \tag{A.14}$$

$$\left(E_P\left[\bar{\psi}_2\left(W,\theta_0,p_0,\lambda_0,G_{2\lambda 0},\eta_{20}\right)^4\right]\right)^{1/4}\leq C_2, \tag{A.15}$$

where $(P_N,\Lambda_N,\mathcal{T}_N)$ are specified in the proof of Theorem 3.1, C_2 is a constant, and

$$\bar{\psi}_2\left(W,\theta,p,\lambda,G_{2\lambda},\eta_2\right)\equiv\frac{1}{\lambda\left(1-\lambda\right)p}\frac{D-g\left(X\right)}{1-g\left(X\right)}\left(\left(T-\lambda\right)Y-\ell_2\left(X\right)\right)-\frac{D\theta}{p}+G_{2\lambda}\left(T-\lambda\right).$$

In fact, we have $E_P\left[\left(\bar{\psi}_2\left(W,\theta_0,p_0,\lambda_0,G_{2\lambda_0},\eta_{20}\right)\right)^2\right]=\Sigma_{20}$. In Step 2, I prove (A.14) and (A.15).

Step 1. Notice that

$$\begin{aligned}\hat{\Sigma}_2=&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\left(\psi_2\left(W,\tilde{\theta},\hat{p}_k,\hat{\eta}_{1k}\right)+\hat{G}_{2p}\left(D-\hat{p}_k\right)+\hat{G}_{2\lambda}\left(T-\hat{\lambda}_k\right)\right)^2\right]\\=&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\bar{\psi}_2\left(W,\tilde{\theta},\hat{p}_k,\hat{\lambda}_k,\hat{G}_{2\lambda},\hat{\eta}_{2k}\right)^2\right],\end{aligned}$$

where the second inequality follows from $\hat{G}_{2p}=-\tilde{\theta}/\hat{p}_k$.

Since K is fixed, which is independent of N, it suffices to show that

$$J_k\equiv\left|\mathbb{E}_{n,k}\left[\bar{\psi}_2\left(W,\tilde{\theta},\hat{p}_k,\hat{\lambda}_k,\hat{G}_{2\lambda},\hat{\eta}_{2k}\right)^2\right]-E_P\left[\bar{\psi}_2\left(W,\theta_0,p_0,\lambda_0,G_{2\lambda 0},\eta_{20}\right)^2\right]\right|=o_P\left(1\right).$$

By the triangle inequality, we have

$$J_k\leq J_{5,k}+J_{6,k},$$

where

$$J_{5,k} \equiv \left| \mathbb{E}_{n,k} \left[\bar{\psi}_2 \left(W, \tilde{\theta}, \hat{p}_k, \hat{\lambda}_k, \hat{G}_{2\lambda}, \hat{\eta}_{2k} \right)^2 \right] - \mathbb{E}_{n,k} \left[\bar{\psi}_2 \left(W, \theta_0, p_0, \lambda_0, G_{2\lambda 0}, \eta_{20} \right)^2 \right] \right|,$$

$$J_{6,k} \equiv \left| \mathbb{E}_{n,k} \left[\bar{\psi}_2 \left(W, \theta_0, p_0, \lambda_0, G_{2\lambda 0}, \eta_{20} \right)^2 \right] - E_P \left[\bar{\psi}_2 \left(W, \theta_0, p_0, \lambda_0, G_{2\lambda 0}, \eta_{20} \right)^2 \right] \right|.$$

Using the same arguments for $I_{3,k}$ and $I_{4,k}$ in the proof of repeated outcomes and the conditions (A.14) and (A.15), we can show $J_{5,k} = o_P(1)$ and $J_{6,k} = o_P(1)$. Hence, $\hat{\Sigma}_2 \xrightarrow{p} \Sigma_{20}$.

Step 2. It remains to show (A.14) and (A.15). Define $\bar{\psi}_{20} \equiv \bar{\psi}_2 (W, \theta_0, p_0, \lambda_0, G_{2\lambda 0}, \eta_{20})$. By the triangle inequality and

$$\bar{\psi}_2 (W, \theta_0, p_0, \lambda_0, G_{2\lambda 0}, \eta_2) - \bar{\psi}_{20} = \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_2) - \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_{20}),$$

we have

$$\begin{aligned} \parallel \bar{\psi}_2 (W, \theta_0, p, \lambda, G_{2\lambda 0}, \eta_2) - \bar{\psi}_{20} \parallel_{P,2} \leq & \parallel \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_2) - \psi_2 (W, \theta_0, p_0, \lambda_0, \eta_{20}) \parallel_{P,2} \\ & + \parallel \partial_\lambda \bar{\psi}_2 \left(W_i, \theta_0, p_0, \bar{\lambda}, G_{2\lambda 0}, \eta_2 \right) \parallel_{P,2} \mid \lambda - \lambda_0 \mid \\ & + \parallel \partial_p \bar{\psi}_2 (W_i, \theta_0, \bar{p}, \lambda, G_{2\lambda 0}, \eta_2) \parallel_{P,2} \mid p - p_0 \mid, \end{aligned}$$

where $\bar{p} \in (p, p_0)$ and $\bar{\lambda} \in (\lambda, \lambda_0)$. The term in the second line is bounded by

$$\begin{aligned} \parallel \partial_\lambda \bar{\psi}_2 \left(W_i, \theta_0, p_0, \bar{\lambda}, G_{2\lambda 0}, \eta_2 \right) \parallel_{P,2} \leq & \frac{\mid 1 - 2\bar{\lambda} \mid}{p_0 \bar{\lambda}^2 \left(1 - \bar{\lambda}\right)^2} \parallel \frac{D - g(X)}{1 - g(X)} ((T - \bar{\lambda}) Y - \ell_2(X)) \parallel_{P,2} \\ & + \frac{1}{p_0 \bar{\lambda} \left(1 - \bar{\lambda}\right)} \parallel \frac{D - g(X)}{1 - g(X)} \times Y \parallel_{P,2} + \mid G_{2\lambda 0} \mid \\ \leq & O(1) \end{aligned}$$

by the same arguments in (A.9)-(A.11) and

$$
\begin{aligned}
\mid G_{2\lambda 0} \mid = \mid E_P \left[- \frac{1-2\lambda_0}{\lambda_0^2 \left(1-\lambda_0\right)^2 p_0} \frac{D-g_0}{1-g_0} \left(\left(T-\lambda_0\right) Y - \ell_{20} \right) - \frac{Y}{\lambda_0 \left(1-\lambda_0\right) p_0} \frac{D-g_0}{1-g_0} \right] \mid \\
\leq & \frac{\mid 1-2\lambda_0 \mid}{\lambda_0^2 \left(1-\lambda_0\right)^2 p_0 \kappa} \mid E_P \left[UV_2\right] \mid + \frac{1}{\lambda_0 \left(1-\lambda_0\right) p_0 \kappa} \mid E_P \left[YU\right] \mid \\
\leq & \frac{\mid 1-2\lambda_0 \mid}{\lambda_0^2 \left(1-\lambda_0\right)^2 p_0 \kappa} C + \frac{1}{\lambda_0 \left(1-\lambda_0\right) p_0 \kappa} C \\
= & O\left(1\right)
\end{aligned}
$$

since $\mid E_P \left[UV_2\right] \mid \leq \parallel UV_2 \parallel_{P,4} \leq C$ and $\mid E_P \left[YU\right] \mid \leq C$. Also, we have

$$
\begin{aligned}
\parallel \partial_p \bar{\psi}_2 \left(W_i, \theta_0, \bar{p}, \lambda, G_{2\lambda 0}, \eta_2\right) \parallel_{P,2} \leq & \frac{1}{\lambda \left(1-\lambda\right) \bar{p}^2} \parallel \frac{D-g\left(X\right)}{1-g\left(X\right)} \left(\left(T-\lambda\right) Y - \ell_2 \left(X\right) \right) \parallel_{P,2} \\
& + \parallel \frac{D\theta_0}{\bar{p}^2} \parallel_{P,2} \\
\leq & O\left(1\right)
\end{aligned}
$$

by the same arguments in (A.9)-(A.11) and

$$
\begin{aligned}
\mid \theta_0 \mid = & \mid E_P \left[\frac{D-g_0\left(X\right)}{p_0' \left(1-g_0\left(X\right)\right)} \left(T-\lambda_0\right) Y \right] \mid \\
\leq & \frac{1}{p_0' \kappa} \mid E_P \left[\left(T-\lambda_0\right) YU \right] \mid \\
= & \frac{1}{p_0 \kappa} \mid E_P \left[\left(\ell_{20}\left(X\right) + V_2\right) U \right] \mid \\
= & \frac{1}{p_0 \kappa} \mid E_P \left[UV_2\right] \mid \\
\leq & \frac{C}{p_0 \kappa}
\end{aligned}
$$

since $\mid E_P\left[UV_2\right] \mid \leq \parallel UV_2 \parallel_{P,4} \leq C$. Together with (A.5), we have

$$\begin{aligned}\parallel \bar{\psi}_2\left(W,\theta_0,p,\lambda,G_{2\lambda 0},\eta_2\right)-\bar{\psi}_{20} \parallel_{P,2} \leq & O\left(\varepsilon_N\right)+O\left(1\right)O\left(N^{-1/2}\right)+O\left(1\right)O\left(N^{-1/2}\right)\\ = & O\left(\varepsilon_N\right),\end{aligned}$$

where I assume that ε_N converges to zero no faster than $N^{-1/2}$.

For (A.15), we have

$$\begin{aligned}\parallel \bar{\psi}_2\left(W,\theta_0,p_0,\lambda_0,G_{2\lambda 0},\eta_{20}\right) \parallel_{P,4} = & \parallel \frac{1}{\lambda_0\left(1-\lambda_0\right)p_0}\frac{UV_2}{1-g_0}-\frac{D\theta_0}{p_0}+G_{2\lambda 0}\left(T-\lambda_0\right) \parallel_{P,4}\\ \leq & \frac{1}{\lambda_0\left(1-\lambda_0\right)p_0\kappa} \parallel UV_2 \parallel_{P,4}+\frac{1}{p_0}\mid \theta_0 \mid + \mid G_{2\lambda 0} \mid\\ \leq & O\left(1\right)\end{aligned}$$

since $\parallel UV_2 \parallel_{P,4} \leq C$.

□

Lemma A.1 (CONDITIONAL CONVERGENCE IMPLIES UNCONDITIONAL)

Let $\{X_m\}$ and $\{Y_m\}$ be sequences of random vectors. (i) If for $\epsilon_m \to 0$, $Pr(\parallel X_m \parallel > \epsilon_m \mid Y_m) \xrightarrow{p} 0$, then $Pr\left(\parallel X_m \parallel > \epsilon_m\right) \to 0$. This occurs if $E\left[\parallel X_m \parallel^q / \epsilon_m^q \mid Y_m\right] \xrightarrow{p} 0$ for some $q \geq 1$, by Markov's inequality. (ii) Let $\{A_m\}$ be a sequence of positive constants. If $\parallel X_m \parallel = O_P\left(A_m\right)$ conditional on Y_m, namely, that for any $\ell_m \to \infty$, $Pr\left(\parallel X_m \parallel > \ell_m A_m \mid Y_m\right) \xrightarrow{p} 0$, then $\parallel X_m \parallel = O_P\left(A_m\right)$ unconditionally, namely, that for any $\ell_m \to \infty$, $Pr\left(\parallel X_m \parallel > \ell_m A_m\right) \to 0$.

PROOF: This lemma is the Lemma 6.1 in Chernozhukov et al. (2018).

Chapter 2

Mode Treatment Effect

2.1 Introduction

The effects of policies on the distribution of outcomes have long been of central interest in many areas of empirical economics. A policy maker might be interested in the difference of the distribution of outcome under treatment and the distribution of outcome in the absence of treatment. The empirical studies of distributional effects include but not are not limited to Freeman (1980), Card (1996), DiNardo, Fortin, & Lemieux (1995), and Bitler, Gelbach, & Hoynes (2006). Most researches use the difference of the averages or quantiles of the treated and untreated distribution, known as average treatment effect and quantile treatment effect, as a summary for the effect of treatment on distribution. The mode of a distribution, which is also an important summary statistics of data, has long been ignored in the literature. This paper fills up the gap by studying the mode treatment effect: the difference of the modes of the treated and untreated distribution. Compared to the average and the quantile treatment effect, the mode treatment effect has two advantages: (1) mode captures the most probable value of the distribution under treatment and in the absence of treatment. It provides a better summary of centrality than average and quantile when the distributions are highly skewed; (2) mode is robust to heavy-tailed distributions where outliers don't follow the same behavior as the majority of a sample. In economic studies, it is especially often to confront a skewed and heavy-tailed distribution when the outcome of interest is income or wage.

This paper discusses the estimation and inference of the mode treatment effect under the Strong Ignorability assumption (Rosenbaum & Rubin, 1983), which states that conditional on a vector of control variables the treatment is randomly assigned. The first estimator I propose is the kernel estimator. I estimate the density function of the outcome distribution using the kernel method and define the maximum of the estimated density function as the estimator of the mode. While the kernel estimator is a straightforward estimator, it requires to estimate the conditional density function in the process, and the estimation of the conditional density function may be difficult in practice when there exist more than two or three control variables, due to the curse of dimension. The kernel estimator is appropriate if there are less than three control variables. In practice, however, researchers may want to include as many control variables as possible in order to make their identification robust. In this circumstance, the curse of dimension may lead to inaccurate estimation and misleading inference.

To address this problem, I propose the ML estimator. The key feature of the proposed ML estimator is that it translates the estimation of the conditional density function into the estimation of conditional expectation, which we can apply a rich set of ML methods, such as Lasso, random forests, neural nets, and etc, to estimate. This feature provides researchers with the flexibility to apply ML methods to estimate the density function of the outcome distribution. By the virtue of ML methods, the proposed ML estimator can handle the situation when there exist many control variables, even the number of control variables is comparable to or more than the sample size. However, it is well-known that the regularization bias embedded in ML methods may lead to the bias of the final estimator and misleading inference (Chernozhukov et al., 2018). To solve this problem, I further derive the Neyman-orthogonal scores (Chernozhukov et al., 2018) for each estimation which requires the first-step estimation of the conditional expectation. These Neyman-orthogonal scores, to my best knowledge, are new results. The proposed ML estimator is built on the newly derived Neyman-orthogonal score, and hence, it is robust to the regularization bias of the first-step ML estimation.

I derive the asymptotic properties for both the proposed kernel and ML estimators. I show that both estimators are consistent and asymptotically normal with the rate of convergence $\sqrt{Nh^3}$, where N is the sample size and h is the bandwidth of the chosen kernel, which is slower than the traditional rate of convergence $\sqrt{N}$ in the estimation of mean and quantile. In fact, this rate of convergence complies the intuition. While the estimators of mean and quantile are the weighted average of all the available observations, only a small portion of observations near the mode provides the information to the estimator of the mode. This explains the slower rate of convergence for the proposed estimators.

This paper contributes to the program evaluation literature which includes the studies of average treatment effect: Rosenbaum & Rubin (1983), Heckman & Robb (1985), Heckman, Ichimura, & Todd (1997), Hahn (1998), and Hirano, Imbens, & Ridder (2003); the studies of quantile treatment effect: Abadie, Angrist, & Imbens (2002), Chernozhukov & Hansen (2005), and Firpo (2007); the studies of mode estimation and mode regression: Parzen (1962), Eddy et al. (1980), Lee (1989), Yao & Li (2014), and Chen, Genovese, Tibshirani, Wasserman, et al. (2016); as well as the causal inference of ML methods: Belloni et al. (2012), Belloni et al. (2014), Chernozhukov et al. (2015), Belloni et al. (2017), Chernozhukov et al. (2018), and Athey et al. (2019). This paper is also closely related to the robustness of average treatment effect estimation discussed in (Robins & Rotnitzky, 1995) and the general discussion in (Chernozhukov, Escanciano, Ichimura, & Newey, 2016). The asymptotic properties of the robust estimators discussed in these papers remain unaffected if only one of the first-step estimation with classical nonparametric method is inconsistent.

Plan of the paper. Section 2 sets up the notation and framework for the discussion of the mode treatment effect. Section 3 discusses the kernel method and derives the asymptotic properties. Section 4 presents the ML estimator for density estimation and the corresponding Neyman-orthogonal score. I combine the Neyman-orthogonal score with the cross-fitting algorithm to propose the ML estimator of the mode treatment effect, and derive its asymptotic properties. Section 5 concludes this paper.

2.2 Notation and Framework

Let Y be a continuous outcome variable of interest, D the binary treatment indicator, and X $d \times 1$ vector of control variables. Denote by Y_1 an individual's potential outcome when $D = 1$ and Y_0 if $D = 0$. Let $f_{Y_1}(y)$ and $f_{Y_0}(y)$ be the marginal probability density function (p.d.f.) of Y_1 and Y_0, respectively. The modes of Y_1 and Y_0 are the values that appear with the highest probability. That is,

$$\theta_1^* \equiv \arg\max_{y \in \mathcal{Y}_1} f_{Y_1}(y) \text{ and } \theta_0^* \equiv \arg\max_{y \in \mathcal{Y}_0} f_{Y_0}(y),$$

where $\mathcal{Y}_1$ and $\mathcal{Y}_0$ are the supports of Y_1 and Y_0. Here I assume that θ_1^* and θ_0^* are unique, meaning that both Y_1 and Y_0 are unimodal. I also assume that the modes θ_1^* and θ_0^* are in the interior of the common supports of Y_1 and Y_0. These conditions are formally stated in the following assumption:

Assumption 2.1. *(Uni-mode)*

- *For all* $\epsilon > 0$*,*

  $$\sup_{y:|y-\theta_1^*|>\varepsilon} f_{Y_1}(y) < f_{Y_1}(\theta_1^*) \text{ for } y \in \mathcal{Y}_1,$$

  and

  $$\sup_{y:|y-\theta_0^*|>\varepsilon} f_{Y_0}(y) < f_{Y_0}(\theta_0^*) \text{ for } y \in \mathcal{Y}_0.$$

- $\theta_1^*, \theta_0^* \in Int(\mathcal{Y}_1 \cap \mathcal{Y}_0)$*.*

Assumption 2.1 has been widely adopted in many studies (Parzen, 1962; Eddy et al., 1980; Lee, 1989; Yao & Li, 2014). Under Assumption 1, the mode treatment effect is uniquely defined as $\Delta^* \equiv \theta_1^* - \theta_0^*$. The following states the strong ignorability assumption (Rosenbaum & Rubin, 1983):

Assumption 2.2. *(strong ignorability)*

- $(Y_0, Y_1) \perp D \mid X$
- $0 < P(D = 1 \mid X) < 1$

The first part of Assumption 2.2 assumes that potential outcomes are independent of treatment after conditioning on the observable covariates X. The second part states that for all values of X, both treatment status occur with a positive probability. Under the strong ignorability condition, both f_{Y_1} and f_{Y_0} can be identified from the observable variables (Y, D, X) since

$$f_{Y|D=1,X}\left(y \mid x\right) = f_{Y_1|D=1,X}\left(y \mid x\right) = f_{Y_1|X}\left(y \mid x\right),$$

and thus

$$f_{Y_1}(y) = E\left[f_{Y_1|X}\left(y \mid X\right)\right] = E\left[f_{Y|D=1,X}\left(y \mid X\right)\right]. \tag{2.1}$$

Similarly, we have

$$f_{Y_0}\left(y\right) = E\left[f_{Y|D=0,X}\left(y \mid X\right)\right]. \tag{2.2}$$

Equation 2.1 and 2.2 shows the identification result of the density function f_{Y_1} and f_{Y_0}. Then it is straightforward to identify their modes θ_1^* and θ_0^*:

$$\theta_1^* = \arg\max_{y \in \mathcal{Y}_1} E\left[f_{Y|D=1,X}\left(y \mid X\right)\right] \text{ and } \theta_0^* = \arg\max_{y \in \mathcal{Y}_0} E\left[f_{Y|D=0,X}\left(y \mid X\right)\right]. \tag{2.3}$$

If both $f_{Y|D=1,X}\left(y \mid X\right)$ and $f_{Y|D=0,X}\left(y \mid X\right)$ are differentiable with respect to y, we can further identify the modes using the first-order conditions under Assumption 2.1:

$$E\left[f^{(1)}_{Y|D=1,X}\left(\theta_1^* \mid X\right)\right] = 0 \text{ and } E\left[f^{(1)}_{Y|D=0,X}\left(\theta_0^* \mid X\right)\right] = 0, \tag{2.4}$$

where $m^{(s)}\left(y, x\right) \equiv \partial^s m\left(y, x\right) / \partial y^s$ denotes the partial derivatives with respect to y.

Equation 2.1-2.4 provide us a direct way to estimate the modes θ_1^* and θ_0^*. Intuitively, we estimate the density functions $f_{Y_1}(y)$ and $f_{Y_0}(y)$ in the first step and use the maximizers of the estimated density functions as the estimators of the modes. Section 3 and 4 presents the kernel and ML estimation method, respectively.

2.3 The Kernel Estimation

In this section, I propose kernel estimators for θ_1^*, θ_0^*, and the mode treatment effect $\Delta^* = \theta_1^* - \theta_0^*$. Let $K(\cdot)$ be a kernel function with bandwidth h. Define the estimators of the density functions $f_{Y_1}(y)$ and $f_{Y_0}(y)$ as,

$$\hat{f}_{Y_1}(y) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{Y|D=1,X}(y \mid X_i),$$

$$\hat{f}_{Y_0}(y) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{Y|D=0,X}(y \mid X_i)$$

with the kernel estimators

$$\hat{f}_{Y|D=1,X}(y \mid x) = \frac{\sum_{j=1}^{n} D_j K_h\left(y - Y_j\right) K_h\left(x - X_j\right)}{\sum_{j=1}^{n} D_j K_h\left(x - X_j\right)},$$

$$\hat{f}_{Y|D=0,X}(y \mid x) = \frac{\sum_{j=1}^{n} \left(1 - D_j\right) K_h\left(y - Y_j\right) K_h\left(x - X_j\right)}{\sum_{j=1}^{n} \left(1 - D_j\right) K_h\left(x - X_j\right)}$$

where $K_h\left(y - Y_j\right) = h^{-1} K\left(\frac{y - Y_j}{h}\right)$ and

$$K_h\left(x - X_j\right) = h^{-d} K\left(\frac{x_1 - X_{j1}}{h}\right) \times \ldots \times K\left(\frac{x_d - X_{jd}}{h}\right).$$

Then it is straightforward to define the estimators of the modes θ_1^* and θ_0^*:

$$\hat{\theta}_1 \equiv \arg\max_y \hat{f}_{Y_1}(y),$$

$$\hat{\theta}_0 \equiv \arg\max_y \hat{f}_{Y_0}(y).$$

The estimator of the mode treatment effect Δ^* is $\hat{\Delta} \equiv \hat{\theta}_1 - \hat{\theta}_0$. Through out the paper, I impose the following conditions on the kernel $K(\cdot)$:

Assumption 2.3.
- $|K(u)| \leq \bar{K} < \infty$.

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

- *$K(u)$ is differentiable.*

The first part of Assumption 2.3 requires that $K(u)$ is bounded. Although the second part implies that $K(u)$ is a first-order kernel, the arguments in this paper can be easily extended to higher-order kernels. We assume the first-order kernel here just for simplicity. The third part imposes enough smoothness on $K(u)$.

Theorem 2.1. *(Consistency) Suppose Assumption 2.1-2.3 hold. Assume that the density functions $f_{Y|D=1,X}(y \mid x)$ and $f_{Y|D=0,X}(y \mid x)$ are (i) continuous in y, (ii) bounded by some function $d(x)$ with $E[d(X)] < \infty$ for all $y \in \mathcal{Y}$, and (iii) $y \in \mathcal{Y}$ and $x \in \mathcal{X}$ with compact $\mathcal{Y}$ and $\mathcal{X}$. We also assume that the density functions $f_{X|D=1}(x)$ and $f_{X|D=0}(x)$ are bounded away from zero. If $n \to \infty$, $h \to 0$, and $\ln n \left(nh^{d+1}\right)^{-1} \to 0$, then we have $\hat{\theta}_1 \xrightarrow{p} \theta_1^*$ and $\hat{\theta}_0 \xrightarrow{p} \theta_0^*$.*

Theorem 2.2. *Suppose that the assumptions of Theorem 2.1 hold. Assume that $f^{(2)}_{Y|X,D=1}(y \mid x)$ and $f^{(2)}_{Y|X,D=0}(y \mid x)$ are continuous at $y = \theta_1^*$ and $y = \theta_0^*$ for all x, respectively. If $n \to \infty$, $h \to 0$, $\sqrt{nh^3}(\ln n)\left(nh^{d+3}\right)^{-1} \to 0$, $(\ln n)\left(nh^{d+5}\right)^{-1} \to 0$, and $\sqrt{nh^3}h^2 \to 0$, then*

$$\sqrt{nh^3}\left(\hat{\theta}_1 - \theta_1^*\right) \xrightarrow{d} N\left(0, M_1^{-1}V_1M_1^{-1}\right)$$

$$\sqrt{nh^3}\left(\hat{\theta}_0 - \theta_0^*\right) \xrightarrow{d} N\left(0, M_0^{-1}V_0M_0^{-1}\right)$$

where

$$M_1 \equiv E\left[f^{(2)}_{Y|X,D=1}\left(\theta_1^* \mid X\right)\right],$$

$$M_0 \equiv E\left[f^{(2)}_{Y|X,D=0}\left(\theta_0^* \mid X\right)\right],$$

$$V_1 = \kappa_0^{(1)} E\left[\frac{f_{Y|X,D=1}\left(\theta_1^* \mid X\right)}{P\left(D = 1 \mid X\right)}\right],$$

$$V_0 = \kappa_0^{(1)} E\left[\frac{f_{Y|X,D=0}\left(\theta_0^* \mid X\right)}{P\left(D=0 \mid X\right)}\right],$$

and $\kappa_0^{(1)} = \int K^{(1)}\left(u\right)^2 du$. *Further, we have*

$$\sqrt{nh^3}\left(\hat{\Delta} - \Delta^*\right) \xrightarrow{d} N\left(0, M_1 V_1 M_1 + M_0 V_0 M_0\right).$$

Theorem 2.1 and 2.2 show that the asymptotic properties of the estimator of the mode treatment effect. We can see that the proposed estimators follows the asymptotic normality but with the rate of convergence slower than the regular rate $\sqrt{N}$. The intuition is that, unlike the estimation of the average and the quantile treatment effect, the estimation of modes only uses a small portion of total observations which are around the modes. The usage rate of observations determines that the rate of convergence is slower than the regular rate $\sqrt{N}$.

To estimate the asymptotic variances, we define $\pi_0\left(X\right) \equiv P\left(D=1 \mid X\right)$ to be the propensity score. The consistent variance estimators are

$$\hat{M}_1 = \frac{1}{n}\sum_{i=1}^{n} \hat{f}^{(2)}_{Y|X,D=1}\left(\hat{\theta}_1 \mid X_i\right),$$

$$\hat{M}_0 = \frac{1}{n}\sum_{i=1}^{n} \hat{f}^{(2)}_{Y|X,D=0}\left(\hat{\theta}_0 \mid X_i\right),$$

$$\hat{V}_1 = \kappa_0^{(1)} \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{f}_{Y|X,D=1}\left(\hat{\theta}_1 \mid X_i\right)}{\hat{\pi}\left(X_i\right)},$$

$$\hat{V}_0 = \kappa_0^{(1)} \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{f}_{Y|X,D=0}\left(\hat{\theta}_0 \mid X_i\right)}{\hat{\pi}\left(X_i\right)}.$$

Theorem 2.3. *(Variance Estimation) Suppose that the assumptions in Theorem 2.2 hold. Let* $\hat{\pi}\left(x\right)$ *be an uniformly consistent estimator for* $\pi_0\left(x\right)$. *If* $n \to \infty$, $h \to 0$, *and* $\ln n(nh^{d+5}) \to 0$, *then* $\hat{M}_1 \xrightarrow{p} M_1$, $\hat{M}_0 \xrightarrow{p} M_0$, $\hat{V}_1 \xrightarrow{p} V_1$, $\hat{V}_0 \xrightarrow{p} V_0$. *Thus we have* $\hat{M}_1^{-1}\hat{V}_1\hat{M}_1^{-1} \xrightarrow{p} M_1^{-1}V_1M_1^{-1}$ *and* $\hat{M}_0^{-1}\hat{V}_0\hat{M}_0^{-1} \xrightarrow{p} M_0^{-1}V_0M_0^{-1}$.

2.4 The Machine Learning Estimation

In this section, I propose the ML estimator of the mode treatment effect. The ML estimator can accommodate a large number of control variables, potentially more than the sample size. This flexibility will enable researcher to include as many control variables they consider important to make their identification assumptions more plausible. The key to implement ML methods is to replace the estimation of the conditional density function with the estimation of the conditional expectation. To begin with, the estimation of the conditional density function in the traditional kernel estimation is

$$\hat{f}_{Y|D=1,X}\left(y \mid x\right) = \frac{\sum_{j=1}^{n} D_j K_h\left(y - Y_j\right) K_h\left(x - X_j\right)}{\sum_{j=1}^{n} D_j K_h\left(x - X_j\right)}.$$

Notice that we can divide both the numerator and the denominator by $\sum_{j=1}^{n} K_h\left(x - X_j\right)$ to obtain

$$\hat{f}_{Y|D=1,X}\left(y \mid x\right) = \frac{\sum_{j=1}^{n} D_j K_h\left(y - Y_j\right) K_h\left(x - X_j\right) / \sum_{j=1}^{n} K_h\left(x - X_j\right)}{\sum_{j=1}^{n} D_j K_h\left(x - X_j\right) / \sum_{j=1}^{n} K_h\left(x - X_j\right)}.$$

The numerator is an kernel estimator of $E\left[DK_h\left(y - Y\right) \mid X\right]$ and the denominator is an kernel estimator of the propensity score $E\left[D \mid X\right] = \pi\left(X\right)$. Hence, $\hat{f}_{Y|D=1,X}\left(y \mid x\right)$ is an estimator of $E\left[DK_h\left(y - Y\right) \mid X\right] / \pi\left(X\right)$. Then the marginal density estimator

$$\hat{f}_{Y_1}\left(y\right) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{Y|D=1,X}\left(y \mid X_i\right)$$

defined in the previous section can be interpreted as an estimator of

$$E\left[\frac{E\left[DK_h\left(y - Y\right) \mid X\right]}{\pi\left(X\right)}\right] = E\left[\frac{DK_h\left(y - Y\right)}{\pi\left(X\right)}\right].$$

Therefore, we can use the machine learning estimator of $E\left[\frac{DK_h(y-Y)}{\pi(X)}\right]$ as an estimator for $f_{Y_1}\left(y\right)$. We have successfully translate the estimation of the conditional density function into the estimation of the conditional expectation, which is the propensity score $\pi(X)$.

Here we pursue a little bit further to construct the Neyman-orthogonal score (Chernozhukov et al., 2018) for the robustness of the first-step estimation:

$$m_1\left(Z, y, \eta_{10}\right)=\frac{D K_h\left(y-Y\right)}{\pi_0\left(X\right)}-\frac{D-\pi_0\left(X\right)}{\pi_0\left(X\right)} E\left[K_h\left(y-Y\right) \mid X, D=1\right], \tag{2.5}$$

where $Z=\left(Y, D, X\right)$ and $\eta_0=\left(\pi_0, g_{10}\right)$ with $g_{10}\left(X\right) \equiv E\left[K_h\left(y-Y\right) \mid X, D=1\right]$. Similary, the Neyman-orthogonal score for $f_{Y_0}\left(y\right)$ is

$$m_2\left(Z, y, \eta_{20}\right)=\frac{\left(1-D\right) K_h\left(y-Y\right)}{1-\pi_0\left(X\right)}-\frac{\pi_0-D\left(X\right)}{1-\pi_0\left(X\right)} E\left[K_h\left(y-Y\right) \mid X, D=0\right], \tag{2.6}$$

where $\eta_{20}=\left(\pi_0, g_{20}\right)$ with $g_{20}\left(X\right) \equiv E\left[K_h\left(y-Y\right) \mid X, D=0\right]$. Equation 2.5 and 2.6, to my best knowledge, should be the new results for density estimation. The Neyman orthogonality will make the estimation of the density functions more robust to the first-step estimation. Now I combine 2.5 and 2.6 with the cross-fitting algorithm (Chernozhukov et al., 2018) to propose the new estimator:

Definition. *(Algorithm)*

(i) Take a K-fold random partition $\left(I_k\right)_{k=1}^{K}$ of $[N]=\{1,...,N\}$ such that the size of each I_k is $n=N/K$. For each $k\in[K]=\{1,...,K\}$, define the auxiliary sample $I_k^c\equiv\{1,...,N\}$.

(ii) For each $k\in[K]$, use the auxiliary sample I_k^c to construct machine learning estimators

$$\hat{\pi}_k\left(x\right), \; \hat{g}_{1k}\left(x\right), \textit{ and } \hat{g}_{2k}\left(x\right)$$

of $\pi_0\left(x\right)$, $g_{10}\left(x\right)$, and $g_{20}\left(x\right)$.

(iii) Construct the estimator of $f_{Y_1}(y)$ *and* $f_{Y_0}(y)$*:*

$$\hat{f}_{Y_1}(y) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[m_1\left(Z, y, \hat{\eta}_{1k}\right)\right] \text{ and } \hat{f}_{Y_0}(y) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[m_2\left(Z, y, \hat{\eta}_{2k}\right)\right]$$

where $\mathbb{E}_{n,k}\left[m\left(Z\right)\right] = n^{-1}\sum_{i\in I_k} m\left(Z_i\right)$.

(iv) Construct the estimator for θ_1^* *and* θ_0^*

$$\hat{\theta}_1 = \arg\max_y \hat{f}_{Y_1}(y) \text{ and } \hat{\theta}_0 = \arg\max_y \hat{f}_{Y_0}(y).$$

(v) Construct the estimator for the mode treatment effect $\hat{\Delta} = \hat{\theta}_1 - \hat{\theta}_0$.

Theorem 2.4. *Suppose that with probability* $1 - o(1)$, $\| \hat{\eta}_{1k} - \eta_{10} \|_{P,2} \leq \varepsilon_N$, $\| \hat{\pi}_k - 1/2 \|_{P,\infty} \leq 1/2 - \kappa$, *and* $\| \hat{\pi}_k - \pi_0 \|_{P,2}^2 + \| \hat{\pi}_k - \pi_0 \|_{P,2} \times \| \hat{g}_{1k} - g_{10} \|_{P,2} \leq \left(\varepsilon_N\right)^2$. *If* $\epsilon_N = o((Nh^3)^{-1/4})$ *and* $Nh^7 \to 0$, *then we have*

$$\sqrt{nh^3}\left(\hat{\theta}_1 - \theta_1^*\right) \xrightarrow{d} N\left(0, M_1^{-1}V_1M_1^{-1}\right),$$

$$\sqrt{nh^3}\left(\hat{\theta}_0 - \theta_0^*\right) \xrightarrow{d} N\left(0, M_0^{-1}V_0M_0^{-1}\right).$$

As for the variance estimation, recall that the kernel estimator of M_1 in the previous section is $\hat{M}_1 = N^{-1}\sum_{i=1}^{N}\hat{f}^{(2)}_{Y|D=1,X}(\hat{\theta}_1 \mid x)$, where

$$\hat{f}^{(2)}_{Y|D=1,X}(y \mid x) = \frac{\sum_{j=1}^{n} D_j K_h\left(y - Y_j\right)^{(2)} K_h\left(x - X_j\right)}{\sum_{j=1}^{n} D_j K_h\left(x - X_j\right)}.$$

Notice that we can divide both the numerator and the denominator by $\sum_{j=1}^{n} K_h\left(x - X_j\right)$ to obtain

$$\hat{f}^{(2)}_{Y|D=1,X}(y \mid x) = \frac{\sum_{j=1}^{n} D_j K_h\left(y - Y_j\right)^{(2)} K_h\left(x - X_j\right) / \sum_{j=1}^{n} K_h\left(x - X_j\right)}{\sum_{j=1}^{n} D_j K_h\left(x - X_j\right) / \sum_{j=1}^{n} K_h\left(x - X_j\right)}.$$

Observe that the numerator is an kernel estimator of $E\left[DK_h\left(y-Y\right)\mid X\right]$ and the denominator is an kernel estimator of the propensity score $E\left[D\mid X\right]=\pi\left(X\right)$. Hence, $\hat{f}_{Y|D=1,X}\left(y\mid x\right)$ is an estimator of $E\left[DK_h^{(2)}\left(y-Y\right)\mid X\right]/\pi\left(X\right)$. Hence, we can use the machine learning estimator of $E\left[\frac{DK_h^{(2)}(y-Y)}{\pi(X)}\right]$ as an estimator for M_1. We can also construct a DML estimator using the Neyman-orthogonal functional form

$$\frac{DK_h^{(2)}\left(y-Y\right)}{\pi\left(X\right)}-\frac{D-\pi_0\left(X\right)}{\pi_0\left(X\right)}E\left[K_h^{(2)}\left(y-Y\right)\mid X,D=1\right]$$

In step 1, we use machine learning methods to estimate $\pi_0(X)$ and $E[K_h^{(2)}\left(\hat{\theta}_1-Y\right)\mid X,D=1]$ using auxiliary sample I_k^c. In Step 2, we construct the DML estimator of M_1:

$$\hat{M}_1=\frac{1}{K}\sum_{k=1}^{K}\sum_{i\in I_k}\frac{D_iK_h^{(2)}\left(\hat{\theta}_1-Y_i\right)}{\hat{\pi}\left(X_i\right)}-\frac{D_i-\hat{\pi}_0\left(X_i\right)}{\hat{\pi}_0\left(X_i\right)}\hat{E}\left[K_h^{(2)}\left(\hat{\theta}_1-Y\right)\mid X_i,D=1\right].$$

By the general DML theory (Chernozhukov et al., 2018), $\hat{M}_1$ is a consistent estimator of M_1. Similarly, we can construct the DML estimators for V_1, M_0, and V_0 using the following table:

Table 2.1: Orthogonal Scores

	Original Form	Equivalent Form
M_1	$E\left[f^{(2)}_{Y\mid X,D=1}\left(\theta_1^*\mid X\right)\right]$	$E\left[\frac{DK_h^{(2)}\left(\theta_1^*-Y\right)}{\pi(X)}-\frac{D-\pi_0(X)}{\pi_0(X)}E\left[K_h^{(2)}\left(y-Y\right)\mid X,D=1\right]\right]$
V_1	$\kappa_0^{(1)}E\left[\frac{f_{Y\mid X,D=1}\left(\theta_1^*\mid X\right)}{P\left(D=1\mid X\right)}\right]$	$E\left[\frac{DK_h\left(\theta_1^*-Y\right)}{\pi(X)^2}-2\frac{D-\pi_0(X)}{\pi_0^2(X)}E\left[K_h\left(y-Y\right)\mid X,D=1\right]\right]$
M_0	$E\left[f^{(2)}_{Y\mid X,D=0}\left(\theta_1^*\mid X\right)\right]$	$E\left[\frac{(1-D)K_h^{(2)}\left(\theta_1^*-Y\right)}{1-\pi(X)}-\frac{\pi_0(X)-D}{1-\pi_0(X)}E\left[K_h^{(2)}\left(y-Y\right)\mid X,D=0\right]\right]$
V_0	$\kappa_0^{(1)}E\left[\frac{f_{Y\mid X,D=0}\left(\theta_1^*\mid X\right)}{P\left(D=0\mid X\right)}\right]$	$E\left[\frac{(1-D)K_h\left(\theta_1^*-Y\right)}{(1-\pi(X))^2}-2\frac{\pi_0(X)-D}{(1-\pi_0(X))^2}E\left[K_h\left(y-Y\right)\mid X,D=0\right]\right]$

2.5 Conclusion

This paper studies the estimation and inference of the mode treatment effect, which has been ignored in the treatment effect literature compared to the estimation of the average and the quantile treatment effect estimation. I propose both kernel and ML estimators to accommodate a variety of data sets faced by researchers. I also derive the asymptotic properties of the proposed estimators. I show that both estimators are consistent and asymptotically normal with the rate of convergence $\sqrt{Nh^3}$.

2.A Appendix

Proof of Theorem 2.1: We only present the proof of the first claim, $\hat{\theta}_1 \xrightarrow{p} \theta_1^*$, since the second claim follows from the same arguments. The proof proceeds in two steps. In Step 1, we show the uniform law of large number holds

$$\sup_y \mid \hat{f}_{Y_1}(y) - f_{Y_1}(y) \mid = o_p(1).$$

In Step 2, we establish the consistency $\hat{\theta}_1 \xrightarrow{p} \theta_1^*$ using the same argument of Theorem 5.7 in Van der Vaart (2000).

Step 1. Notice that we have the decomposition

$$\begin{aligned}
\hat{f}_{Y_1}(y) - f_{Y_1}(y) &= \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{Y|D=1,X}(y \mid X_i) - E\left[f_{Y|D=1,X}(y \mid X)\right] \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_{Y|D=1,X}(y \mid X_i) - f_{Y|D=1,X}(y \mid X_i)\right)}_{A(y)} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_{Y|D=1,X}(y \mid X_i) - E\left[f_{Y|D=1,X}(y \mid X)\right]}_{B(y)}.
\end{aligned}$$

Hence,

$$\sup_{y} \mid \hat{f}_{Y_1}(y) - f_{Y_1}(y) \mid \leq \sup_{y} \mid A(y) \mid + \sup_{y} \mid B(y) \mid$$

By Theorem 6 in Hansen (2008) (uniform rates of convergence of kernel estimators), the first term $\sup_y \mid A(y) \mid$ is bounded by

$$\begin{aligned}
\sup_{y} |A(y)| &\leq \sup_{y} \frac{1}{n} \sum_{i=1}^{n} \left| \hat{f}_{Y|D=1,X}(y \mid X_i) - f_{Y|D=1,X}(y \mid X_i) \right| \\
&\leq \sup_{y} \sup_{x} \left| \hat{f}_{Y|D=1,X}(y \mid x) - f_{Y|D=1,X}(y \mid x) \right| \\
&\leq \sup_{x,y} \left| \hat{f}_{Y|D=1,X}(y \mid x) - f_{Y|D=1,X}(y \mid x) \right| \\
&= O_p\left(\sqrt{\frac{\ln n}{nh^{d+1}}} + h^2 \right) \\
&= o_p(1).
\end{aligned}$$

On the other hand, by Lemma 1 of Tauchen (1985) (uniform law of large numbers), we have

$$\sup_{y \in \mathcal{Y}} |B(y)| = \sup_{y \in \mathcal{Y}} \left| \frac{1}{n} \sum_{i=1}^{n} f_{Y|D=1,X}(y \mid X_i) - E\left[f_{Y|D=1,X}(y \mid X) \right] \right| \xrightarrow{p} 0.$$

Combining the results of $\sup_y \mid A(y) \mid$ and $\sup_y \mid B(y) \mid$ gives

$$\sup_{y} \mid \hat{f}_{Y_1}(y) - f_{Y_1}(y) \mid = o_p(1).$$

Step 2. The definition of $\hat{\theta}_1$ implies that $\hat{f}_{Y_1}\left(\hat{\theta}_1\right) \geq \hat{f}_{Y_1}(\theta_1^*)$. Therefore, we have

$$\begin{aligned}
f_{Y_1}(\theta_1^*) - f_{Y_1}\left(\hat{\theta}_1\right) &= f_{Y_1}(\theta_1^*) - \hat{f}_{Y_1}(\theta_1^*) + \hat{f}_{Y_1}(\theta_1^*) - f_{Y_1}\left(\hat{\theta}_1\right) \\
&\leq f_{Y_1}(\theta_1^*) - \hat{f}_{Y_1}(\theta_1^*) + \hat{f}_{Y_1}\left(\hat{\theta}_1\right) - f_{Y_1}\left(\hat{\theta}_1\right) \\
&\leq 2 \sup_{y} \mid \hat{f}_{Y_1}(y) - f_{Y_1}(y) \mid.
\end{aligned}$$

By Step 1, we have that for any $\delta > 0$,

$$P\left(f_{Y_1}\left(\theta_1^*\right) - f_{Y_1}\left(\hat{\theta}_1\right) > \delta\right) \leq P\left(\sup_y \mid \hat{f}_{Y_1}(y) - f_{Y_1}(y) \mid > \delta/2\right) \to 0.$$

Further, Assumption 2.1 implies that for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\sup_{y:|y-\theta_1^*|>\varepsilon} f_{Y_1}(y) < f_{Y_1}\left(\theta_1^*\right) - \delta.$$

Then the following inequality holds

$$\begin{aligned} P\left(\mid \hat{\theta}_1 - \theta_1^* \mid > \varepsilon\right) &\leq P\left(f_{Y_1}\left(\hat{\theta}_1\right) < f_{Y_1}\left(\theta_1^*\right) - \delta\right) \\ &\leq P\left(f_{Y_1}\left(\theta_1^*\right) - f_{Y_1}\left(\hat{\theta}_1\right) > \delta\right) \to 0. \end{aligned}$$

Thus, we prove the consistency $\hat{\theta}_1 \xrightarrow{p} \theta_1^*$.

Proof of Theorem 2.2: Here we focus on the result for $\hat{\theta}_1$ only. Notice that the first-order condition for $\hat{\theta}_1$ gives

$$\begin{aligned} 0 = \hat{f}_{Y_1}^{(1)}\left(\hat{\theta}_1\right) &= \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{Y|X,D=1}^{(1)}\left(\hat{\theta}_1 \mid X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{Y|X,D=1}^{(1)}\left(\theta_1^* \mid X_i\right) + \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{Y|X,D=1}^{(2)}\left(\tilde{\theta}_1 \mid X_i\right)\left(\hat{\theta}_1 - \theta_1^*\right), \end{aligned}$$

where $\tilde{\theta}_1 \in \left(\hat{\theta}_1, \theta_1^*\right)$. Then we have

$$\sqrt{nh^3}\left(\hat{\theta}_1 - \theta_1^*\right) = -\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{Y|X,D=1}^{(2)}\left(\tilde{\theta}_1 \mid X_i\right)\right]^{-1}\left(\frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n} \hat{f}_{Y|X,D=1}^{(1)}\left(\theta_1^* \mid X_i\right)\right).$$

The proof proceeds in six steps. In Step 1, we show that the first term of r.h.s converges to $M_1 = E\left[f_{Y|X,D=1}^{(2)}\left(\theta_1^* \mid X\right)\right]$ in probability. In Step 2-5, we show the asymptotic normality of the second term. Then, by Slutsky's theorem, we can show the asymptotic normality for $\hat{\theta}_1$. In Step 6, we show the asymptotic normality for $\hat{\Delta}$.

For convenience, we define $\gamma_{10}(x) \equiv f^{(1)}_{Y,X|D=1}(\theta_1^*, x)$, $\gamma_{20}(x) \equiv f_{X|D=1}(x)$, and

$$\begin{aligned}\hat{\gamma}_1(x) &\equiv \frac{1}{n}\sum_{j=1}^{n}\frac{D_j K_h^{(1)}\left(\theta_1^* - Y_j\right) K_h\left(x - X_j\right)}{P(D=1)}\\ \hat{\gamma}_2(x) &\equiv \frac{1}{n}\sum_{j=1}^{n}\frac{D_j K_h\left(x - X_j\right)}{P(D=1)}.\end{aligned}$$

In these notations, we can express $\hat{f}^{(1)}_{Y|X,D=1}(\theta_1^* \mid x)$ and $f^{(1)}_{Y|X,D=1}(\theta_1^* \mid x)$ as $\hat{\gamma}_1(x)/\hat{\gamma}_2(x)$ and $\gamma_{10}(x)/\gamma_{20}(x)$, respectively. Also, let $\gamma_0 = (\gamma_{10}, \gamma_{20})'$ and $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2)'$.

Step 1. In this step, we show that $n^{-1}\sum_{i=1}^{n}\hat{f}^{(2)}_{Y|X,D=1}\left(\tilde{\theta}_1 \mid X_i\right) \xrightarrow{p} E\left[f^{(2)}_{Y|X,D=1}(\theta_1^* \mid X)\right]$. Notice that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{f}^{(2)}_{Y|X,D=1}\left(\tilde{\theta}_1 \mid X_i\right) = \frac{1}{n}\sum_{i=1}^{n}f^{(2)}_{Y|X,D=1}(\theta_1^* \mid X_i) + A_1 + A_2$$

where

$$A_1 = \frac{1}{n}\sum_{i=1}^{n}\hat{f}^{(2)}_{Y|X,D=1}\left(\tilde{\theta}_1 \mid X_i\right) - f^{(2)}_{Y|X,D=1}\left(\tilde{\theta}_1 \mid X_i\right)$$

and

$$A_2 = \frac{1}{n}\sum_{i=1}^{n}f^{(2)}_{Y|X,D=1}\left(\tilde{\theta}_1 \mid X_i\right) - f^{(2)}_{Y|X,D=1}(\theta_1^* \mid X_i).$$

Since $\frac{1}{n}\sum_{i=1}^{n}f^{(2)}_{Y|X,D=1}(\theta_1^* \mid X_i) \xrightarrow{p} E\left[f^{(2)}_{Y|X,D=1}(\theta_1^* \mid X)\right]$ by the law of large numbers, we only have to show that $A_1 = o_p(1)$ and $A_2 = o_p(1)$. Note that

$$\begin{aligned}|A_1| &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\hat{f}^{(2)}_{Y|X,D=1}\left(\tilde{\theta}_1 \mid X_i\right) - f^{(2)}_{Y|X,D=1}\left(\tilde{\theta}_1 \mid X_i\right)\right|\\ &\leq \sup_{y,x}\left|\hat{f}^{(2)}_{Y|X,D=1}(y \mid x) - f^{(2)}_{Y|X,D=1}(y \mid x)\right|\\ &= O_p\left(\sqrt{\frac{\ln n}{nh^{d+5}}} + h^2\right)\\ &= o_p(1),\end{aligned}$$

where the first equality follows from the uniform rates of convergence of kernel estimators (Hansen, 2008). For A_2, we use the argument in Lemma 4.3 of Newey & McFadden (1994). By consistency of $\hat{\theta}_1$, and thus $\tilde{\theta}_1$, there is $\delta_n \to 0$ such that $\left\| \tilde{\theta}_1 - \theta_1^* \right\| \leq \delta_n$ with probability approaching to one. Define

$$\Delta_n \left(X_i \right) = \sup_{\left\| y - \theta_1^* \right\| \leq \delta_n} \left\| f_{Y|X,D=1}^{(2)} \left(y \mid X_i \right) - f_{Y|X,D=1}^{(2)} \left(\theta_1^* \mid X_i \right) \right\| .$$

By the continuity of $f_{Y|X,D=1}^{(2)} \left(y \mid X_i \right)$ at θ_1^*, $\Delta_n \left(X_i \right) \xrightarrow{p} 0$. Hence, by the dominated convergence theorem, we have $E \left[\Delta_n \left(X_i \right) \right] \to 0$. Then, by Markov's inequality,

$$P \left(\frac{1}{n} \sum_{i=1}^{n} \Delta_n \left(X_i \right) > \epsilon \right) \leq E \left[\Delta_n \left(X_i \right) \right] / \epsilon \to 0.$$

Therefore, we have

$$\left| A_2 \right| \leq \frac{1}{n} \sum_{i=1}^{n} \Delta_n \left(X_i \right) + o_p \left(1 \right) = o_p \left(1 \right) .$$

Step 2. In this step, we show

$$\frac{\sqrt{nh^3}}{n} \sum_{i=1}^{n} \hat{f}_{Y|X,D=1}^{(1)} \left(\theta_1^* \mid X_i \right) = \frac{\sqrt{nh^3}}{n} \sum_{i=1}^{n} f_{Y|X,D=1}^{(1)} \left(\theta_1^* \mid X_i \right) + \frac{\sqrt{nh^3}}{n} \sum_{i=1}^{n} G \left(Z_i, \hat{\gamma} - \gamma_0 \right) + o_p \left(1 \right) ,$$

where $G \left(z, \gamma \right) = \gamma_{20} \left(x \right)^{-1} \left[1, -\frac{\gamma_{10}(x)}{\gamma_{20}(x)} \right] \gamma \left(x \right)$ and $z = \left(y, x, d \right)$ denotes data observation. To do this, it suffices to show

$$\frac{\sqrt{nh^3}}{n} \sum_{i=1}^{n} \left[\hat{f}_{Y|X,D=1}^{(1)} \left(\theta_1^* \mid X_i \right) - f_{Y|X,D=1}^{(1)} \left(\theta_1^* \mid X_i \right) - G \left(Z_i, \hat{\gamma} - \gamma_0 \right) \right] = o_p \left(1 \right) .$$

Using the notation of γ, we have

$$\hat{f}_{Y|X,D=1}^{(1)} \left(\theta_1^* \mid x \right) - f_{Y|X,D=1}^{(1)} \left(\theta_1^* \mid x \right) = \frac{\hat{\gamma}_1 \left(x \right)}{\hat{\gamma}_2 \left(x \right)} - \frac{\gamma_{10} \left(x \right)}{\gamma_{20} \left(x \right)} .$$

The following argument follows from Newey & McFadden (1994). Consider the algebra relation $\tilde{a}/\tilde{b} - a/b = b^{-1}\left[1 - \tilde{b}^{-1}\left(\tilde{b} - b\right)\right]\left[\tilde{a} - a - (a/b)\left(\tilde{b} - b\right)\right]$. The linear part of the r.h.s is $b^{-1}\left[\tilde{a} - a - (a/b)\left(\tilde{b} - b\right)\right]$, and the remaining term is of higher order. By letting $a = \gamma_{10}$, $\tilde{a} = \hat{\gamma}_1$, $b = \gamma_{20}$, and $\tilde{b} = \hat{\gamma}_2$, this linear term corresponds to the linear functional $G\left(Z_i, \hat{\gamma} - \gamma_0\right)$. The remaining higher-order term will satisfy

$$
\begin{aligned}
&\mid \frac{\gamma_1(x)}{\gamma_2(x)} - \frac{\gamma_{10}(x)}{\gamma_{20}(x)} - G\left(z, \gamma - \gamma_0\right) \mid \\
&\leq \mid \gamma_2(x) \mid^{-1} \gamma_{20}(x)^{-1}\left[1 + \frac{\gamma_{10}(x)}{\gamma_{20}(x)}\right]\left[\left(\gamma_1(x) - \gamma_{10}(x)\right)^2 + \left(\gamma_2(x) - \gamma_{20}(x)\right)^2\right] \\
&\leq C \sup_{x \in \mathcal{X}} \left\|\gamma(x) - \gamma_0(x)\right\|^2
\end{aligned}
$$

for some constant C if γ_2 and γ_{20} are bounded away from zero. Hence Lemma 1 holds if $\sqrt{nh^3}\sup_{x\in\mathcal{X}}\left\|\hat{\gamma}(x) - \gamma_0(x)\right\|^2 \xrightarrow{p} 0$. By the uniform rates of convergence of kernel estimators (Hansen, 2008), we have

$$
\begin{aligned}
\sup_{x \in \mathcal{X}} \left\|\hat{\gamma}(x) - \gamma_0(x)\right\|^2 &= \sup_{x \in \mathcal{X}}\left(\left(\hat{\gamma}_1(x) - \gamma_{10}(x)\right)^2 + \left(\hat{\gamma}_1(x) - \gamma_{10}(x)\right)^2\right) \\
&\leq \sup_{x \in \mathcal{X}}\left(\hat{\gamma}_1(x) - \gamma_{10}(x)\right)^2 + \sup_{x \in \mathcal{X}}\left(\hat{\gamma}_2(x) - \gamma_{20}(x)\right)^2 \\
&= O_p\left[(\ln n)\left(nh^{d+3}\right)^{-1} + h^4\right] + O_p\left[(\ln n)\left(nh^{d}\right)^{-1} + h^4\right] \\
&= O_p\left[(\ln n)\left(nh^{d+3}\right)^{-1} + h^4\right].
\end{aligned}
$$

The rates of h and n imply that $\sqrt{nh^3}\sup_{x\in\mathcal{X}}\left\|\hat{\gamma}(x) - \gamma_0(x)\right\|^2 \xrightarrow{p} 0$.

Step 3. In this step, we show

$$
\begin{aligned}
\frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n}\hat{f}^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right) = \frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n}f^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right) + \sqrt{nh^3}\int G\left(z, \hat{\gamma} - \gamma_0\right)dF_0(z) \\
+ o_p(1),
\end{aligned}
$$

where F_0 is the c.d.f. of z. To do this, it suffices to show that

$$\sqrt{nh^3}\left\{\frac{1}{n}\sum_{i=1}^{n}G\left(Z_i,\hat{\gamma}-\gamma_0\right)-\int G\left(z,\hat{\gamma}-\gamma_0\right)dF_0\left(z\right)\right\}=o_p\left(1\right).$$

Let $\bar{\gamma}\equiv E\left[\hat{\gamma}\right]$ and by the linearity of $G\left(z,\gamma\right)$, we have the decomposition

$$G\left(z,\hat{\gamma}-\gamma_0\right)=G\left(z,\hat{\gamma}-\bar{\gamma}\right)+G\left(z,\bar{\gamma}-\gamma_0\right).$$

Therefore we just need to show that

$$\sqrt{nh^3}\left\{\frac{1}{n}\sum_{i=1}^{n}G\left(Z_i,\hat{\gamma}-\bar{\gamma}\right)-\int G\left(z,\hat{\gamma}-\bar{\gamma}\right)dF_0\left(z\right)\right\}=o_p\left(1\right)$$

and

$$\sqrt{nh^3}\left\{\frac{1}{n}\sum_{i=1}^{n}G\left(Z_i,\bar{\gamma}-\gamma_0\right)-\int G\left(z,\bar{\gamma}-\gamma_0\right)dF_0\left(z\right)\right\}=o_p\left(1\right).$$

The second condition holds by the central limit theorem since

$$\sqrt{nh^3}\left\{\frac{1}{n}\sum_{i=1}^{n}G\left(Z_i,\bar{\gamma}-\gamma_0\right)-\int G\left(z,\bar{\gamma}-\gamma_0\right)dF_0\left(z\right)\right\}=\sqrt{nh^3}O_p\left(n^{-1/2}\right)=o_p\left(1\right).$$

It remains to show the first condition. We follow the arguments in Newey & McFadden (1994). Define $q_j\equiv\left(\frac{D_jK_h^{(1)}\left(\theta_1^*-Y_j\right)}{P(D=1)},\frac{D_j}{P(D=1)}\right)'$, we can rewrite

$$\hat{\gamma}\left(x\right)=\begin{bmatrix}\hat{\gamma}_1\left(x\right)\\ \hat{\gamma}_2\left(x\right)\end{bmatrix}=\frac{1}{n}\sum_{j=1}^{n}q_jK_h\left(x-X_j\right).$$

We also define

$$\begin{aligned}
m\left(Z_i, Z_j\right) &= G\left[Z_i, q_j K_h\left(\cdot - X_j\right)\right] \\
m_1(z) &= \int m(z, \tilde{z})\, dF_0(\tilde{z}) = G(z, \bar{\gamma}) \\
m_2(z) &= \int m(\tilde{z}, z)\, dF_0(\tilde{z}) = \int G\left[\tilde{z}, qK_h(\cdot - X)\right] dF_0(\tilde{z}).
\end{aligned}$$

Then the l.h.s. of the first condition equals

$$\begin{aligned}
&\sqrt{nh^3}\left\{\frac{1}{n}\sum_{i=1}^{n} G(Z_i, \hat{\gamma} - \bar{\gamma}) - \int G(z, \hat{\gamma} - \bar{\gamma})\, dF_0(z)\right\} \\
&= \sqrt{nh^3}\left\{\frac{1}{n}\sum_{i=1}^{n} G(z, \hat{\gamma}) - \frac{1}{n}\sum_{i=1}^{n} G(z, \bar{\gamma}) - \int G(z, \hat{\gamma})\, dF_0(z) + \int G(z, \bar{\gamma})\, dF_0(z)\right\} \\
&= \sqrt{nh^3}\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} m\left(Z_i, Z_j\right) - \frac{1}{n}\sum_{i=1}^{n} m_1(Z_i) - \frac{1}{n}\sum_{i=1}^{n} m_2(Z_i) + E\left[m_1(z)\right]\right\} \\
&= \sqrt{nh^3} \times O_p\left\{E\left[\left\|m(Z_1, Z_1)\right|\right]/n + \left(E\left[\left\|m(Z_1, Z_2)\right|^2\right]\right)^{1/2}/n\right\},
\end{aligned}$$

where the last equality follows from Lemma 8.4 of Newey & McFadden (1994). The last term converges to zero in probability if we can control the convergence rates of $E\left[\left\|m(Z_1, Z_1)\right|\right]$ and $E\left[\left\|m(Z_1, Z_2)\right|^2\right]$. Notice that we have $\left|G(z, \gamma)\right| \le b(z)\left\|\gamma\right\|_2$ with

$$b(z) = \left\|f_{X|D=1}(x)^{-1}\left[1, -f^{(1)}_{Y|X,D=1}(\theta_1^*, x)\right]\right\|_2$$

where $\left\|\cdot\right\|_2$ denotes the ℓ_2 norm. Then $E\left[\left\|G(z, qK_h(\cdot - x))\right|\right] \le b(z)\, h^{-d}\left\|q\right\|_2$ by the boundedness of $K(u)$. By that $f_{X|D=1}(x)$ is bounded away from zero and $f_{Y|X,D=1}(\theta_1^*, x)$ is bounded from above, we have that $E\left[b(z)^2\right] \leq \infty$. Therefore, we have

$$\begin{aligned}
&\sqrt{nh^3} \times O_p\left\{E\left[\left|m(Z_1, Z_1)\right|\right]/n + \left(E\left[\left|m(Z_1, Z_2)\right|^2\right]\right)^{1/2}/n\right\} \\
&= \sqrt{nh^3} \times O_p\left\{E\left[\|q\|_2 b(Z_1)\right]/n + \left(E\left[\|q\|_2^2 b(Z_1)^2\right]\right)^{1/2}\left(nh^d\right)^{-1}\right\} \\
&= \sqrt{nh^3} \times O_p\left(n^{-1}h^{-d-2}\right) \\
&= o_p(1)
\end{aligned}$$

by the assumptions on n and h. The additional h^{-2} in the rates of convergence follows from that q contains $K_h^{(1)}(u) = h^{-2}K^{(1)}(u/h)$ with bounded $K^{(1)}(u)$.

Step 4. In this step, we show that

$$\frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n}\hat{f}_{Y|X,D=1}^{(1)}(\theta_1^* \mid X_i) = \frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n}f_{Y|X,D=1}^{(1)}(\theta_1^* \mid X_i) + \frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n}v(X_i)\,q_i + o_p(1),$$

where $v(X_i) = \frac{P(D=1)}{P\left(D=1|X_i\right)}\left[1, -\frac{\gamma_{10}(X_i)}{\gamma_{20}(X_i)}\right]$ and $q_i = \left(\frac{D_i K_h^{(1)}\left(\theta_1^* - Y_i\right)}{P(D=1)}, \frac{D_i}{P(D=1)}\right)'$. To do this, it suffices to show that

$$\sqrt{nh^3}\int G(z, \hat{\gamma} - \gamma_0)\,dF_0(z) - \frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n}v(X_i)\,q_i = o_p(1)$$

Notice that

$$\begin{aligned}
\int G(z,\gamma)\,dF_0(z) &= \int \gamma_{20}(x)^{-1}\left[1, -\frac{\gamma_{10}(x)}{\gamma_{20}(x)}\right]\gamma(x)\,f_X(x)\,dx \\
&= \int f_{X|D=1}(x)^{-1}\left[1, -\frac{\gamma_{10}(x)}{\gamma_{20}(x)}\right]\gamma(x)\,f_X(x)\,dx \\
&= \int \frac{P(D=1)}{P(D=1 \mid X=x)}\left[1, -\frac{\gamma_{10}(x)}{\gamma_{20}(x)}\right]\gamma(x)\,dx \\
&= \int v(x)\,\gamma(x)\,dx,
\end{aligned}$$

where $f_X(x)$ is the density function of X and $v(x) = \frac{P(D=1)}{P\left(D=1|X=x\right)} \left[1, -\frac{\gamma_{10}(x)}{\gamma_{20}(x)}\right]$. Also, we have

$$\begin{aligned} v(x)\gamma_0(x) &= \frac{P(D=1)}{P(D=1 \mid X=x)} \left[1, -\frac{\gamma_{10}(x)}{\gamma_{20}(x)}\right] \begin{bmatrix} \gamma_{10}(x) \\ \gamma_{20}(x) \end{bmatrix} \\ &= \frac{P(D=1)}{P(D=1 \mid X=x)} \left(\gamma_{10}(x) - \gamma_{10}(x)\right) \\ &= 0. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \int G(z, \hat{\gamma} - \gamma_0)\, dF_0(z) = \int v(x)\hat{\gamma}(x)\, dx - \int v(x)\gamma_0(x)\, dx &= \int v(x)\hat{\gamma}(x)\, dx \\ &= \frac{1}{n}\sum_{i=1}^{n} \int v(x) q_i K_h(x - X_i)\, dx \\ &= \frac{1}{n}\sum_{i=1}^{n} v(X_i) q_i + \left(\frac{1}{n}\sum_{i=1}^{n} \int v(x) q_i K_h(x - X_i)\, dx - \frac{1}{n}\sum_{i=1}^{n} v(X_i) q_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} v(X_i) q_i + \frac{1}{n}\sum_{i=1}^{n} \left[\int v(x) K_h(x - X_i)\, dx - v(X_i)\right] q_i. \end{aligned}$$

By Chebyshev's inequality, sufficient conditions for $\sqrt{nh^3}$ times the second term in the last line converging to zero in probability are that

$$\sqrt{nh^3} E\left[\left(\int v(x) K_h(x - X_i)\, dx - v(X_i)\right) q_i\right] \to 0$$

and

$$E\left[\|q_i\|^2 \left\|\int v(x) K_h(x - X_i)\, dx - v(X_i)\right\|^2\right] \to 0.$$

The expectation in the first condition is the difference of $E\left[\left(\int v(x) K_h(x - X_i)\, dx\right) q_i\right]$ and

$E\left[v\left(X_i\right)q_i\right]$. We begin with the second term $E\left[v\left(X_i\right)q_i\right]$. Notice that

$$\begin{aligned}
E\left[v\left(X_i\right)q_i\right] &= E\Big[v\left(X_i\right)E\left[q_i \mid X_i\right]\Big] \\
&= E\left[v\left(X_i\right)E\left[\begin{pmatrix} \frac{D_i K_h^{(1)}\left(\theta_1^* - Y_i\right)}{P(D=1)} \\ \frac{D_i}{P(D=1)} \end{pmatrix} \mid X_i\right]\right] \\
&= E\left[v\left(X_i\right)\frac{P\left(D=1 \mid X_i\right)}{P\left(D=1\right)}E\left[\begin{pmatrix} K_h^{(1)}\left(\theta_1^* - Y_i\right) \\ 1 \end{pmatrix} \mid X_i, D=1\right]\right]
\end{aligned}$$

by the law of iterated expectations. The inner conditional expectation in the last line satisfies

$$\begin{aligned}
E\left[K_h^{(1)}\left(\theta_1^* - Y_i\right) \mid X_i, D=1\right] &= \frac{1}{h^2}E\left[K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right) \mid X_i, D=1\right] \\
&= \frac{1}{h^2}\int K^{(1)}\left(\frac{\theta_1^* - y}{h}\right) f_{Y|X,D=1}\left(y \mid X_i\right)dy \\
&= \frac{1}{h}\int K\left(\frac{\theta_1^* - y}{h}\right) f^{(1)}_{Y|X,D=1}\left(y \mid X_i\right)dy \\
&= \int K\left(u\right) f^{(1)}_{Y|X,D=1}\left(\theta_1^* + hu \mid X_i\right)du \\
&= \int K\left(u\right) f^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right)du \\
&+ \int huK\left(u\right) f^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right)du \\
&+ \int \frac{h^2u^2}{2}K\left(u\right) f^{(3)}_{Y|X,D=1}\left(\tilde{\theta}_1 \mid X_i\right)du \\
&= f^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right) + \frac{h^2}{2}\kappa_2 f^{(3)}_{Y|X,D=1}\left(\tilde{\theta}_1 \mid X_i\right)
\end{aligned}$$

with $\tilde{\theta}_1 \in \left(\theta_1^*, \theta_1^* + hu\right)$ and $\kappa_2 = \int u^2K\left(u\right)du$. The third equality follows from integration by parts and the forth from change of variables. Hence,

$$
\begin{aligned}
E\left[v\left(X_i\right)q_i\right] &= E\left[v\left(X_i\right)\frac{P\left(D=1\mid X_i\right)}{P\left(D=1\right)}\begin{pmatrix} f^{(1)}_{Y|X,D=1}\left(\theta_1^*\mid X_i\right) \\ 1 \end{pmatrix}\right] \\
&\quad + \frac{h^2}{2}\kappa_2 E\left[v\left(X_i\right)\frac{P\left(D=1\mid X_i\right)}{P\left(D=1\right)}\begin{pmatrix} f^{(3)}_{Y|X,D=1}\left(\tilde{\theta}_1\mid X_i\right) \\ 0 \end{pmatrix}\right] \\
&= E\left[v\left(X_i\right)\frac{P\left(D=1\mid X_i\right)}{P\left(D=1\right)}\begin{pmatrix} f^{(1)}_{Y|X,D=1}\left(\theta_1^*\mid X_i\right) \\ 1 \end{pmatrix}\right] + O\left(h^2\right) \\
&= \int v\left(x\right)\frac{P\left(D=1\mid X_i=x\right)}{P\left(D=1\right)}\begin{pmatrix} f^{(1)}_{Y|X,D=1}\left(\theta_1^*\mid X_i=x\right) \\ 1 \end{pmatrix} f_X\left(x\right)dx + O\left(h^2\right) \\
&= \int v\left(x\right)\begin{pmatrix} f^{(1)}_{Y|X,D=1}\left(\theta_1^*\mid X_i=x\right) \\ 1 \end{pmatrix} f_{X|D=1}\left(x\right)dx + O\left(h^2\right) \\
&= \int v\left(x\right)\begin{pmatrix} f^{(1)}_{Y,X|D=1}\left(\theta_1^*\mid X_i=x\right) \\ f_{X|D=1}\left(x\right) \end{pmatrix} dx + O\left(h^2\right) \\
&= \int v\left(x\right)\gamma_0\left(x\right)dx + O\left(h^2\right).
\end{aligned}
$$

Using the same arguments, we can also show that

$$
\begin{aligned}
E\left[\left(\int v\left(x\right)K_h\left(x-X_i\right)dx\right)q_i\right] &= E\left[\left(\int v\left(X_i+hu\right)K\left(u\right)du\right)q_i\right] \\
&= \int\left(\int v\left(x+hu\right)K\left(u\right)du\right)\gamma_0\left(x\right)dx + O\left(h^2\right)
\end{aligned}
$$

Then the first condition equals

$$
\begin{aligned}
&\sqrt{nh^3}\left\|E\left[\left(\int v\left(x\right)K_h\left(x-X_i\right)dx - v\left(X_i\right)\right)q_i\right]\right\| \\
&= \sqrt{nh^3}\left\|\int\left(\int v\left(x+hu\right)K\left(u\right)du\right)\gamma_0\left(x\right)dx - \int v\left(x\right)\gamma_0\left(x\right)dx + O\left(h^2\right)\right\|.
\end{aligned}
$$

Following the argument in Theorem 8.11 of Newey & McFadden (1994), the last line satisfies

$$\begin{aligned}
&\sqrt{nh^3}\left\|\int\int v\left(x\right)K\left(u\right)\gamma_0\left(x-hu\right)dudx-\int v\left(x\right)\gamma_0\left(x\right)dx+O\left(h^2\right)\right\| \\
&=\sqrt{nh^3}\left\|\int v\left(x\right)\left\{\int\left[\gamma_0\left(x-hu\right)-\gamma_0\left(x\right)\right]du\right\}dx+O\left(h^2\right)\right\| \\
&\leq\sqrt{nh^3}\int\left\|v\left(x\right)\right\|\left\|\int\left[\gamma_0\left(x-hu\right)-\gamma_0\left(x\right)\right]du\right\|dx+O\left(\sqrt{nh^3}h^2\right) \\
&\leq\sqrt{nh^3}Ch^2\int\left\|v\left(x\right)\right\|dx+O\left(\sqrt{nh^3}h^2\right) \\
&=O\left(\sqrt{nh^3}h^2\right).
\end{aligned}$$

Therefore the first condition holds if $\sqrt{nh^3}h^2\to 0$.

Recall that the second condition we would like to show is

$$E\left[\left\|q_i\right\|^2\left\|\int v\left(x\right)K_h\left(x-X_i\right)dx-v\left(X_i\right)\right\|^2\right]\to 0.$$

By Cauchy Schwartz inequality, it suffices to show that

$$E\left[\left\|\int v\left(x\right)K_h\left(x-X_i\right)dx-v\left(X_i\right)\right\|^4\right]\to 0.$$

By the continuity of $v\left(x\right)$, $v\left(x+hu\right)\to v\left(x\right)$ for all x and u as $h\to 0$. By the dominated convergence theorem, $\int v\left(x\right)K_h\left(x-x_i\right)dx=\int v\left(x+hu\right)K\left(u\right)du\to\int v\left(x\right)K\left(u\right)du=v\left(x\right)$ for all x. Therefore we have

$$E\left[\left\|\int v\left(x\right)K_h\left(x-X_i\right)dx-v\left(X_i\right)\right\|^4\right]=E\left[\left\|\int v\left(X_i+hu\right)K\left(u\right)du-v\left(X_i\right)\right\|^4\right]\to 0.$$

Step 5. By Step 4 and the definition of $v(X_i)$ and q_i, we have

$$\begin{aligned}
\frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n}\hat{f}^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right) &= \frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n} f^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right) + \frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n} v\left(X_i\right) q_i + o_p\left(1\right) \\
&= \frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n} f^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right) + o_p\left(1\right) \\
&+ \frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n} \frac{D_i}{P\left(D=1 \mid X_i\right)}\left[K_h^{(1)}\left(\theta_1^* - Y_i\right) - f^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right)\right] \\
&= \frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n} f^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right)\left[1 - \frac{D_i}{P\left(D=1 \mid X_i\right)}\right] \\
&+ \frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n} \frac{D_i}{P\left(D=1 \mid X_i\right)} K_h^{(1)}\left(\theta_1^* - Y_i\right) + o_p\left(1\right).
\end{aligned}$$

Since we have $E\left[f^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right)\left[1 - \frac{D_i}{P\left(D=1|X_i\right)}\right]\right] = 0$ by the law of iterated expectations, the central limit theorem holds for the first term of r.h.s. Hence,

$$\begin{aligned}
\frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n}\hat{f}^{(1)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right) &= O_p\left(\sqrt{nh^3}n^{-1/2}\right) + \frac{\sqrt{nh^3}}{nh^2}\sum_{i=1}^{n}\frac{D_i}{P\left(D=1 \mid X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right) \\
&+ o_p\left(1\right) \\
&= O_p\left(\sqrt{h^3}\right) + \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\frac{D_i}{P\left(D=1 \mid X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right) + o_p\left(1\right) \\
&= \frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\frac{D_i}{P\left(D=1 \mid X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right) + o_p\left(1\right).
\end{aligned}$$

In this step, we show that

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\frac{D_i}{P\left(D=1 \mid X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right) \xrightarrow{d} N\left(0, V\right),$$

where $V = \kappa_0^{(1)} E\left[\frac{f_{Y|X,D=1}\left(\theta_1^*|X\right)}{P\left(D=1|X\right)}\right]$ and $\kappa_0^{(1)} = \int K^{(1)}\left(u\right)^2 du$.

For convenience, we define $\hat{g}\left(\theta_1^*\right) \equiv \left(nh^2\right)^{-1}\sum_{i=1}^{n}\frac{D_i}{P\left(D=1|X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right)$. Then it is equivalent to show that

$$\sqrt{nh^3}\left(\hat{g}\left(\theta_1^*\right)-0\right)\xrightarrow{d}N\left(0,V\right).$$

To use central limit theorem, we have to calculate $E\left[\hat{g}\left(\theta_1^*\right)\right]$ and $Var\left(\hat{g}\left(\theta_1^*\right)\right)$.

$$\begin{aligned}
E\left[\hat{g}\left(\theta_1^*\right)\right] &= \frac{1}{h^2}E\left[\frac{D_i}{P\left(D=1\mid X_i\right)}K^{(1)}\left(\frac{\theta_1^*-Y_i}{h}\right)\right]\\
&= \frac{1}{h^2}E\left[\frac{1}{P\left(D=1\mid X_i\right)}E\left[D_iK^{(1)}\left(\frac{\theta_1^*-Y_i}{h}\right)\mid X_i\right]\right]\\
&= \frac{1}{h^2}E\left[E\left[K^{(1)}\left(\frac{\theta_1^*-Y_i}{h}\right)\mid X_i, D=1\right]\right].
\end{aligned}$$

Since $h^{-2}E\left[K^{(1)}\left(\frac{\theta_1^*-Y_i}{h}\right)\mid X_i, D=1\right]=f^{(1)}_{Y|X,D=1}\left(\theta_1^*\mid X_i\right)+\frac{h^2}{2}\kappa_2 f^{(3)}_{Y|X,D=1}\left(\tilde{\theta}_1\mid X_i\right)$ from the calculation in Step 4, then

$$E\left[\hat{g}\left(\theta_1^*\right)\right]=E\left[f^{(1)}_{Y|X,D=1}\left(\theta_1^*\mid X_i\right)\right]+O\left(h^2\right)=0+O\left(h^2\right).$$

For the variance,

$$
\begin{aligned}
Var\left(\hat{g}\left(\theta_1^*\right)\right) &= \frac{1}{nh^4} Var\left(\frac{D_i}{P\left(D=1 \mid X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right)\right) \\
&= \frac{1}{nh^4} E\left[\left(\frac{D_i}{P\left(D=1 \mid X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right)\right)^2\right] \\
&- \frac{1}{nh^4}\left(E\left[\frac{D_i}{P\left(D=1 \mid X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right)\right]\right)^2 \\
&= \frac{1}{nh^4} E\left[\left(\frac{D_i}{P\left(D=1 \mid X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right)\right)^2\right] + \frac{1}{nh^4} O\left(h^4\right) \\
&= \frac{1}{nh^4} E\left[\frac{1}{P\left(D=1 \mid X_i\right)^2} E\left[D_i^2 K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right)^2 \mid X\right]\right] + \frac{1}{nh^4} O\left(h^4\right) \\
&= \frac{1}{nh^4} E\left[\frac{1}{P\left(D=1 \mid X_i\right)} E\left[K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right)^2 \mid X, D=1\right]\right] + \frac{1}{nh^4} O\left(h^4\right).
\end{aligned}
$$

The inner expectation in the last line equals

$$
\begin{aligned}
E\left[K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right)^2 \mid X, D=1\right] &= \int K^{(1)}\left(\frac{\theta_1^* - y}{h}\right)^2 f_{Y|X,D=1}\left(y \mid X\right) dy \\
&= h \int K^{(1)}\left(u\right)^2 f_{Y|X,D=1}\left(\theta_1^* + hu \mid X\right) du \\
&= h f_{Y|X,D=1}\left(\theta_1^* \mid X\right) \int K^{(1)}\left(u\right)^2 du \\
&+ h^2 f_{Y|X,D=1}^{(1)}\left(\tilde{\theta}_1 \mid X\right) \int u K^{(1)}\left(u\right)^2 du,
\end{aligned}
$$

where $\tilde{\theta}_1 \in \left(\theta_1^*, \theta_1^* + hu\right)$. Define $\kappa_0^{(1)} = \int K^{(1)}\left(u\right)^2 du$ and $\kappa_1^{(1)} = \int u K^{(1)}\left(u\right)^2 du$, the variance equals

$$\begin{aligned}
Var\left(\hat{g}\left(\theta_1^*\right)\right) &= \frac{1}{nh^4} E\left[\frac{h}{P\left(D=1 \mid X_i\right)} \kappa_0^{(1)} f_{Y \mid X, D=1}\left(\theta_1^* \mid X\right)\right] \\
&+ \frac{1}{nh^4} E\left[\frac{h^2}{P\left(D=1 \mid X_i\right)} \kappa_1^{(1)} f_{Y \mid X, D=1}^{(1)}\left(\theta_1^* \mid X\right)\right] + \frac{1}{nh^4} O\left(h^4\right) \\
&= \frac{1}{nh^3}\left(\kappa_0^{(1)} E\left[\frac{1}{P\left(D=1 \mid X_i\right)} f_{Y \mid X, D=1}\left(\theta_1^* \mid X\right)\right] + O(h) + O\left(h^3\right)\right) \\
&= \frac{1}{nh^3}\left(V + O(h) + O\left(h^3\right)\right).
\end{aligned}$$

Then we are ready to apply the central limit theorem.

Let

$$Z_{n,i} \equiv (nh)^{-1/2}\left(\frac{D_i}{P\left(D=1 \mid X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right) - E\left[\frac{D_i}{P\left(D=1 \mid X_i\right)} K^{(1)}\left(\frac{\theta_1^* - Y_i}{h}\right)\right]\right),$$

then $E\left[Z_{n,i}\right] = 0$ and $Var\left(Z_{n,i}\right) = h^3 Var\left(\hat{g}\left(\theta_1^*\right)\right) = n^{-1}V + o\left(n^{-1}\right)$. Then

$$\begin{aligned}
\sqrt{nh^3}\left(\hat{g}\left(\theta_1^*\right) - 0\right) &= \sqrt{nh^3}\left(\hat{g}\left(\theta_1^*\right) - E\left[\hat{g}\left(\theta_1^*\right)\right]\right) + \sqrt{nh^3}\left(E\left[\hat{g}\left(\theta_1^*\right)\right] - 0\right) \\
&= \sqrt{nh^3}\left(\hat{g}\left(\theta_1^*\right) - E\left[\hat{g}\left(\theta_1^*\right)\right]\right) + \sqrt{nh^3} O\left(h^2\right) \\
&= \sum_{i=1}^{n} Z_{n,i} + \sqrt{nh^3} O\left(h^2\right) \\
&\xrightarrow{d} N(0, V)
\end{aligned}$$

by Liapunov CLT and $\sqrt{nh^3}h^2 \to 0$.

Step 6. In this step, we show that

$$\sqrt{nh^3}\begin{bmatrix} \hat{\theta}_1 - \theta_1^* \\ \hat{\theta}_0 - \theta_0^* \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} M_1 V_1 M_1 & 0 \\ 0 & M_0 V_0 M_0 \end{bmatrix}\right)$$

and thus, by the delta method, we have

$$\sqrt{nh^3}\left(\hat{\Delta}-\Delta^*\right)\xrightarrow{d} N\left(0, M_1V_1M_1+M_0V_0M_0\right).$$

To show the joint distribution we adopt vector notations. The first-order conditions of $\hat{\theta}_1$ and $\hat{\theta}$ give

$$\begin{bmatrix}0\\0\end{bmatrix}=\begin{bmatrix}\hat{f}_{Y_1}\left(\hat{\theta}_1\right)\\\hat{f}_{Y_0}\left(\hat{\theta}_0\right)\end{bmatrix}=\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}\hat{f}^{(1)}_{Y|X,D=1}\left(\hat{\theta}_1\mid X_i\right)\\\hat{f}^{(1)}_{Y|X,D=0}\left(\hat{\theta}_0\mid X_i\right)\end{bmatrix}=\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}\hat{f}^{(1)}_{Y|X,D=1}\left(\theta_1^*\mid X_i\right)\\\hat{f}^{(1)}_{Y|X,D=0}\left(\theta_0^*\mid X_i\right)\end{bmatrix}+J_n\begin{bmatrix}\hat{\theta}_1-\theta_1^*\\\hat{\theta}_0-\theta_0^*\end{bmatrix}$$

where

$$\begin{aligned}J_n&=\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}\frac{\partial\hat{f}^{(1)}_{Y|X,D=1}\left(\tilde{\theta}_1|X_i\right)}{\partial\theta_1}&\frac{\partial\hat{f}^{(1)}_{Y|X,D=1}\left(\tilde{\theta}_1|X_i\right)}{\partial\theta_0}\\\frac{\partial\hat{f}^{(1)}_{Y|X,D=0}\left(\tilde{\theta}_0|X_i\right)}{\partial\theta_1}&\frac{\partial\hat{f}^{(1)}_{Y|X,D=0}\left(\tilde{\theta}_0|X_i\right)}{\partial\theta_0}\end{bmatrix}\\&=\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}\hat{f}^{(2)}_{Y|X,D=1}\left(\tilde{\theta}_1\mid X_i\right)&0\\0&\hat{f}^{(2)}_{Y|X,D=0}\left(\tilde{\theta}_0\mid X_i\right)\end{bmatrix}.\end{aligned}$$

Hence we have

$$\begin{aligned}\sqrt{nh^3}\begin{bmatrix}\hat{\theta}_1-\theta_1^*\\\hat{\theta}_0-\theta_0^*\end{bmatrix}&=\left(J_n^{-1}\right)\frac{\sqrt{nh^3}}{n}\sum_{i=1}^{n}\begin{bmatrix}\hat{f}^{(1)}_{Y|X,D=1}\left(\hat{\theta}_1\mid X_i\right)\\\hat{f}^{(1)}_{Y|X,D=0}\left(\hat{\theta}_0\mid X_i\right)\end{bmatrix}\\&=\left(J_n^{-1}\right)\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\begin{bmatrix}\frac{D_i}{P\left(D=1|X_i\right)}K^{(1)}\left(\frac{\theta_1^*-Y_i}{h}\right)\\\frac{1-D_i}{P\left(D=0|X_i\right)}K^{(1)}\left(\frac{\theta_0^*-Y_i}{h}\right)\end{bmatrix}+\begin{bmatrix}o_p\left(1\right)\\o_p\left(1\right)\end{bmatrix}\end{aligned}$$

where the last equality follows from the Step 5 in the proof of Theorem 2.1. Since

$$J_n\xrightarrow{p}\begin{bmatrix}M_1&0\\0&M_0\end{bmatrix}$$

and

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\begin{bmatrix}\frac{D_i}{P\left(D=1|X_i\right)}K^{(1)}\left(\frac{\theta_1^*-Y_i}{h}\right)\\ \frac{1-D_i}{P\left(D=0|X_i\right)}K^{(1)}\left(\frac{\theta_0^*-Y_i}{h}\right)\end{bmatrix}\xrightarrow{d}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}V_1 & 0\\ 0 & V_0\end{bmatrix}\right),$$

then by Slutsky's theorem we have

$$\sqrt{nh^3}\begin{bmatrix}\hat{\theta}_1-\theta_1^*\\ \hat{\theta}_0-\theta_0^*\end{bmatrix}\xrightarrow{d}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}M_1V_1M_1 & 0\\ 0 & M_0V_0M_0\end{bmatrix}\right).$$

Proof of Theorem 2.3. It is enough to show the results of $\hat{M}_1$ and $\hat{V}_1$. We first show that $\hat{M}_1\xrightarrow{p}M_1$. By adding and subtracting additional terms, we have

$$\hat{M}_1=\frac{1}{n}\sum_{i=1}^{n}\hat{f}^{(2)}_{Y|X,D=1}\left(\hat{\theta}_1\mid X_i\right)=\frac{1}{n}\sum_{i=1}^{n}f^{(2)}_{Y|X,D=1}\left(\theta_1^*\mid X_i\right)+A_1+A_2$$

where

$$A_1=\frac{1}{n}\sum_{i=1}^{n}\hat{f}^{(2)}_{Y|X,D=1}\left(\hat{\theta}_1\mid X_i\right)-f^{(2)}_{Y|X,D=1}\left(\hat{\theta}_1\mid X_i\right)$$

and

$$A_2=\frac{1}{n}\sum_{i=1}^{n}f^{(2)}_{Y|X,D=1}\left(\hat{\theta}_1\mid X_i\right)-f^{(2)}_{Y|X,D=1}\left(\theta_1^*\mid X_i\right).$$

If we can show that $A_1=o_p(1)$ and $A_2=o_p(1)$, then $\hat{M}\xrightarrow{p}M$ by law of large numbers.

Note that

$$\begin{aligned}
|A_1| &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\hat{f}^{(2)}_{Y|X,D=1}\left(\hat{\theta}_1 \mid X_i\right) - f^{(2)}_{Y|X,D=1}\left(\hat{\theta}_1 \mid X_i\right)\right| \\
&\leq \sup_{y,x}\left|\hat{f}^{(2)}_{Y|X,D=1}\left(y \mid x\right) - f^{(2)}_{Y|X,D=1}\left(y \mid x\right)\right| \\
&= O_p\left(\sqrt{\frac{\ln n}{nh^{d+5}}} + h^2\right) \\
&= o_p\left(1\right),
\end{aligned}$$

where the first equality follows from the uniform rates of convergence of kernel estimators (Hansen, 2008). For A_2, we use the argument in Lemma 4.3 of Newey & McFadden (1994). By consistency of $\hat{\theta}_1$ there is $\delta_n \to 0$ such that $\left\|\hat{\theta}_1 - \theta_1^*\right\| \leq \delta_n$ with probability approaching to one. Define $\Delta_n\left(Z_i\right) = \sup_{\left\|y-\theta_1^*\right\|\leq\delta_n}\left\|f^{(2)}_{Y|X,D=1}\left(y \mid X_i\right) - f^{(2)}_{Y|X,D=1}\left(\theta_1^* \mid X_i\right)\right\|$. By the continuity of $f^{(2)}_{Y|X,D=1}\left(y \mid X_i\right)$ at θ_1^*, $\Delta_n\left(Z_i\right) \xrightarrow{p} 0$. By the dominated convergence theorem, we have $E\left[\Delta_n\left(Z_i\right)\right] \to 0$. By Markov inequality, $P\left(n^{-1}\sum_{i=1}^{n}\Delta_n\left(Z_i\right) > \epsilon\right) \leq E\left[\Delta_n\left(Z_i\right)\right]/\epsilon \to 0$. Therefore, we have

$$|A_2| \leq \frac{1}{n}\sum_{i=1}^{n}\Delta_n\left(Z_i\right) = o_p\left(1\right).$$

Next we show that $\hat{V}_1 \xrightarrow{p} V_1$. We can rewrite

$$\hat{V}_1/\kappa_0^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{f}_{Y|X,D=1}\left(\hat{\theta}_1 \mid X_i\right)}{\hat{\pi}\left(X_i\right)} = \frac{1}{n}\sum_{i=1}^{n}\frac{f_{Y|X,D=1}\left(\theta_1^* \mid X_i\right)}{\pi\left(X_i\right)} + B_1 + B_2$$

with

$$B_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{f}_{Y|X,D=1}\left(\hat{\theta}_1 \mid X_i\right)}{\hat{\pi}\left(X_i\right)} - \frac{f_{Y|X,D=1}\left(\hat{\theta}_1 \mid X_i\right)}{\pi\left(X_i\right)}$$

and

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}\frac{f_{Y|X,D=1}\left(\hat{\theta}_1 \mid X_i\right)}{\pi\left(X_i\right)} - \frac{f_{Y|X,D=1}\left(\theta_1^* \mid X_i\right)}{\pi\left(X_i\right)}.$$

It remains to show that $B_1 = o_p\left(1\right)$ and $B_2 = o_p\left(1\right)$. The result of B_2 follows from the same arguments as in the proof of A_2 if $f_{Y|X,D=1}(y \mid X_i)$ is continuous at θ_1^*. Thus, we only focus on B_1. For conenience, define $f(y \mid x) = f_{Y|X,D=1}(y \mid x)$. For π bounded away from zero, we have

$$
\begin{aligned}
\frac{\hat{f}(y \mid x)}{\hat{\pi}(x)} - \frac{f(y \mid x)}{\pi(x)} &= \frac{\pi(x)\hat{f}(y \mid x) - \hat{\pi}(x) f(y \mid x)}{\hat{\pi}(x)\pi(x)} \\
&= \frac{\pi(x)\hat{f}(y \mid x) - \pi(x) f(y \mid x) + \pi(x) f(y \mid x) - \hat{\pi}(x) f(y \mid x)}{\hat{\pi}(x)\pi(x)} \\
&= \frac{\hat{f}(y \mid x) - f(y \mid x)}{\hat{\pi}(x)} + \frac{f(y \mid x)}{\hat{\pi}(x)\pi(x)}(\hat{\pi}(x) - \pi(x)) \\
&\leq C\left(\left(\hat{f}(y \mid x) - f(y \mid x)\right) + (\hat{\pi}(x) - \pi(x))\right)
\end{aligned}
$$

for some $C > 0$. By the uniform rates of convergence of kernel estimators (Hansen, 2008), we have

$$
\begin{aligned}
|B_1| &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{f}_{Y|X,D=1}\left(\hat{\theta}_1 \mid X_i\right)}{\hat{\pi}(X_i)} - \frac{f_{Y|X,D=1}\left(\hat{\theta}_1 \mid X_i\right)}{\pi(X_i)}\right| \\
&\leq C\sup_{y,x}\left(\left|\hat{f}(y \mid x) - f(y \mid x)\right| + |\hat{\pi}(x) - \pi(x)|\right) \\
&= O_p\left(\sqrt{\frac{\ln n}{nh^{d+1}}} + h^2\right) + \sup_x |\hat{\pi}(x) - \pi(x)| \\
&= o_p(1)
\end{aligned}
$$

by the rates of n and h and the uniform convergence of $\hat{\pi}(x)$.

Proof of Theorem 2.4: Suppose that

$$
\hat{f}_{Y_1}(y) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}[m_1(Z, y, \hat{\eta}_{1k})]
$$

is differentiable with respect to y. Define

$$\hat{f}_{Y_1}^{(1)}(y) \equiv \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}[m_1^{(1)}(Z, y, \hat{\eta}_{1k})]$$

where $m_1^{(1)}(Z, y, \hat{\eta}_{1k}) \equiv \partial m_1(Z, y, \hat{\eta}_{1k})/\partial y$.

By the definition of $\hat{\theta}_1$, we have

$$\begin{aligned}
0 = \hat{f}_{Y_1}^{(1)}(\hat{\theta}_1) &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}[m_1^{(1)}(Z, \hat{\theta}_1, \hat{\eta}_{1k})] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}[m_1^{(1)}(Z, \theta_1^*, \hat{\eta}_{1k})] + \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}[m_1^{(2)}(Z, \tilde{\theta}_1, \hat{\eta}_{1k})](\hat{\theta}_1 - \theta_1^*)
\end{aligned}$$

and

$$\sqrt{Nh^3}(\hat{\theta}_1 - \theta_1^*) = -\left[\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}[m_1^{(2)}(Z, \tilde{\theta}_1, \hat{\eta}_{1k})]\right]^{-1}\left(\frac{\sqrt{Nh^3}}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}[m_1^{(1)}(Z, \theta_1^*, \hat{\eta}_{1k})]\right).$$

In Step 1 and 2 below, we will show that

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}[m_1^{(2)}(Z, \tilde{\theta}_1, \hat{\eta}_{1k})] \xrightarrow{p} M_1$$

and

$$\frac{\sqrt{Nh^3}}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}[m_1^{(1)}(Z, \theta_1^*, \hat{\eta}_{1k})] \xrightarrow{d} N(0, V_1),$$

respectively. Hence, we can obtain the final result

$$\sqrt{Nh^3}(\hat{\theta}_1 - \theta_1^*) \xrightarrow{d} N(0, M_1^{-1}V_1M_1^{-1}).$$

Step 1. Since K is a fixed integer, which is independent of N, it suffices to show that for each $k \in [K]$,

$$\mathbb{E}_{n,k}[m_1^{(2)}(Z, \tilde{\theta}_1, \hat{\eta}_{1k})] \xrightarrow{p} M_1.$$

Then we can show this convergence using the same argument in Step 1 in the proof of Theorem 2.2.

Step 2. Since K is a fixed integer, which is independent of N, it is enough to consider the convergence of $\mathbb{E}_{n,k}[m_1^{(1)}(Z, \theta_1^*, \hat{\eta}_{1k})]$. Notice that

$$\mathbb{E}_{n,k}[m_1^{(1)}(Z, \theta_1^*, \hat{\eta}_{1k})] = \frac{1}{n} \sum_{i \in I_k} m_1^{(1)}(Z, \theta_1^*, \eta_{10}) + R_{2k}$$

where

$$R_{2,k} = \mathbb{E}_{n,k}[m_1^{(1)}(Z, \theta_1^*, \hat{\eta}_{1k})] - \frac{1}{n} \sum_{i \in I_k} m_1^{(1)}(Z, \theta_1^*, \eta_{10}).$$

Then by triangular inequality,

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

where

$$I_{1,k} \equiv \left\| \mathbb{G}_{n,k} \left[m_1^{(1)}(Z, \theta_1^*, \hat{\eta}_{1k}) \right] - \mathbb{G}_{n,k} \left[m_1^{(1)}(Z, \theta_1^*, \eta_{10}) \right] \right\|,$$

$$I_{2,k} \equiv \sqrt{n} \left\| E_P \left[m_1^{(1)}(Z, \theta_1^*, \hat{\eta}_{1k}) \mid (W_i)_{i \in I_k^c} \right] - E_P \left[m_1^{(1)}(Z, \theta_1^*, \eta_{10}) \right] \right\|.$$

Two auxiliary results will be used to bound $I_{1,k}$ and $I_{2,k}$:

$$\sup_{\eta_1 \in \mathcal{T}_N} \left(E \left[\| \, m_1^{(1)} \left(Z, \theta_1^*, \eta_1 \right) - m_1^{(1)} \left(Z, \theta_1^*, \eta_{10} \right) \|^2 \right] \right)^{1/2} \leq \varepsilon_N, \tag{A.1}$$

$$\sup_{r \in (0,1), \eta_1 \in \mathcal{T}_N} \| \, \partial_r^2 E \left[m_1^{(1)} \left(Z, \theta_1^*, \eta_{10} + r \left(\eta_1 - \eta_{10} \right) \right) \right] \| \leq \left(\varepsilon_N \right)^2, \tag{A.2}$$

where $\mathcal{T}_N$ is the set of all $\eta_1 = (\pi_0, g_{10})$ consisting of square-integrable functions π_0 and g_{10} such that

$$\| \, \eta_1 - \eta_{10} \, \|_{P,2} \leq \varepsilon_N,$$

$$\parallel \pi - 1/2 \parallel_{P,\infty} \leq 1/2 - \kappa,$$

$$\parallel \pi - \pi_0 \parallel_{P,2}^2 + \parallel \pi - \pi_0 \parallel_{P,2} \times \parallel g_1 - g_{10} \parallel_{P,2} \leq (\varepsilon_N)^2 .$$

Then by assumption, we have $\hat{\eta}_{1k} \in \mathcal{T}_N$ with probability $1 - o(1)$.

To bound $I_{1,k}$, note that conditional on $(W_i)_{i \in I_k^c}$ the estimator $\hat{\eta}_{1k}$ is nonstochastic. Under the event that $\hat{\eta}_{1k} \in \mathcal{T}_N$, we have

$$\begin{aligned} E_P\left[I_{1,k}^2 \mid (W_i)_{i \in I_k^c}\right] =& E_P\left[\parallel m_1^{(1)}(Z, \theta_1^*, \hat{\eta}_{1k}) - m_1^{(1)}(Z, \theta_1^*, \eta_{10}) \parallel^2 \mid (W_i)_{i \in I_k^c}\right] \\ &\leq \sup_{\eta_1 \in \mathcal{T}_N} E_P\left[\parallel m_1^{(1)}(Z, \theta_1^*, \eta_1) - m_1^{(1)}(Z, \theta_1^*, \eta_{10}) \parallel^2 \mid (W_i)_{i \in I_k^c}\right] \\ &= \sup_{\eta_1 \in \mathcal{T}_N} E_P\left[\parallel m_1^{(1)}(Z, \theta_1^*, \eta_1) - m_1^{(1)}(Z, \theta_1^*, \eta_{10}) \parallel^2\right] \\ &= (\varepsilon_N)^2 \end{aligned}$$

by (A.1). Hence, $I_{1,k} = O_P(\varepsilon_N)$. To bound $I_{2,k}$, define the following function

$$f_k(r) = E_P\left[m_1^{(1)}(Z, \theta_1^*, \eta_{10} + r(\hat{\eta}_{1k} - \eta_{10})) \mid (W_i)_{i \in I_k^c}\right] - E\left[m_1^{(1)}(Z, \theta_1^*, \eta_{10})\right]$$

for $r \in [0, 1)$. By Taylor series expansion, we have

$$f_k(1) = f_k(0) + f_k'(0) + f_k''(\tilde{r})/2, \text{for some } \tilde{r} \in (0, 1).$$

Note that $f_k(0) = E\left[m_1^{(1)}(Z, \theta_1^*, \eta_{10}) \mid (W_i)_{i \in I_k^c}\right] = E\left[m_1^{(1)}(Z, \theta_1^*, \eta_{10})\right] = O(h^2)$ by the calculation in Step 4 in the proof of Theorem 2.2. Further, on the event $\hat{\eta}_{1k} \in \mathcal{T}_N$,

$$\parallel f_k'(0) \parallel = \parallel \partial_{\eta_1} E[m_1^{(1)}(Z, \theta_1^*, \eta_{10})] [\hat{\eta}_{1k} - \eta_{10}] \parallel = 0$$

by the orthogonality. Also, on the event $\hat{\eta}_{1k} \in \mathcal{T}_N$,

$$\parallel f_k''(\tilde{r}) \parallel \leq \sup_{r \in (0,1)} \parallel f_k''(r) \parallel \leq (\varepsilon_N)^2$$

by (A.2). Thus,

$$I_{2,k} = \sqrt{n} \parallel f_k\left(1\right) \parallel = O_P\left(\sqrt{n}\left(\varepsilon_N\right)^2 + \sqrt{n}h^2\right).$$

Together with the result on $I_{1,k}$, we have

$$\begin{aligned}\left\|R_{2,k}\right\| \leq & \frac{I_{1,k} + I_{2,k}}{\sqrt{n}} \\ = & O_P\left(n^{-1/2}\varepsilon_N + \left(\varepsilon_N\right)^2 + h^2\right)\end{aligned}$$

Hence,

$$\sqrt{Nh^3}\left\|R_{2,k}\right\| = O_P(\sqrt{h^3}\epsilon_N + \sqrt{Nh^3}\epsilon_N^2 + \sqrt{Nh^3}h^2) = o_P(1)$$

by the assumptions on the rate of convergence that $\epsilon_N = o((Nh^3)^{-1/4})$ and $Nh^7 \to 0$. Therefore,

$$\mathbb{E}_{n,k}[m_1^{(1)}(Z, \theta_1^*, \hat{\eta}_{1k})] = \frac{1}{n}\sum_{i \in I_k} m_1^{(1)}(Z, \theta_1^*, \eta_{10}) + o_P(1) \xrightarrow{d} N(0, V_1).$$

Chapter 3

Double/debiased Machine Learning for Quantile Treatment Effect

3.1 Introduction

In program evaluation, economists have long been interested in the effect of a treatment or policy intervention on outcomes beyond simple average. Examples include but are not limited to the distributional effects on job training programs (LaLonde, 1995), unions (Card, 1996), and minimum wages (DiNardo et al., 1995). One of the most popular way to capture the distributional effect of a treatment is to compute the difference of the quantiles of the outcome distribution before and after the treatment, which is called quantile treatment effect.

Like many studies in program evaluation, the main challenge in the estimation of quantile treatment effect is that the selection for treatment usually affects individual's potential outcomes. In order to address the endogenous effect of selection, researchers usually estimate the treatment effect conditioning on a vector of control variables. Firpo (2007) proposed a two-step estimator to include control variables nonparametrically in the estimation of the quantile treatment effect. In the paper, he proposed to estimate the first-step infinite-dimensional nuisance parameter using the series estimator and plug it into a check function to obtain a final estimator for the quantile treatment effect. Firpo (2007) also derived in detail the asymptotic properties and showed the efficiency of the proposed semiparametric estimator.

This paper discusses an orthogonal extension of the semiparametric estimator proposed in Firpo (2007). The proposed series estimator in the first-step estimation works well when the sample size is large compared to the number of control variables. In practice, researcher may want to include many potential control variables in order to exclude potential endogeneity. The number of control variables can be comparable to or even larger than the sample size. In this situation, the series estimator in the first-step estimation would suffer from the curse of dimension. Researcher may replace series estimator with machine learning (ML) methods such as Lasso, random forests, neural nets, and etc, in the first-step estimation. As noted in Chernozhukov et al. (2018), however, the asymptotic properties derived in Firpo (2007) may fail if researchers use ML methods in the first-step estimation since the regularization bias embedded in ML methods would lead to the bias of the final estimator.

To address this problem, we discuss in detail the estimation and inference of the double machine learning (DML) estimator for quantile treatment effect. The key is to replace the check function in the second step in Firpo (2007) with a newly derived score function. The new score function enjoys the Neyman-orthogonal property (Chernozhukov et al., 2018), which means that the first-order derivative of the score function with respect to the nuisance parameter is zero. With the property, the regularization bias within ML methods would only have second or higher order effects on the final estimator. The final estimator obtained based on this Neyman-orthogonal score can achieve $\sqrt{N}$-consistency and asymptotic normality as long as the first-step ML estimator converges to its true value with a rate faster than $N^{-1/4}$, which is a rate can be achieved by many ML estimators such as Lasso, random forests, neural nets, and etc.

The result in this paper relies heavily on the recent high-dimensional and ML literature: Belloni et al. (2012), Belloni et al. (2014), Chernozhukov et al. (2015), Belloni et al. (2017), and Chernozhukov et al. (2018). In Belloni et al. (2017), they provided a very general framework to derive the Neyman-orthogonal score for many treatment effect estimations, including quantile treatment effect. This paper complements their paper by presenting in detail the functional form of the Neyman-orthogonal score and the steps of estimation procedure.

Plan of the paper. Section 2 sets up the notation and framework for the discussion of the mode treatment effect. Section 3 derives the Neyman-orthogonal score for quantile treatment effect and I combine it with the cross-fitting algorithm to propose the DML estimaor and derive its asymptotic properties. Section 5 concludes this paper.

3.2 Notation and Framework

Let Y be a continuous outcome variable of interest, D the binary treatment indicator, and X a $d \times 1$ vector of control variables. Denote by $Y(1)$ an individual's potential outcome when

$D = 1$ and $Y(0)$ if $D = 0$. Then we have $Y = Y(1)D + Y(0)(1 - D)$. Also, we define the propensity score $m_0(X) := E[D|X]$ and for $\tau \in (0, 1)$ and $j \in \{0, 1\}$, the τ-th quantile for potential outcome $Y(j)$ is defined as $q_{j,\tau} := \inf\{q : Pr(Y(j) \leq q) \geq \tau\}$. We are interested in the quantile treatment effect

$$\theta_{0,\tau} := q_{1,\tau} - q_{0,\tau}$$

Assumption 3.1. *(strong ignorability)*

- $(Y(0), Y(1)) \perp D \mid X$
- $0 < m_0(X) < 1$

The first part of Assumption 3.1 assumes that potential outcomes are independent of treatment after conditioning on the observable covariates X. The second part states that for all values of X, both treatment status occur with a positive probability. With the strong ignorability, Firpo (2007) identified $q_{1,\tau}$ and $q_{0,\tau}$ by the following moment conditions:

$$E[\frac{T}{m_0(X)}\Big(\mathbb{1}\{Y \leq q_{1,\tau}\} - \tau\Big)] = 0, \tag{3.1}$$

$$E[\frac{1-T}{1-m_0(X)}\Big(\mathbb{1}\{Y \leq q_{0,\tau}\} - \tau\Big)] = 0. \tag{3.2}$$

In the moment condition 3.1 and 3.2, there is only one unknown nuisance parameter $m_0(X)$. An direct way to apply 3.2 and 3.2 to estimate $q_{1,\tau}$ and $q_{0,\tau}$ is to estimate $m_0(X)$ in the first-step and then use the estimator of $m_0(X)$ with the moment condition 3.1 and 3.2 to obtain the final estimator of $q_{1,\tau}$ and $q_{0,\tau}$. Specifically,

Definition (Direct Estimator).

(i) Obtain an nonparametric estimator of $m_0(X)$, denoted by $\hat{m}(X)$.

(ii) The estimator of $q_{1,\tau}$ and $q_{0,\tau}$ is $\hat{q}_{1,\tau}$ and $\hat{q}_{0,\tau}$ where $\hat{q}_{1,\tau}$ and $\hat{q}_{0,\tau}$ satisfies

$$\frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\hat{m}(X_i)}(\mathbb{1}\{Y_i \leq \hat{q}_{1,\tau}\} - \tau) = 0,$$

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1-T_i}{1-\hat{m}(X_i)}(\mathbb{1}\{Y_i \leq \hat{q}_{0,\tau}\} - \tau) = 0.$$

If $\hat{m}(X)$ is a kernel or a series estimator, by the classical semiparametric estimation results, we can show that both $\hat{q}_{1,\tau}$ and $\hat{q}_{0,\tau}$ are $\sqrt{N}$-consistent and asymptotically normal.

In many cases, however, the dimension of X may be large and researchers do not have enough observations to obtain an accurate kernel or series estimator because of the curse of dimension. In this situation, researchers may turn to use ML methods such as Lasso, random forests, boosting, neural nets, and etc, to build $\hat{m}(X)$ and then plug it into the Step 2 above. Unfortunately, as noted in Chernozhukov et al. (2018), the estimator of $q_{1,\tau}$ and $q_{0,\tau}$ obtained in this manner may be biased because the regularization bias embedded in the ML estimator $\hat{m}(X)$ would result in the bias of $\hat{q}_{1,\tau}$ and $\hat{q}_{0,\tau}$. In the next section, we derive the Neyman-orthogonal score of 3.1 and 3.2 and combine them with the cross-fitting algorithm (Chernozhukov et al., 2018) to propose the DML estimator of the quantile treatment effect.

3.3 Estimation

Based on the moment condition 3.1 and 3.2, we find the corresponding Neyman-orthogonal scores:

$$\Psi_1(w;q,\eta_1) = \frac{t}{m_0(x)}\Big(\mathbb{1}\{y \leq q\} - \tau\Big) - \frac{t - m_0(x)}{m_0(x)_0}g_1(x), \tag{3.3}$$

$$\Psi_2(w;q,\eta_2) = \frac{1-t}{1-m_0(x)}\Big(\mathbb{1}\{y \leq q\} - \tau\Big) - \frac{m_0(x) - t}{1-m_0(x)}g_2(x), \tag{3.4}$$

where $W = (Y, D, X)$, $\eta_1 = (m_0, g_1)$, $\eta_2 = (m_0, g_2)$, and

$$g_1(x) = E[\mathbb{1}\{Y \leq q_{1,\tau}\} - \tau | T = 1, X = x]$$

and

$$g_2(x) = E[\mathbb{1}\{Y \leq q_{0,\tau}\} - \tau | T = 0, X = x].$$

Notice that 3.3 and 3.4 are valid scores since

$$E[\Psi_1(W; q_{1,\tau}, \eta_1)] = 0,$$

$$E[\Psi_2(W; q_{0,\tau}, \eta_2)] = 0$$

by the law of iterated expectation. The key property of 3.3 and 3.4 is that the first-order Gateaux derivatives of Ψ_1 and Ψ_2 with respect to the nuisance parameters η_1 and η_2, respectively, are zero. This is the Neyman orthogonaliy proposed in Chernozhukov et al. (2018). In contrast, the first-order derivative of 3.3 and 3.4 with respect to their nuisance parameter m_0 does not equal to zero, and hence, 3.3 and 3.4 do not satisfy the Neyman-orthogonal property. The following lemma establishes the Neyman orthogonality of 3.3 and 3.4.

Lemma 3.1. *The scores 3.3 and 3.4 are Neyman-orthogonal.*

The quantile treatment effect can be estimated in three steps.

- STEP 1. Non-parametric estimation (machine learning) of the nuisance parameter η_j. Here we adopt cross-fitting.
- STEP 2. With the first step estimates, estimate $q_{j,\tau}$ using the orthogonal scores. The resulting plug-in type estimators $\hat{q}_{j,\tau}$ are called the double machine learning (DML) estimators.
- STEP 3. Take difference and get $\hat{\theta} = \hat{q}_{1,\tau} - \hat{q}_{0,\tau}$.

Below we provide a formal definition of the DML estimator which reflects the first two steps described above. Suppose we have i.i.d. observations $\{W_i\}_{i=1}^N$. Notice that in the estimation of η, we have to estimate g_1 where $g_1(x) = E[\mathbb{1}\{Y \leq q_{1,\tau}\} - \tau | T = 1, X = x]$ and there is an unknown paremeter $q_{1,\tau}$, which happens to be our target parameter. Hence, we need a preliminary estimator for $q_{j,\tau}$.

Definition. *(The preliminary estimator for $q_{j,\tau}$)*

- *Step 1. Nonparametric or ML estimation of the propensity score m_0.*
- *Step 2. The preliminary estimator $\hat{q}_{1,\tau}$ and $\hat{q}_{0,\tau}$ solve*

$$\frac{1}{N}\sum_{i=1}^N \frac{T_i}{\hat{m}(X_i)}(\mathbb{1}\{Y_i \leq \hat{q}_{1,\tau}\} - \tau) = 0,$$

$$\frac{1}{N}\sum_{i=1}^N \frac{1-T_i}{1-\hat{m}(X_i)}(\mathbb{1}\{Y_i \leq \hat{q}_{0,\tau}\} - \tau) = 0.$$

Definition (DML estimator for quantiles of potential outcome). *(a) Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $[N] = \{1, ..., N\}$ such that the size of each fold I_k is $n = N/K$. Also, for each $k \in [K] = \{1, ..., K\}$, define $I_k^c := \{1, ..., N\} \setminus I_k$. (b) For each $k \in [K]$, use the preliminary estimator in Definition 1 to construct Machine Learning (ML) estimators*

$$\hat{\eta}_{j,k} = \hat{\eta}_{j,k}((W_i)_{i \in I_k^c})$$

of η_j, $j \in \{0, 1\}$, where $\hat{\eta}_j$ is a random element in T, and where randomness depends only on the subset of data indexed by I_k^c. (c) For each $k \in [K]$, construct the estimators $\check{q}_{j,k}$,$j \in \{0, 1\}$, as the solution of the following equation:

$$\mathbb{E}_{n,k}[\Psi_j(W; \check{q}_{j,k}, \hat{\eta}_{j,k})] = 0$$

where Ψ is the Neyman orthogonal score, and $\mathbb{E}_{n,k}$ is the empirical expectation over the kth *fold of the data; (4) Aggregate the estimators:*

$$\hat{q}_j = \frac{1}{K}\sum_{k=1}^{K} \check{q}_{j,k}$$

3.3.1 Asymptotic Properties

This section studies the property of the estimator under high-level assumptions on the first step ML estimator.

Assumption 3.2 (regularity conditions)**.** *(a) The true parameter value obeys 3.1 and 3.2, and is contained in a ball in Θ. (b) map $(\theta, \eta) \mapsto E_P[\Psi(W; \theta, \eta)]$ is twice continuously Gateaux-differentable on $\Theta \times T$. (c) The marginal density function $f_{Y(1)}(q)$ does not equal to zero over its support. (d) $2|F_{Y(1)}(q) - \tau| \geq |f_{Y(1)}(q_{1,\tau})(q - q_{1,\tau})| \wedge c_0$.*

Theorem 3.1. *[Asymptotic normality of the DML estimator] Suppose Assumption 1 holds and with probability $1 - o(1)$, $\| \hat{\eta}_{1k} - \eta_{10} \|_{P,2} \leq \varepsilon_N$, $\| \hat{m}_k - 1/2 \|_{P,\infty} \leq 1/2 - \kappa$, and $\| \hat{m}_k - m_0 \|_{P,2}^2 + \| \hat{m}_k - m_0 \|_{P,2} \times \| \hat{g}_{j1k} - g_{j0} \|_{P,2} \leq (\varepsilon_N)^2$. We have*

$$\sqrt{N}(\hat{q}_j - q_j) \xrightarrow{d} N(0, f_{Y(j)}^{-1}(q) V_j f_{Y(j)}^{-1}(q))$$

where $f_{Y(j)}(q)$ is pdf of $Y(j)$ and $V_j = E[\Psi_j(W; q_{j,\tau}, \eta_j)^2]$.

Theorem 3.2. *Construct the estimator of the asymptotic variance as*

$$\hat{V}_j = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{n,k}[\Psi_j(W; \tilde{q}_{j,\tau}, \hat{\eta}_j)^2]$$

where where $\mathbb{E}_{n,k}\left[f\left(W\right)\right] = n^{-1}\sum_{i \in I_k} f\left(W_i\right)$. If the assumptions of Theorem 1 hold, $\hat{V}_j = V_j + o_P\left(1\right)$

Theorem 3.1 shows that DML $\tilde{q}_j$ can achieve $\sqrt{N}$-consistency and asymptotic normality if the first-step estimators of the infinite dimensional nuisance parameters converge at a rate faster than $N^{-1/4}$. This rate of convergence can be achieved by many ML methods. Theorem

3.2 provides consistent estimators for the asymptotic variance of $\tilde{q}_j$. The proofs of Theorem 3.1 and Theorem 3.2 can be found in the appendix.

3.4 Conclusion

This paper studies the estimation and inference of the orthogonal extension of the semiparametric estimation proposed in Firpo (2007) and propose a DML quantile treatment effect estimator. The proposed estimator can achieve $\sqrt{N}$-consistency and asymptotic normality when researchers apply ML methods in the first-step estimation. It also provides flexibility for empirical researchers to explore a broader set of popular estimation methods and analyze more types of data sets.

3.A Appendix

Proof of 3.1

with respect to g, want to show

$$\partial_r E[\Psi_j(W; q_{j,\tau}, m_0, g_j + r(g - g_j))]\Big|_{r=0} = 0, \text{ for all } g \in G \tag{3.5}$$

for $j = 1$,

$$\begin{aligned} &\Psi_1(W; q_{1,\tau}, m_0, g_1 + r(g - g_1)) && (3.6)\\ =&\frac{T}{m_0(X)}\Big(\mathbb{1}\{Y \leq q\} - \tau\Big) - \frac{T - m_0(X)}{m_0(X)}[g_1(X) + r(g(X) - g_1(X))] && (3.7)\end{aligned}$$

$$\partial_r E[\Psi_1(W; q_{1,\tau}, m_0, g_1 + r(g - g_1))] \tag{3.8}$$

$$= E\Big[- \frac{T - m_0(x)}{m_0(x)} [g(X) - g_1(X)]\Big] \tag{3.9}$$

$$= E\Big[- E[T - m_0(X) \mid X] \frac{1}{m_0(X)} [g(X) - g_1(X)]\Big] \tag{3.10}$$

$$= 0 \tag{3.11}$$

with respect to m, want to show

$$\partial_r E[\Psi_j(W; q_{j,\tau}, m_0 + r(m - m_0), g_j)]\Big|_{r=0} = 0, \text{ for all } g \in G \tag{3.12}$$

for $j = 1$,

$$\Psi_1(W; q_{1,\tau}, m_0 + r(m - m_0), g_1) \tag{3.13}$$

$$= \frac{T}{m_0(X) + r[m(X) - m_0(X)]} \Big(\mathbb{1}\{Y \leq q\} - \tau \Big) - \frac{T - m_0(X) - r[m(X) - m_0(X)]}{m_0(X) + r[m(X) - m_0(X)]} g_1(X) \tag{3.14}$$

$$\partial_r E[\Psi_1(W; q_{1,\tau}, m_0 + r(m - m_0), g_1)] \tag{3.15}$$

$$= E\Big[- \frac{T[m(X) - m_0(X)]}{(m_0(X) + r[m(X) - m_0(X)])^2} \Big(\mathbb{1}\{Y \leq q\} - \tau \Big) \tag{3.16}$$

$$+ \frac{T[m(X) - m_0(X)]}{(m_0(X) + r[m(X) - m_0(X)])^2} g_1(X) \Big] \tag{3.17}$$

recall that

$$g_1(X) = E[\mathbb{1}\{Y \leq q_{1,\tau}\} - \tau | T = 1, X = x] \tag{3.18}$$

Hence the result follows from law of iterated expectation and law of total probability.

proof of Theorem 3.1

Assumption 3.2 should lead to the conditions in Assumption 3.3 and 3.4 in CCDDHNR.

Conditions of Assumption 3.3.

- (a) *The true parameter value obeys (2.1), and* Θ *contains a ball ....*

  Neyman orthogonal scores are indeed score. Parameter space set to $\mathbb{R}$.

- (b) *map* $(\theta, \eta) \mapsto E_P[\Psi(W; \theta, \eta)]$ *is twice continuously Gateaux-differentable on* $\Theta \times T$.

  Directly assume this to our score functions. However, there's an indicator, so this could be tricky?

- (c) *identification relation*

$$J_0 := \partial_\theta E_P[\Psi(W; \theta, \eta_0)]\Big|_{\theta=\theta_0}$$

in our setting,

$$
\begin{aligned}
&\partial_q E[\Psi_1(W; q, m_0, g_1)] \\
=&\partial_q E[\frac{T}{m_0(X)}\Big(\mathbb{1}\{Y \leq q\} - \tau\Big) - \frac{T - m_0(X)}{m_0(X)} g_j(X)] \\
=&\partial_q E[\frac{T}{m_0(X)}\Big(\mathbb{1}\{Y \leq q\} - \tau\Big)] \\
=&\partial_q E[E[\frac{T}{m_0(X)}\Big(\mathbb{1}\{Y \leq q\} - \tau\Big) \mid X]] \\
=&\partial_q E[E[T\Big(\mathbb{1}\{Y \leq q\} - \tau\Big) \mid X]\frac{1}{m_0(X)}] \\
=&\partial_q E[E[\mathbb{1}\{Y(1) \leq q\} - \tau \mid X, T = 1]] \\
=&\partial_q E[E[\mathbb{1}\{Y(1) \leq q\} - \tau \mid X]] \\
=&\partial_q E[\mathbb{1}\{Y(1) \leq q\} - \tau] \\
=&\partial_q E[\mathbb{1}\{Y(1) \leq q\}] \\
=&\partial_q F_{Y(1)}(q) \\
=&f_{Y(1)}(q)
\end{aligned}
$$

Where $F_{Y(1)}(q)$ and $f_{Y(1)}(q)$ are cdf and pdf of $Y(1)$ **Assume potential outcomes are continuously distributed.** Now we need

$$2\|E[\Psi_1(W; q, m_0, g_1)]\| \geq \|f_{Y(1)}(q_{1,\tau})(q - q_{1,\tau})\| \wedge c_0$$

$$
\begin{aligned}
E[\Psi_1(W; q, m_0, g_1)] =&E[\frac{T}{m_0(X)}\Big(\mathbb{1}\{Y \leq q\} - \tau\Big) - \frac{T - m_0(X)}{m_0(X)} g_1(X)] \\
=&E[\frac{T}{m_0(X)}\Big(\mathbb{1}\{Y \leq q\} - \tau\Big)] \\
=&F_{Y(1)}(q) - \tau
\end{aligned}
$$

Hence we are assuming

$$2|F_{Y(1)}(q)-\tau| \geq |f_{Y(1)}(q_{1,\tau})(q-q_{1,\tau})| \wedge c_0$$

- (d) *Neyman-orthogonal*

  see above.

Conditions of Assumption 3.4.

- (a) *Exists $\mathcal{T}_n$, and $\Delta_n \to 0$, so that $\hat{\eta}_0$ falls into it with probability* $1-\Delta_n$

  Assume our first step estimators $\hat{\eta}_1$ satisfies the following conditions with probability $1-\Delta_n$ (MIGHT NOT NEED ALL OF THEM)

$$\begin{aligned} \|\hat{\eta}_1-\eta_1\|_{P,2} \leq& \delta'_N \\ \|\hat{m}_0-\frac{1}{2}\|_{P,\infty} \leq& \frac{1}{2}-\varepsilon \\ \|\hat{m}_0-m_0\|_{P,2} \times \|\hat{g}_0-g_1\|_{P,2} \leq& \delta'_N N^{-1/2} \end{aligned}$$

  Now define the set $\mathcal{T}_n$ as the set of η that satisfies the following conditions

$$\begin{aligned} \|\eta-\eta_1\|_{P,2} \leq& \delta'_N \\ \|m-\frac{1}{2}\|_{P,\infty} \leq& \frac{1}{2}-\varepsilon \\ \|m-m_0\|_{P,2} \times \|g-g_1\|_{P,2} \leq& \delta'_N N^{-1/2} \end{aligned}$$

  Then by the assumption and the definition of $\mathcal{T}_n$, this condition is satisfied.

- (b) ***EMPIRICAL PROCESS ASSUMPTION***

- (c) *There exists positive sequences* $\{\delta_N\}_{N\geq 1}$, $\{\nu_N\}_{N\geq 1}$ *that converge to zero and*

$$r_N := \sup_{\eta\in\mathcal{T}_n,\theta\in\Theta} \|E[\Psi(W;\theta,\eta)] - E[\Psi(W;\theta,\eta_0)]\| \leq \delta_N\nu_N$$

$$r'_N := \sup_{\eta\in\mathcal{T}_n,\|\theta-\theta_0\|\leq\nu_N} (E[\|\Psi(W;\theta,\eta) - \Psi(W;\theta_0,\eta_0)\|^2])^{1/2} \textit{ and } r'_N \log^{1/2}(1/r'_N) \leq \delta_N$$

$$\lambda'_N := \sup_{r\in(0,1),\eta\in\mathcal{T}_n,\|\theta-\theta_0\|\leq\nu_N} \|\partial_r^2 E[\Psi(W;\theta_0 + r(\theta-\theta_0),\eta_0 + r(\eta-\eta_0))]\| \leq \delta_N N^{-1/2}$$

First, $r_N \lesssim \delta'_N$, since for any q and $\eta\in\mathcal{T}_n$,

$$\begin{aligned}
&E[\Psi_1(W;q,\eta)] - E[\Psi_1(W;q,\eta_0)]\\
=&E\Big[\frac{T}{m(X)}\Big(\mathbb{1}\{Y\leq q\} - \tau\Big) - \frac{T-m(X)}{m(X)}g(X) - \frac{T}{m_0(X)}\Big(\mathbb{1}\{Y\leq q\} - \tau\Big)\\
&+ \frac{T-m_0(X)}{m_0(X)}g_1(X)\Big]\\
\leq&\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3\\
\lesssim&\delta'_N
\end{aligned}$$

where

$$
\begin{aligned}
\mathcal{I}_1 =& E[|\Big(\mathbb{1}\{Y \leq q\} - \tau\Big)\frac{T(m_0 - m)}{m_0 m}|] \\
\leq& \frac{1}{\varepsilon^2} E[|m_0 - m|] \\
\leq& \frac{1}{\varepsilon^2} \|m_0 - m\|_{P,2} \\
\lesssim& \delta'_N \\
\mathcal{I}_2 =& E[|g_1 - g|] \\
\leq& \|g_1 - g\|_{P,2} \\
\lesssim& \delta'_N \\
\mathcal{I}_3 =& E[|\frac{T}{m_0} g_1 - \frac{T}{m} g|] \\
=& E[|T \frac{g_1 m - m_0 g}{m m_0}|] \\
\leq& \frac{1}{\varepsilon^2} E[|m_0(g_1 - g) + g_1(m - m_0)|] \\
\leq& \frac{1}{\varepsilon^2} E[|m_0(g_1 - g)|] + E[|g_1(m - m_0)|] \\
\leq& \frac{1}{\varepsilon^2} E[|g_1 - g|] + E[|m - m_0|] \\
\leq& \frac{1}{\varepsilon^2} \|g_1 - g\|_{P,2} + \|m - m_0\|_{P,2} \\
\lesssim& \delta'_N
\end{aligned}
$$

Where we used that $|m_0(X)| < 1$ and $|g_1(X)| \leq 1$ almost surely because $m_0(X) = E[T|X]$ and $g_1(X) = E[\mathbb{1}\{Y \leq q_1\} - \tau | T = 1, X]$.

Next we show that $r'_N \lesssim \delta'_N + \nu_N^{1/2}$. For all $q \in \{q : |q - q_{1,\tau}| \leq \nu_N\}$, $\eta \in \mathcal{T}_N$

$$
\begin{aligned}
&\|\Psi_1(W; q, \eta) - \Psi_1(W; q_1, \eta_1)\|_{P,2} \\
=&\|\frac{T}{m(X)}\Big(\mathbb{1}\{Y \leq q\} - \tau\Big) - \frac{T - m(X)}{m(X)}g(X) - \frac{T}{m_0(X)}\Big(\mathbb{1}\{Y \leq q_{1,\tau}\} - \tau\Big) \\
&+ \frac{T - m_0(X)}{m_0(X)}g_1(X)\|_{P,2} \\
\leq&\mathcal{I}_4 + \mathcal{I}_5 + \mathcal{I}_6 + \mathcal{I}_7
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{I}_4 =& \| \frac{T}{m} \mathbb{1}\{Y \leq q\} - \frac{T}{m_0} \mathbb{1}\{Y \leq q_{1,\tau}\} \|_{P,2} \\
=& \| T \frac{\mathbb{1}\{Y \leq q\} m_0 - \mathbb{1}\{Y \leq q_{1,\tau}\} m}{m_0 m} \|_{P,2} \\
\leq& \frac{1}{\varepsilon^2} \| \mathbb{1}\{Y \leq q\} m_0 - \mathbb{1}\{Y \leq q_{1,\tau}\} m \|_{P,2} \\
=& \frac{1}{\varepsilon^2} \| \mathbb{1}\{Y \leq q_{1,\tau}\}(m_0 - m) + m_0(\mathbb{1}\{Y \leq q\} - \mathbb{1}\{Y \leq q_{1,\tau}\}) \|_{P,2} \\
\lesssim& \| \mathbb{1}\{Y \leq q_{1,\tau}\}(m_0 - m) \|_{P,2} + \| m_0(\mathbb{1}\{Y \leq q\} - \mathbb{1}\{Y \leq q_{1,\tau}\}) \|_{P,2} \\
\leq& \| m - m_0 \|_{P,2} + \| \mathbb{1}\{Y \leq q\} - \mathbb{1}\{Y \leq q_{1,\tau}\} \|_{P,2} \\
=& \| m - m_0 \|_{P,2} + \Big[Pr(q \wedge q_{1,\tau} \leq Y \leq q \vee q_{1,\tau}) \Big]^{1/2} \\
\leq& \delta'_N + [\nu_N \sup f_Y(y)]^{1/2} \\
\lesssim& \delta'_N + \nu_N^{1/2} \\
\mathcal{I}_3 =& \| \tau(\frac{T}{m_0} - \frac{T}{m}) \|_{P,2} \\
=& \| \tau T \frac{m - m_0}{m m_0} \|_{P,2} \\
\leq& \frac{1}{\varepsilon^2} \| m - m_0 \|_{P,2} \\
\lesssim& \delta'_N \\
\mathcal{I}_4 =& \| g - g_1 \|_{P,2} \\
\lesssim& \delta'_N \\
\mathcal{I}_5 =& \| \frac{T}{m_0} g_1 - \frac{T}{m} g \|_{P,2} \\
=& \| T \frac{g_1 m - g m_0}{m_0 m} \|_{P,2} \\
\leq& \frac{1}{\varepsilon^2} \| g_1(m - m_0) + m_0(g_1 - g) \|_{P,2} \\
\lesssim& \| g_1(m - m_0) \|_{P,2} + \| m_0(g_1 - g) \|_{P,2} \\
\lesssim& \delta'_N
\end{aligned}
$$

At last we show that

$$\begin{aligned}
f(r) :=& E\Big[\frac{T}{m_0 + r(m-m_0)}\mathbb{1}\{Y \leq q_1 + r(q-q_1)\} \\
&+ \frac{T - m_0 - r(m-m_0)}{m_0 + r(m-m_0)}(g_1 + r(g-g_1))\Big] \\
=& g(r) + h(r)
\end{aligned}$$

where

$$\begin{aligned}
g(r) :=& E\Big[\frac{T}{m_0 + r(m-m_0)}\mathbb{1}\{Y \leq q_1 + r(q-q_1)\}\Big] \\
h(r) :=& E\Big[\frac{T - m_0 - r(m-m_0)}{m_0 + r(m-m_0)}(g_1 + r(g-g_1))\Big]
\end{aligned}$$

$$\begin{aligned}
g(r) =& E\Big[\frac{m_0}{m_0 + r(m-m_0)}\frac{1}{m_0}E[T\mathbb{1}\{Y \leq q_1 + r(q-q_1)\} \mid X]\Big] \\
=& E\Big[\frac{m_0}{m_0 + r(m-m_0)}E[\mathbb{1}\{Y(1) \leq q_1 + r(q-q_1)\} \mid X]\Big] \\
=& E\Big[\frac{m_0}{m_0 + r(m-m_0)}F_{Y(1)|X}(q_1 + r(q-q_1))\Big] \\
\partial_r g(r) =& E\Big[\frac{-m_0(m-m_0)}{(m_0 + r(m-m_0))^2}F_{Y(1)|X}(q_1 + r(q-q_1)) \\
&+ \frac{m_0}{m_0 + r(m-m_0)}f_{Y(1)|X}(q_1 + r(q-q_1))(q-q_1)\Big] \\
\partial_r^2 g(x) :=& E\Big[\frac{2m_0(m-m_0)^2}{(m_0 + r(m-m_0))^3}F_{Y(1)|X}(q_1 + r(q-q_1))\Big] \\
&+ E\Big[\frac{-m_0(m-m_0)}{(m_0 + r(m-m_0))^2}f_{Y(1)|X}(q_1 + r(q-q_1))(q-q_1)\Big] \\
&+ E\Big[\frac{-m_0(m-m_0)}{(m_0 + r(m-m_0))^2}f_{Y(1)|X}(q_1 + r(q-q_1))(q-q_1)\Big] \\
&+ E\Big[\frac{m_0}{m_0 + r(m-m_0)}\dot{f}_{Y(1)|X}(q_1 + r(q-q_1))(q-q_1)^2\Big]
\end{aligned}$$

$$
\begin{aligned}
h(r) =& E\Big[\frac{T}{m_0 + r(m-m_0)}(g_1 + r(g-g_1)) - (g_1 + r(g-g_1))\Big] \\
\partial_r h(r) =& E\Big[\frac{-T(m-m_0)}{(m_0 + r(m-m_0))^2}(g_1 + r(g-g_1)) + \frac{T(g-g_1)}{m_0 + r(m-m_0)} - (g-g_1)\Big] \\
\partial_r^2 h(r) =& E\Big[\frac{2T(m-m_0)^2}{(m_0 + r(m-m_0))^3}(g_1 + r(g-g_1))\Big] \\
&+ E\Big[\frac{-T(m-m_0)(g-g_1)}{(m_0 + r(m-m_0))^2}\Big] \\
&+ E\Big[\frac{-T(g-g_1)(m-m_0)}{(m_0 + r(m-m_0))^2}\Big]
\end{aligned}
$$

Proof of Theorem 2

We assume the following auxiliary conditions:

$$
\sup_{\eta_1 \in \mathcal{T}_N} \left(E_P \left[\| \ \Psi_j(W; q_{j,\tau}, \eta) - \Psi_j(W; q_{j,\tau}, \eta_j) \ \|^2 \right] \right)^{1/2} \leq \varepsilon_N, \tag{A.12}
$$

$$
\left(E_P \left[\Psi_j(W; q_{j,\tau}, \eta_j)^4 \right] \right)^{1/4} \leq C_1, \tag{A.13}
$$

where $\mathcal{T}_N$ are specified in the proof of Theorem 1, C_1 is a constant. Since K is fixed, which is independent of N, it suffices to show that for each $k \in [k]$,

$$
I_k \equiv \left| \mathbb{E}_{n,k} \left[\Psi_j(W; \hat{q}_{j,\tau}, \hat{\eta}_j)^2 \right] - E_P \left[\Psi_j(W; q_{j,\tau}, \eta_j)^2 \right] \right| = o_P(1).
$$

By the triangle inequality, we have

$$
I_k \leq I_{3,k} + I_{4,k},
$$

where

$$
I_{3,k} \equiv \left| \mathbb{E}_{n,k} \left[\Psi_j(W; \hat{q}_{j,\tau}, \hat{\eta}_j)^2 \right] - \mathbb{E}_{n,k} \left[\Psi_j(W; q_{j,\tau}, \eta_j)^2 \right] \right|,
$$

$$I_{4,k} \equiv \left| \mathbb{E}_{n,k}\left[\Psi_j(W; q_{j,\tau}, \eta_j)^2\right] - E_P\left[\Psi_j(W; q_{j,\tau}, \eta_j)^2\right] \right|.$$

To bound $I_{4,k}$, we have

$$\begin{aligned} E_P\left[I_{4,k}^2\right] \leq & n^{-1} E_P\left[\Psi_j(W; q_{j,\tau}, \eta_j)^4\right] \\ \leq & n^{-1} C_1^4, \end{aligned}$$

where the last inequality follows from (A.13). Then we have $I_{4,k} = O_P\left(n^{1/2}\right)$.

Next, we bound $I_{3,k}$. This part is essentially identical to the proof of Theorem 3.2 in Chernozhukov et al. (2018), I reproduce it here for reader's convenience. Observe that for any number a and δa,

$$\mid (a+\delta a)^2 - a^2 \mid \leq 2(\delta a)(a + \delta a).$$

Denote $\psi_i = \Psi_j(W; q_{j,\tau}, \eta_j)$ and $\hat{\psi}_i = \Psi_j(W; \hat{q}_{j,\tau}, \hat{\eta}_j)$, and $a \equiv \psi_i$, $a + \delta a \equiv \hat{\psi}_i$. Then

$$\begin{aligned} I_{3,k} = & \mid \frac{1}{n} \sum_{i \in I_k} \left(\hat{\psi}_i\right)^2 - (\psi_i)^2 \mid \leq \frac{1}{n} \sum_{i \in I_k} \mid \left(\hat{\psi}_i\right)^2 - (\psi_i)^2 \mid \\ \leq & \frac{2}{n} \sum_{i \in I_k} \mid \hat{\psi}_i - \psi_i \mid \times \left(\mid \psi_i \mid + \mid \hat{\psi}_i - \psi_i \mid\right) \\ \leq & \left(\frac{2}{n} \sum_{i \in I_k} \mid \hat{\psi}_i - \psi_i \mid^2\right)^{1/2} \left(\frac{2}{n} \sum_{i \in I_k} \left(\mid \psi_i \mid + \mid \hat{\psi}_i - \psi_i \mid\right)^2\right)^{1/2} \\ \leq & \left(\frac{2}{n} \sum_{i \in I_k} \mid \hat{\psi}_i - \psi_i \mid^2\right)^{1/2} \left[\left(\frac{2}{n} \sum_{i \in I_k} \mid \psi_i \mid^2\right)^{1/2} + \left(\frac{2}{n} \sum_{i \in I_k} \mid \hat{\psi}_i - \psi_i \mid^2\right)^{1/2}\right]. \end{aligned}$$

Thus,

$$I_{3,k}^2 \lesssim S_N \times \left(\frac{1}{n} \sum_{i \in I_k} \parallel \Psi_j(W; q_{j,\tau}, \eta_j) \parallel^2 + S_N\right),$$

where

$$S_N \equiv \frac{1}{n} \sum_{i \in I_k} \parallel \Psi_j(W; \hat{q}_{j,\tau}, \hat{\eta}_j) - \Psi_j(W; q_{j,\tau}, \eta_j) \parallel^2.$$

Since $\frac{1}{n} \sum_{i \in I_k} \parallel \Psi_j(W; q_{j,\tau}, \eta_j) \parallel^2 = O_P(1)$, it suffices to bound S_N. Under the event that

$\hat{\eta}_{1k} \in \mathcal{T}_N$, we have

$$
\begin{aligned}
& E_P \left[\| \, \bar{\psi}_1 \left(W_i, \theta_0, \hat{p}_k, \hat{\eta}_{1k} \right) - \bar{\psi}_1 \left(W_i, \theta_0, p_0, \eta_{10} \right) \, \|^2 | \, \left(W_i \right)_{i \in I_k^c} \right] \\
\leq & \sup_{p \in \mathcal{P}_N, \eta_1 \in \mathcal{T}_N} E_P \left[\| \, \bar{\psi}_1 \left(W_i, \theta_0, p, \eta_1 \right) - \bar{\psi}_1 \left(W_i, \theta_0, p_0, \eta_{10} \right) \, \|^2 \right] = \left(\varepsilon_N \right)^2
\end{aligned}
$$

by (A.12). It follows that $S_N = O_P \left(N^{-1} + \left(\varepsilon_N \right)^2 \right)$. Therefore, we obtain

$$
I_k = O_P \left(N^{-1/2} \right) + O_P \left(N^{-1/2} + \varepsilon_N \right) = o_P \left(1 \right).
$$

Hence, $\hat{V}_1 \xrightarrow{p} V_{10}$.

References

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

Abadie, A., Angrist, J., & Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, *70*(1), 91–117.

Akee, R., Copeland, W., Costello, E. J., & Simeonova, E. (2018). How does household income affect child personality traits and behaviors? *American Economic Review*, *108*(3), 775–827.

Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of public economics*, *1*, 323–338.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429.

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

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls†. *The Review of Economic Studies*, *81*(2), 608-650.

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, *96*(4), 988–1012.

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *ILR Review*, *43*(2), 245–257.

Card, D. (1996). The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica: Journal of the Econometric Society*, 957–979.

Card, D., & Krueger, A. (1994). Minimum wages and employment: a case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, *84*(4), 772–793.

Chen, Y.-C., Genovese, C. R., Tibshirani, R. J., Wasserman, L., et al. (2016). Nonparametric modal regression. *The Annals of Statistics*, *44*(2), 489–514.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., & Newey, W. K. (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Chernozhukov, V., & Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica*, *73*(1), 245–261.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annu. Rev. Econ.*, *7*(1), 649–688.

Clotfelter, C. T. (1983). Tax evasion and tax rates: An analysis of individual returns. *The review of economics and statistics*, 363–373.

DiNardo, J., Fortin, N. M., & Lemieux, T. (1995). *Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach* (Tech. Rep.). National bureau of economic research.

Eddy, W. F., et al. (1980). Optimum kernel estimators of the mode. *The Annals of Statistics*, *8*(4), 870–882.

Feinstein, J. S. (1991). An econometric analysis of income tax evasion and its detection. *The RAND Journal of Economics*, 14–35.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, *75*(1), 259–276.

Freeman, R. B. (1980). Unionism and the dispersion of wages. *ILR Review*, *34*(1), 3–23.

Fuest, C., Peichl, A., & Siegloch, S. (2018). Do higher corporate taxes reduce wages? micro evidence from germany. *American Economic Review*, *108*(2), 393–418.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 315–331.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, *24*(3), 726–748.

Heckman, J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, *64*(4), 605–654.

Heckman, J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics*, *30*(1-2), 239–267.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189.

LaLonde, R. J. (1995). The promise of public sector-sponsored training programs. *Journal of Economic perspectives*, *9*(2), 149–168.

Lee, M.-J. (1989). Mode regression. *Journal of Econometrics*, *42*(3), 337–349.

Li, F. (2019). Double-robust estimation in difference-in-differences with an application to traffic safety evaluation. *arXiv preprint arXiv:1901.02152*.

Lu, C., Nie, X., & Wager, S. (2019). Robust nonparametric difference-in-differences estimation. *arXiv preprint arXiv:1905.11622*.

Meyer, B. D., Viscusi, W. K., & Durbin, D. L. (1995). Workers' compensation and injury duration: evidence from a natural experiment. *The American economic review*, 322–340.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 1349–1382.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, *33*(3), 1065–1076.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

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

Sant'Anna, P. H., & Zhao, J. B. (2019). Doubly robust difference-in-differences estimators. *Available at SSRN 3293315*.

Sequeira, S. (2016). Corruption, trade costs, and gains from tariff liberalization: evidence from southern africa. *American Economic Review*, *106*(10), 3029–63.

Sequeira, S., & Djankov, S. (2014). Corruption and firm behavior: Evidence from african ports. *Journal of International Economics*, *94*(2), 277–294.

Slemrod, J., & Yitzhaki, S. (2002). Tax avoidance, evasion, and administration. In *Handbook of public economics* (Vol. 3, pp. 1423–1470). Elsevier.

Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, *30*(1-2), 415–443.

Van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, *36*(2), 614–645.

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

Yao, W., & Li, L. (2014). A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, *41*(3), 656–671.

Zimmert, M. (2019). Efficient difference-in-differences estimation with high-dimensional common trend confounding. *arXiv preprint arXiv:1809.01643*.