Tests for Heterogeneous Treatment Effects

Fangzhou Yu*

March, 2025

Abstract

Recent advances in causal machine learning have facilitated reliable estimators of the average treatment effect (ATE) with valid statistical inference. However, applying similar techniques to conditional average treatment effects (CATE) poses significant inferential challenges. To bridge the gap between inference on the ATE and heterogeneity analysis, we propose three hypothesis tests to detect the existence of heterogeneous treatment effects. These tests inform researchers whether the treatment effect is constant across subpopulations defined by the covariates, thereby bridging the gap between inference on the ATE and the more ambitious task of fully characterizing those heterogeneities. Our tests build on three causal parameters: the projection of CATE on covariates, the variance of the CATE, and the variance difference between the potential outcomes. The test statistics are derived from the influence functions of the proposed parameters and are illustrated through Monte Carlo simulations and two empirical applications.

*Department of Economics, UNSW. fangzhou.yu@student.unsw.edu.au

1 Introduction

The analysis of causal effects has traditionally centered on the Average Treatment Effect (ATE), which measures the mean impact of a treatment or policy across the entire population. While nonparametric estimators for the ATE with valid statistical inference are well-established under the unconfoundedness assumption (e.g., Robins et al., 1994; Chernozhukov et al., 2018), the ATE often masks substantial heterogeneity in individual responses. Recognizing this heterogeneity is crucial for understanding the underlying mechanisms of a treatment and for designing optimal policies that tailor interventions to specific subpopulations (Heckman et al., 1997; Athey and Imbens, 2017).

Recent advances in causal machine learning have facilitated flexible estimation of the Conditional Average Treatment Effect (CATE) function, $\tau(x) = E[Y(1) - Y(0)|X = x]$, even in high-dimensional settings (e.g., Wager and Athey, 2018; Nie and Wager, 2021). However, achieving valid statistical inference on the CATE function remains challenging. The complexity of modern machine learning algorithms often precludes the use of classical empirical process theory (e.g., Donsker conditions), and the adaptive nature of these methods can introduce substantial biases when estimating the CATE function directly. Consequently, while these methods provide valuable point estimates, constructing reliable confidence intervals for the CATE or formally testing hypotheses about its structure remains difficult.

To bridge the gap between the robust inference available for the ATE and the more ambitious goal of characterizing heterogeneity, we propose two distinct hypothesis tests designed to detect the existence of heterogeneous treatment effects. These tests serve as a critical prerequisite in the empirical workflow: if the null hypothesis of homogeneity is rejected, researchers are justified in undertaking a more granular analysis of the CATE; if not, the ATE may suffice as a summary of the treatment's impact.

This paper develops these two complementary tests by leveraging the framework of Double/Debiased Machine Learning (DML) (Chernozhukov et al., 2018). DML utilizes Neyman-orthogonal influence functions and cross-fitting to provide $\sqrt{n}$-consistent and asymptotically normal estimators, even when the underlying nuisance functions (such as propensity scores and conditional outcome models) are estimated using flex-

ible machine learning methods that converge at slower rates (e.g., $o_P(n^{-1/4})$).

Our first contribution is the CATE Variance Test (CVT). This test examines the null hypothesis that the CATE is constant across all subpopulations defined by the covariates, $H_0 : Var(\tau(X)) = 0$. The variance of the CATE is a natural, omnibus measure of heterogeneity that summarizes potentially complex interactions into a scalar parameter. We derive the efficient influence function for the CATE variance, enabling robust inference under weak conditions. The CVT provides a necessary and sufficient test for whether observed covariates moderate the treatment effect.

Our second contribution is the Potential Outcome Variance Test (POVT). This test moves beyond effects moderated by covariates to consider the impact of the treatment on the distribution of outcomes. We examine the difference between the variances of the treated and untreated potential outcomes, $\dot{\lambda}_0 = Var(Y(1)) - Var(Y(0))$. This parameter captures the Treatment Effect on Variance, which is intrinsically valuable for understanding how interventions affect outcome dispersion or inequality. Furthermore, the POVT serves as a test for the existence of Individual Treatment Effect (ITE) heterogeneity. Since a constant ITE implies equal variances, rejecting $H_0 : \dot{\lambda}_0 = 0$ provides sufficient evidence that treatment effects vary at the individual level.

The POVT is the most informative test for ITE heterogeneity that relies solely on the identifiable marginal distributions of potential outcomes, avoiding untestable assumptions on their joint dependence (Heckman et al., 1997). However, it is a conservative test, as it cannot detect ITE heterogeneity if the variance of the treatment effect is exactly offset by a negative covariance between the effect and the baseline outcome. We address this limitation by introducing an ancillary parameter: the CATE-Baseline Covariance, $\kappa_0 = Cov(\tau(X), E[Y(0)|X])$. We develop a DML estimator for this identifiable component of the covariance, which helps diagnose whether the treatment is compensatory or amplifying, thereby aiding the interpretation of the POVT.

Our work contributes to the growing literature on testing for heterogeneous treatment effects. Unlike projection-based methods that test the significance of coefficients in a linear approximation of the CATE (e.g., Crump et al., 2008; Semenova and Chernozhukov, 2021), the CVT provides an omnibus test powerful against nonlinear alternatives. The POVT extends ideas previously explored in randomized experiments

(Ding et al., 2016) to observational settings by employing DML to adjust for confounding in high dimensions.

The remainder of the paper is organized as follows. Section 2 establishes the econometric framework and formally defines the concepts of heterogeneity. Section **??** reviews the related literature. Section 3.2 introduces the CATE Variance Test. Section 3.3 presents the Potential Outcome Variance Test, the ancillary CATE-Baseline Covariance parameter, and their joint interpretation. Section 4 details the Monte Carlo simulation results. Section 5 provides the empirical application, and Section 6 concludes. All proofs are collected in the Appendix.

In the literature of empirical treatment effect analysis, most of the papers focus on the estimation and inference of the average treatment effect (ATE) identified under the unconfoundedness. Nonparametric estimators with valid statistical inference have been developed to reduce reliance on model assumptions (e.g. Van Der Laan and Rubin, 2006; Chernozhukov et al., 2018). These practices evaluate the treatment effect by its average over the whole population but may overlook the heterogeneity in the treatment effect. Understanding heterogeneous treatment effects is crucial for two main objectives. First, it provides insights into the key drivers of the treatment effect and informs mechanism analysis. Second, it helps policy makers find the optimal treatment assignment rule by identifying whether an individual would be better off with or without the treatment.

To complement the established methods for ATE inference and advance our understanding of heterogeneous treatment effects, we propose three hypothesis tests for detecting the existence of heterogeneity. We focus on the null hypothesis that the conditional average treatment effect (CATE) is constant across subpopulations defined by the covariates, against the alternative hypothesis of its negation. These tests streamline the empirical workflow by helping researchers determine when more detailed CATE analysis would be beneficial.

Our hypothesis tests are built on three causal parameters that summarize the heterogeneity of the treatment effect: the projection of CATE on covariates, the variance of the CATE, and the variance difference between the potential outcomes. These parameters offer a significant advantage over the infinite-dimensional CATE function, as

their low-dimensional nature allows us to leverage recent advances in causal machine learning for statistical inference. In developing our approach, we appeal to the influence functions of these parameters to construct $\sqrt{n}$-consistent and asymptotically normal estimators. The resulting estimators are equivalent to those obtained from double/debiased machine learning methods (Chernozhukov et al., 2018). For inference, we provide a set of conditions that unify the classical Donsker condition on the complexity of nonparametric estimators with the modern approach of cross-fitting.

There has been an emerging literature on nonparametric tests for heterogeneous treatment effects, (e.g. Crump et al., 2008; Chang et al., 2015; Hsu, 2017; Sant'Anna, 2021; Dai et al., 2023), under different definitions of heterogeneity and hypotheses. Our approach distinguishes itself from prior work primarily through its integration of machine learning methods to address high-dimensionality and nonlinearity of the data.

Our first test, the CATE Projection Test (CPT), examines the joint significance of projection coefficients of the CATE on a set of covariates. The idea of direct examination of the CATE function is similar to Crump et al. (2008), who proposed a test based on series estimation of the regression functions of the treated and untreated potential outcomes. Their method was later generalized by Sant'Anna (2021) to test for heterogeneity in duration outcomes. The CPT also shares theoretical connections with machine learning estimation of the CATE function. CATE estimation is challenging because canonical machine learning methods are designed for outcome prediction by minimizing the mean squared error loss. The literature has developed two main strategies for CATE estimation. Athey and Imbens (2016) and Wager and Athey (2018) modify regression trees to optimize for the loss of CATE estimation. Another strand decomposes CATE estimation into a sequence of sub-regression problems and solves them using off-the-shelf machine learning methods (Zimmert and Lechner, 2019; Nie and Wager, 2021; Semenova and Chernozhukov, 2021; Fan et al., 2022). Our approach aligns with the latter strategy but diverges from these methods in its primary objective. Rather than pursuing accurate CATE function estimates by series or kernel estimators, we project the CATE onto a set of covariates to detect heterogeneity, which substantially simplifies the estimation and inference procedure.

Our second and third tests introduce variance-based parameters for detecting treat-

ment effect heterogeneity. The CATE Variance Test (CVT) examines the variance of the conditional effects, offering a novel measure of treatment effect variation across subpopulations. The Potential Outcome Variance Test (POVT) stems from the insight that the heterogeneity of the treatment effect manifests in the variance change of potential outcomes. This fundamental idea was previously explored by Ding et al. (2016), who proposed a randomization test based on the variance ratio of potential outcomes. These variance-based parameters have received limited attention in the literature, perhaps because they do not characterize the specific function of the CATE. However, they prove especially valuable for hypothesis testing. An additional advantage of these two parameters is that they are identified without strong assumptions on the joint distribution of potential outcomes (Heckman et al., 1997). We contribute to the literature by developing identification strategies and machine learning estimators for these causal variance parameters.

The rest of the paper is organized as follows. Section **??** reviews the identification of CATE and presents the hypotheses of interest. In Section 3, we develop three tests for heterogeneous treatment effects. For each test, we systematically present: the definition of the causal parameter, its identification, the derivation of the corresponding influence function, and the construction of the test statistic. Section 4 presents Monte Carlo simulation results of the performance of our tests. In Section 5, we apply our proposed tests to the survey data from the Chinese Family Panel Studies (CFPS) regarding the effect of being the only child on the mental health of only children, and the survey data of the 401(k) plan regarding the effect of 401(k) participation on net financial assets of the participants. Some concluding remarks are provided in Section 6. All proofs are collected in Appendix A.

2 Framework and Definitions of Heterogeneity

This section establishes the econometric framework, formally defines the distinct concepts of treatment effect heterogeneity, and outlines the identification strategies for the parameters of interest.

2.1 Setup and Notation

We operate within the potential outcomes framework (Rubin, 1974). Suppose we observe n independent and identically distributed (i.i.d.) realizations $O_i = (Y_i, D_i, X_i)$, $i = 1, ..., n$, drawn from an unknown distribution P_0. $D_i \in \{0, 1\}$ is a binary treatment indicator, and $X_i \in \mathcal{X}$ is a vector of pre-treatment covariates, which may be high-dimensional. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes under treatment and control, respectively. The observed outcome $Y_i \in \mathbb{R}$ is linked to the potential outcomes via the consistency equation:

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

The Individual Treatment Effect (ITE) is defined as $\tau_i = Y_i(1) - Y_i(0)$. As $Y_i(1)$ and $Y_i(0)$ are never simultaneously observed, τ_i is fundamentally unobservable. We therefore focus on population-level parameters. The Conditional Average Treatment Effect (CATE) is defined as:

$$\tau(x) = \mathbb{E}[\tau_i | X_i = x].$$

The Average Treatment Effect (ATE) is the expectation over the population: $\tau_0 = \mathbb{E}[\tau_i]$.

We denote the key nuisance functions required for identification as follows: the propensity score $e_0(x) = P(D_i = 1 | X_i = x)$, and the conditional expectation functions (CEFs) of the potential outcomes $\mu_0(d, x) = \mathbb{E}[Y_i(d) | X_i = x]$ for $d \in \{0, 1\}$.

2.2 Identification Assumptions

To identify the causal parameters from the observed data distribution P_0, we maintain the following standard assumptions, which allow for the identification of the marginal distributions of the potential outcomes.

Assumption 1 (Unconfoundedness)**.** $D_i \perp (Y_i(1), Y_i(0)) | X_i$.

Assumption 3 states that, conditional on the covariates X_i, the treatment assignment is independent of the potential outcomes. This requires that X_i includes all relevant confounders.

Assumption 2 (Overlap)**.** *There exists a constant $\xi > 0$ such that $\xi \leq e_0(x) \leq 1 - \xi$ almost surely over the support $\mathcal{X}$.*

Assumption 4 ensures that for any realization of the covariates, there is a positive probability of receiving either treatment or control. Under Assumptions 3 and 4, the CATE function is identified as $\tau(x) = \mu_0(1, x) - \mu_0(0, x)$.

2.3 Identification via Pseudo-Outcomes

A crucial construct for robust estimation and inference in the presence of high-dimensional covariates is the Augmented Inverse Probability Weighted (AIPW) pseudo-outcome (Robins et al., 1994). The AIPW pseudo-outcome for the ATE is defined as:

$$\psi(O_i) = \mu_0(1, X_i) - \mu_0(0, X_i) + \frac{D_i(Y_i - \mu_0(1, X_i))}{e_0(X_i)} - \frac{(1 - D_i)(Y_i - \mu_0(0, X_i))}{1 - e_0(X_i)}. \quad (1)$$

The key property of the AIPW pseudo-outcome is that its conditional expectation identifies the CATE:

$$\mathbb{E}[\psi(O_i)|X_i = x] = \tau(x). \quad (2)$$

This identity implies that $\psi(O_i)$ can be viewed as an unbiased, albeit noisy, observation of $\tau(x)$, which forms the basis for the tests developed in this paper. Furthermore, $\psi(O_i)$ corresponds to the efficient influence function for the ATE, possessing desirable robustness properties.

2.4 Defining Heterogeneity

We distinguish between two fundamental types of treatment effect heterogeneity.

Definition 1 (ITE Homogeneity)**.** *The individual treatment effect is homogeneous if $\tau_i = \tau_0$ almost surely.*

ITE homogeneity is a strong condition implying that every individual responds identically to the treatment. The corresponding null hypothesis is $H_0^{ITE} : Var(\tau_i) = 0$.

Definition 2 (CATE Homogeneity)**.** *The conditional average treatment effect is homogeneous if* $\tau(x) = \tau_0$ *almost surely over* $\mathcal{X}$*.*

CATE homogeneity implies that the observed covariates X_i do not moderate the average treatment effect. The corresponding null hypothesis is $H_0^{CATE} : Var(\tau(X_i)) = 0$.

The relationship between these two concepts can be formalized using the Law of Total Variance:

$$Var(\tau_i) = \underbrace{Var(\mathbb{E}[\tau_i|X_i])}_{Var(\tau(X_i))} + \underbrace{\mathbb{E}[Var(\tau_i|X_i)]}_{\text{Unexplained Variation}} . \tag{3}$$

The total variance of the ITE is decomposed into the variance of the CATE (variation explained by X_i) and the expected conditional variance of the ITE (variation unexplained by X_i).

It follows directly that H_0^{ITE} implies H_0^{CATE}. However, the converse is not true. It is possible for the CATE to be constant ($Var(\tau(X_i)) = 0$) while the ITE varies ($\mathbb{E}[Var(\tau_i|X_i)] > 0$), provided that the individual variations are orthogonal to X_i.

2.5 Challenges and Target Parameters

A central challenge in testing for H_0^{ITE} is the non-identifiability of $Var(\tau_i)$. The variance of the ITE can be written as:

$$Var(\tau_i) = Var(Y_i(1)) + Var(Y_i(0)) - 2Cov(Y_i(1), Y_i(0)).$$

While the marginal variances $Var(Y_i(d))$ are identified under Assumptions 3 and 4, the covariance term $Cov(Y_i(1), Y_i(0))$ is not, as it depends on the unobservable dependence structure between the potential outcomes (Heckman et al., 1997).

To circumvent this fundamental identification problem while enabling robust inference, we focus on three identifiable parameters that summarize different aspects of

treatment effect heterogeneity:

1. **CATE Variance (θ_0):** $\theta_0 = Var(\tau(X_i))$. This parameter directly tests H_0^{CATE}.
2. **Potential Outcome Variance Difference (λ_0):** $\lambda_0 = Var(Y_i(1)) - Var(Y_i(0))$. This parameter captures the treatment effect on the variance. Since H_0^{ITE} implies $\lambda_0 = 0$, finding $\lambda_0 \neq 0$ provides sufficient evidence for ITE heterogeneity.
3. **CATE-Baseline Covariance (κ_0):** $\kappa_0 = Cov(\tau(X_i), \mu_0(0, X_i))$. This ancillary parameter measures the association between the treatment effect and the baseline outcome, as explained by covariates, aiding the interpretation of λ_0.

In the subsequent sections, we develop DML estimators and corresponding hypothesis tests for these three parameters.

Suppose we observe n i.i.d. data $O_i = (Y_i, D_i, X_i)$ with $i = 1, ..., n$, distributed according to an unknown distribution P_0. D_i is a dummy variable indicating the status of treatment in a population of interest with $D_i = 1$ if individual i receives treatment and $D_i = 0$ otherwise. X_i is a vector of covariates that is potentially high-dimensional. Following the potential outcome framework or Rubin causal model by Rubin (1974), we define $Y_i(1)$ as the potential outcome of individual i with treatment and $Y_i(0)$ as the corresponding potential outcome without treatment. The observed outcome Y_i can be written as

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

The CATE is defined as

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x].$$

Further define the conditional expectations of the potential outcomes for the treated and control groups by $\mu_0(d, x) = \mathbb{E}[Y_i | D_i = d, X_i = x]$ with $d = 0, 1$. To achieve identification of the CATE, we maintain the following two assumptions.

Assumption 3 (Unconfoundedness)**.** $D_i \perp (Y_i(1), Y_i(0)) | X_i$.

Assumption 4 (Overlap)**.**

$$\exists\ \xi > 0,\ \ s.t.\ \xi \leq e_0(x) \leq 1 - \xi$$

where $e_0(x) = P(D_i = 1|X_i = x)$ *is the propensity score.*

Apply the AIPW (Robins et al., 1994) transformation and define the pseudo-outcome $\psi(o)$ as

$$\psi(O_i) = D_i \frac{Y_i - \mu_0(1, X_i)}{e_0(X_i)} - (1 - D_i)\frac{Y_i - \mu_0(0, X_i)}{1 - e_0(X_i)} + \mu_0(1, X_i) - \mu_0(0, X_i). \tag{4}$$

Under Assumption 3 and Assumption 4, standard results (e.g. Abrevaya et al., 2015; Semenova and Chernozhukov, 2021) ensure that the CATE is identified by the conditional expectation of the pseudo-outcome, namely,

$$\tau(x) = \mathbb{E}[\psi(O_i)|X_i = x]. \tag{5}$$

In this paper, we focus on the question of whether the CATE is identical for all subpopulations defined by the covariates and consider the following hypotheses for the CATE:

$$\begin{aligned} &H_0 : \tau(x) \text{ is constant for all } x, \\ &H_a : \tau(x) \text{ is not constant for some } x. \end{aligned} \tag{6}$$

Under the null hypothesis, the conditional average treatment effect is constant for all values of covariates. Throughout the paper, we refer to H_a in Equation (6) as treatment effect heterogeneity, which is particularly relevant for applications in policy evaluation and optimal treatment assignment. Heterogeneity may exist beyond the conditional average effect, and we will come back to this point in Section 3.3. As mentioned in Section 1, direct tests on $\tau(x)$ are difficult to implement and therefore, we will instead develop parameters that summarize the heterogeneity of the CATE and construct corresponding test statistics.

3 Tests for Heterogeneous Treatment Effect

3.1 CATE Projection Test (CPT)

Our first test is based on the projection coefficients of the CATE on the covariates. Let $b(X_i)$ be a function of the covariates which includes a constant. We will discuss the restrictions on $b(X_i)$[1] and some potentially useful specifications later.

Definition 3 (CATE Projection Coefficients)**.** *The projection coefficients of the CATE on $b(X_i)$ are*

$$\dot{\beta}_0 = \mathbb{E}[b(X_i)'b(X_i)]^{-1}\mathbb{E}[b(X_i)'(Y_i(1) - Y_i(0))]$$

Throughout the paper, we use parameters with dots above them to represent causal parameters, which are defined by potential outcomes. The projection of the CATE provides a linear approximation of the true CATE function, and projection coefficients are general targets of inference. A similar parameter is also considered by Semenova and Chernozhukov (2021), where $b(x)$ is set to be a vector of basis functions of x and they propose a series estimator of the CATE by minimizing the squared approximation error $\mathbb{E}[(\tau(X_i) - b(X_i)\dot{\beta}_0)^2]$.

In this paper, we focus on testing the existence of heterogeneity instead of accurately estimating the CATE. The change of the goalpost gives flexibility in specifying $b(x)$ and also largely reduces the complexity of the inference procedure. In the case when X_i is high-dimensional, for $\dot{\beta}_0$ to exist, we require $b(X_i)$ to be low-dimensional, which means that $b(X_i)$ needs to be a small subset of X_i. This imposes the assumption that the treatment effect is identified by a high-dimensional vector of covariates while the heterogeneity is only driven by a small proportion of the covariates.

Proposition 1 (Identification of CATE Projection Coefficients)**.** *Under Assumption 3 and Assumption 4 and assuming that $\mathbb{E}[b(X_i)'b(X_i)]$ is positive definite, the*

[1] In this paper X represents a $n \times q$ matrix where n is the number of observations and q is the number of covariates, while X_i represents the i-th row of X which is a q-dimensional row vector. The same layout applies to O and O_i.

projection coefficients $\dot{\beta}_0$ *is identified by*

$$\beta_0 = \mathbb{E}[b(X_i)'b(X_i)]^{-1}\mathbb{E}[b(X_i)'\psi(O_i)] = \dot{\beta}_0 \tag{7}$$

Given the identification result of $\dot{\beta}_0$ in Proposition 1, we now focus on the estimation and inference of β_0. We partition β_0 as $(\beta_c, \beta_x')'$ where β_c is the intercept and β_x is the rest of the projection coefficients, and then the hypotheses in Equation (6) can be translated into the following hypotheses.

$$\begin{aligned} H_0 &: \beta_x = 0, \\ H_a &: \beta_x \neq 0. \end{aligned} \tag{8}$$

We note that $\beta_x = 0$ is only a necessary condition for the null hypothesis of no heterogeneity but not a sufficient one, which means that rejecting $H_0 : \beta_x = 0$ implies the existence of heterogeneity, however, accepting H_0 does not imply there is no heterogeneous treatment effect. There exist special cases where β_0 fails to capture treatment effect heterogeneity. For example, the CATE function is parabolic in $b(x)$, leading to a zero projection coefficient. Consequently, the test based on β_0 will lose power against certain directions in the alternative space. In such cases, it suffices to include polynomials or discretized X_i in $b(X_i)$. Moreover, we will introduce two other tests that do not suffer from this issue in Section 3.2 and Section 3.3.

If a consistent and asymptotically normal estimator of β_0 is available, the null hypothesis in Equation (8) can be easily tested by a Wald test. In order to use machine learning methods to deal with potentially high-dimensional X_i and nonlinear functions of e_0 and μ_0, we appeal to recent developments in influence-function-based semiparametric estimation.

Proposition 2 (Influence Function of CATE Projection Coefficients)**.** *The influence function (IF) of* β_0 *is given by*

$$IF_\beta(O_i) = \mathbb{E}[b(X_i)'b(X_i)]^{-1}b(X_i)'\left(\psi(O_i) - b(X_i)\beta_0\right). \tag{9}$$

The IF in Equation (9) is the key to construct an estimator and perform hypothe-

sis tests on β_0. There are several semiparametric methods relying on the IF, such as one-step estimation (e.g. Pfanzagl and Wefelmeyer, 1985; Bickel et al., 1993), estimating equation method (e.g. Laan and Robins, 2003; Chernozhukov et al., 2018)[2], and targeted maximum likelihood estimation (Van Der Laan and Rubin, 2006). In this paper, we use the estimating equation method and develop conditions under which the estimator is $\sqrt{n}$-consistent.

Let $\hat{\mu}$ and $\hat{e}$ denote consistent estimators of μ_0 and e_0, and

$$\hat{\psi}(O_i) = D_i \frac{Y_i - \hat{\mu}(1, X_i)}{\hat{e}(X_i)} - (1 - D_i) \frac{Y_i - \hat{\mu}(0, X_i)}{1 - \hat{e}(X_i)} + \hat{\mu}(1, X_i) - \hat{\mu}(0, X_i). \tag{10}$$

The estimating equation estimator of β_0 is given by solving $n^{-1} \sum_{i=1}^{n} IF_\beta(O_i) = 0$, namely,

$$\hat{\beta} = \left(\sum_{i=1}^{n} b(X_i)' b(X_i) \right)^{-1} \left(\sum_{i=1}^{n} b(X_i)' \hat{\psi}(O_i) \right) \tag{11}$$

To develop the asymptotic properties of $\hat{\beta}$, we adopt the following notations. Let $\|\cdot\|$ be the L_2 norm of a vector, $\|\cdot\|_P = \mathbb{E}[(\cdot)^2]^{1/2}$ be the $L_2(P_0)$ norm of a function.

Proposition 3 (Asymptotic Distribution of $\hat{\beta}$)**.** *Assume that the matrices $\mathbb{E}[b(X_i)'b(X_i)]$ and $\sum_{i=1}^{n} b(X_i)'b(X_i)$ are positive definite, there exists constants $\xi, K \in (0, \infty)$ such that $\|b(X_i)\| < K$, $\mathbb{E}[(Y_i - \hat{\mu})^2 | X_i] < K$, $\hat{e}(X_i) \in (\xi, 1 - \xi)$ almost surely, and $\|e_0 - \hat{e}\|_P \|\mu_0 - \hat{\mu}\|_P = o_P(n^{-1/2})$. Suppose one of the two conditions holds:*

(i) Donsker condition: the quantities $\hat{\mu}(D_i, X_i)$, $D_i(Y_i - \hat{\mu}(1, X_i))/\hat{e}(X_i)$, and $(1 - D_i)(Y_i - \hat{\mu}(0, X_i))/(1 - \hat{e}(X_i))$ fall within a P-Donsker class with probability approaching 1.

(ii) Cross-fitting: The sample used to estimate $\hat{e}(x)$ and $\hat{\mu}(x)$ is independent of the sample used to construct $\hat{\beta}$.

Then $\hat{\beta}$ is a regular asymptotically linear estimator of β_0 with IF in Equation (9).

[2]The estimating equation method is called double/debiased machine learning in Chernozhukov et al. (2018), where the IF is recognized as the Neyman orthogonal score.

Hence,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_\beta)$$

with $V_\beta = \mathbb{E}[IF_\beta(O_i)IF_\beta(O_i)']$.

Proposition 3 provides a set of conditions under which $\hat{\beta}$ is asymptotically normal. The Donsker condition requires not fitting overly complex models and is usually assumed in the traditional nonparametric literature, such as Robins et al. (1994). However, the Donsker condition is not guaranteed to hold when modern machine learning methods are used. Alternatively, the cross-fitting condition proposed by Chernozhukov et al. (2018) is able to incorporate a large class of machine learning methods like lasso, random forest, and neural networks, allowing us to handle high-dimensional X_i. The condition $\|e_0 - \hat{e}\|_P \|\mu_0 - \hat{\mu}\|_P = o_P(n^{-1/2})$ is standard in the debiased machine learning estimator (Chernozhukov et al., 2018) or the augmented inverse propensity estimator of the ATE. This condition is also recognized as a double robustness property since it implies that $\hat{\beta}$ is $\sqrt{n}$-consistent if either $\hat{\mu}$ or $\hat{e}$ converges sufficiently fast. It is not surprising to see that the CATE projection coefficients, which are a linear transformation of the ATE, inherit the double robustness property.

A straightforward way to satisfy the cross-fitting condition is to first randomly split the data into two halves, then obtain $\hat{\mu}$ and $\hat{e}$ by fitting the machine learning methods on one half and calculate $\hat{\beta}$ on the other half. This typically results in the loss of efficiency, as the sample size used to estimate $\hat{\beta}$ is halved. A more efficient algorithm is provided in Chernozhukov et al. (2018), which works as follows. First, randomly partition the n observations into K equal folds indexed by $(I_k)_{k=1}^K$. Each fold has observations $n_K = n/K$. Let I_k^c denote the observations not in fold I_k. Using each I_k^c, we estimate the nuisance parameter μ_0 and e_0 by some machine learning method to obtain $\hat{\mu}_k$ and $\hat{e}_k$. Second, we construct $\hat{\psi}_k$ by plugging $\hat{\mu}_k$ and $\hat{e}_k$ into Equation (10) and calculate

$$\hat{\beta}_k = \left(\sum_{i \in I_k} b(X_i)'b(X_i)\right)^{-1} \left(\sum_{i \in I_k} b(X_i)'\hat{\psi}_k\right), \quad \hat{\beta} = \frac{1}{K}\sum_{k=1}^{K} \hat{\beta}_k. \tag{12}$$

Theorem 1 (CATE Projection Test)**.** *If Assumption 3, Assumption 4, and the con-*

ditions in Proposition 3 hold, under $H_0 : \beta_x = 0$*, the Wald statistic*

$$T = \hat{\beta}_x'(\hat{V}_{\beta_x}/n)^{-1}\hat{\beta}_x \xrightarrow{d} \chi^2_{dim(\beta_x)}$$

where $\hat{\beta}_x$ *is the estimator in Equation* (11) *or Equation* (12) *without intercept, and* $\hat{V}_{\beta_x}$ *is a consistent estimator of the submatrix of* V_β *corresponding to* β_x*.*

The CPT in Theorem 1 is essentially the joint test on the significance of CATE projection coefficients. It is convenient to implement by constructing $\hat{\psi}(O_i)$ and then regressing on $b(X_i)$ as if $\hat{\psi}(O_i)$ is observed. Given the formulas of $\hat{\beta}$ and V_β, the test statistic is readily available in the standard output of statistical software and no adjustment is needed for plugging in the machine learning estimators. For the α-level test, we reject the null hypothesis when T is greater than or equal to the $(1-\alpha)$-quantile of the $\chi^2_{\dim(\beta_x)}$ distribution. Another benefit of the CPT is that when Assumption 3 does not hold but a binary instrument is available, an extension to heterogeneity in the local average treatment effect is simple.

3.1.1 Extension to Binary IV

Suppose we observe data $O_i = (Y_i, D_i, X_i, Z_i)$ where Z_i is a binary instrument for the treatment D_i. Define the pseudo-outcome for some random variable R_i as

$$\psi^R(O_i) = Z_i\frac{R_i - r_0(1, X_i)}{q_0(X_i)} - (1 - Z_i)\frac{R_i - r_0(0, X_i)}{1 - q_0(X_i)} + r_0(1, X_i) - r_0(0, X_i)$$

where $r_0(z, x) = \mathbb{E}[R_i|Z_i = z, X_i = x]$ and $q_0(x) = \mathbb{E}[Z_i|X_i = x]$. Under the conditional local average treatment effect (CLATE) set up (e.g. Abadie, 2002), the CLATE $\tau_{late}(x)$ is identified by

$$\tau_{late}(x) = \frac{\mathbb{E}[\psi^Y(O_i)|X_i = x]}{\mathbb{E}[\psi^D(O_i)|X_i = x]}.$$

We focus on testing the null hypothesis of constant CLATE $H_0 : \tau_{late}(x)$ is constant for all x. Let $\alpha_0 = (\alpha_c, \alpha_x')'$ denote the projection coefficients of $\psi^Y(O_i)$ on $b(X_i)$, and $\gamma_0 = (\gamma_c, \gamma_x')'$ denote the projection coefficients of $\psi^D(O_i)$ on $b(X_i)$. The null

hypothesis can be translated into

$$H_0 : \alpha_x = \frac{\alpha_c}{\gamma_c} \cdot \gamma_x. \tag{13}$$

Consider the regression of stacked $\psi^Y(O_i)$ and $\psi^D(O_i)$

$$\begin{bmatrix} \psi^Y(O_i) \\ \psi^D(O_i) \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & b(X) & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & b(X) \end{bmatrix} \begin{bmatrix} \alpha_c \\ \gamma_c \\ \alpha_x \\ \gamma_x \end{bmatrix} + \begin{bmatrix} \epsilon^Y \\ \epsilon^D \end{bmatrix} = \tilde{b}(X)\beta_0 + \epsilon$$

where ϵ is the projection error of the stacked pseudo-outcome. We see that the coefficients in the above regression take the same form as in Equation (7). Therefore, Proposition 3 directly applies. We can construct the design matrix $\tilde{b}(X)$ and run the stacked regression, then the null hypothesis in Equation (13) can be tested by combining Theorem 1 with the delta method. Specifically, the Wald statistic under the null is given by

$$T_{late} = l(\hat{\beta})'(\nabla l(\hat{\beta})' \cdot (\hat{V}/n) \cdot \nabla l(\hat{\beta}))^{-1} l(\hat{\beta}) \xrightarrow{d} \chi^2_{\dim(\alpha_x)} \tag{14}$$

where $l(\hat{\beta}) = \hat{\alpha}_x - \frac{\hat{\alpha}_c}{\hat{\gamma}_c} \cdot \hat{\gamma}_x$, $\nabla l(\hat{\beta})$ is the gradient of $l(\hat{\beta})$, and $\hat{V}$ is a clustered consistent estimator of the covariance matrix of $\hat{\beta}$ with clusters defined by whether the stacked pseudo-outcome is $\psi^Y(O_i)$ or $\psi^D(O_i)$.

3.2 CATE Variance Test (CVT)

The CPT is an intuitive and practical test for detecting the presence of treatment effect heterogeneity. However, as discussed in Section 3.1, the test may lose power against certain directions in the alternative space and requires low-dimensional heterogeneity. To address these issues, we propose a variance test for the CATE.

Definition 4 (CATE Variance)**.** *The variance of the CATE is defined by*

$$\dot{\theta}_0 = Var(\tau(X_i)).$$

Assuming $\|\tau(X_i)\| < \infty$, $\dot{\theta}_0$ is well-defined. The variance of the CATE is a natural parameter for measuring the heterogeneity, which informs the extent to which the heterogeneity of the treatment effect is explained by covariates. The benefit of considering $\dot{\theta}_0$ in hypothesis testing is that it summarizes the potentially high-dimensional heterogeneity into a scalar parameter, which avoids the issue of high-dimensional inference as we discussed in CPT.

Proposition 4 (Identification of the CATE Variance)**.** *Let* $\nu_0(x) = \mu_0(1, x) - \mu_0(0, x)$. *Under Assumption 3 and 4, the variance of the CATE* $\dot{\theta}_0$ *is identified by*

$$\theta_0 = \mathbb{E}\left[(\nu_0(X_i) - \mathbb{E}[\nu_0(X_i)])^2\right] = \dot{\theta}_0.$$

A necessary and sufficient condition for a heterogeneous CATE is that its variance is not equal to zero. Then the hypotheses in Equation (6) can be translated into:

$$\begin{aligned} H_0 &: \theta_0 = 0, \\ H_a &: \theta_0 > 0. \end{aligned} \tag{15}$$

Now we proceed to develop consistent and asymptotically normal estimators of θ_0 based on the IF.

Proposition 5 (Influence Function for the CATE Variance)**.** *The IF of* θ_0 *is given by*

$$IF_\theta(O_i) = (\psi(O_i) - \tau_0)^2 - (\psi(O_i) - \nu_0(X_i))^2 - \theta_0 \tag{16}$$

where $\tau_0 = \mathbb{E}[\nu_0(X_i)]$ *is the ATE.*

Based on $IF_\theta(O_i)$, the estimating equation estimator of θ_0 is given by

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left((\hat{\psi}(O_i) - \hat{\tau})^2 - (\hat{\psi}(O_i) - \hat{\nu}(X_i))^2\right) \tag{17}$$

where $\hat{\tau}$ is a consistent estimator of τ_0. One choice of $\hat{\tau}$ is the augmented inverse propensity weighting estimator $n^{-1}\sum_{i=1}^{n}\hat{\psi}(O_i)$ (Chernozhukov et al., 2018).

Equation (17) also gives the intuition for the CATE variance. Since $\tau(X_i)$ works as

the conditional expectation of $\psi(O_i)$, we can estimate $Var(\tau(X_i))$ by subtracting the variance of the "error" term $\psi(O_i) - \nu_0(X_i)$ from $Var(\psi(O_i))$.

Proposition 6 (Asymptotic Distribution of $\hat{\theta}$)**.** *Assume that there exists constants* $\xi, K \in (0, \infty)$ *such that* $Var(\psi(O_i)|X_i) < K$, $\mathbb{E}[(Y_i - \hat{\mu})^2|X_i] < K$, $(\hat{\nu}(X_i) - \hat{\tau})^2 < K$, $\hat{e}(X_i) \in (\xi, 1-\xi)$ *almost surely. Also assume that* $\|e_0 - \hat{e}\|_P \|\mu_0 - \hat{\mu}\|_P = o_P(n^{-1/2})$, $\|\tau_0 - \hat{\tau}\|_P$ *and* $\|\nu_0 - \hat{\nu}\|_P$ *are both* $o_P(n^{-1/4})$. *Suppose one of the two conditions holds:*

(i) Donsker condition: the quantities $(\hat{\psi}(O_i) - \hat{\nu}(X_i))^2$, $(\hat{\psi}(O_i) - \hat{\tau})^2$, *and* $\hat{\psi}(O_i)(\hat{\nu}(X_i) - \hat{\tau})$ *fall within a P-Donsker class with probability approaching 1.*

(ii) Cross-fitting: The sample used to estimate $\hat{e}(x)$, $\hat{\mu}(x)$, $\hat{\nu}(x)$, *and* $\hat{\tau}$ *is independent of the sample used to construct* $\hat{\theta}$.

Then $\hat{\theta}$ *is a regular asymptotically linear estimator of* θ_0 *with IF in Equation (16). Hence,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V_\theta)$$

with $V_\theta = \mathbb{E}[IF_\theta(O_i)^2]$.

Compared to Proposition 3, an additional convergence rate condition is imposed by $\|\nu_0 - \hat{\nu}\|_P = o_P(n^{-1/4})$. By definition, $\hat{\nu}(x)$ can be constructed by $\hat{\mu}(1, x) - \hat{\mu}(0, x)$. The condition is satisfied if $\|\mu_0 - \hat{\mu}\|_P = o_P(n^{-\kappa})$ with $\kappa \geq 1/4$. Then $\|e_0 - \hat{e}\|_P \|\mu_0 - \hat{\mu}\|_P = o_P(n^{-1/2})$ implies that $\|e_0 - \hat{e}\|_P = o_P(n^{-1/2+\kappa})$. It suggests that $\hat{e}$ is allowed to converge at a slower rate but $\hat{\mu}$ has to converge faster. A better way to construct $\hat{\nu}$ is to use $\hat{\psi}$, such that only the condition $\|e_0 - \hat{e}\|_P \|\mu_0 - \hat{\mu}\|_P = o_P(n^{-1/2})$ binds and the double robustness property is preserved.

The construction of a cross-fitting estimator of $\hat{\theta}$ follows a similar procedure to that described in Section 3.1, which we omit for brevity. Given the result in Proposition 6, it is straightforward to construct a Z-test for the null hypothesis in Equation (15).

Theorem 2 (CATE Variance Test)**.** *If Assumption 3, Assumption 4, and the conditions in Proposition 6 hold, under* $H_0 : \theta_0 = 0$, *the Z-statistic*

$$Z_\theta = \frac{\hat{\theta}}{\sqrt{\hat{V}_\theta / n}} \xrightarrow{d} N(0, 1)$$

where $\hat{V}_\theta$ is a consistent estimator of V_θ.

Since the variance cannot be negative, a one-sided test should be used. The null hypothesis is rejected at significance level α if $Z_\theta > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution.

3.3 Potential Outcome Variance Test (POVT)

In this section, we propose a test based on the difference between the variances of the potential outcomes.

Definition 5 (Variance Difference of Potential Outcomes)**.** *The variance difference of the potential outcomes is defined by*

$$\dot{\lambda}_0 = Var(Y_i(1)) - Var(Y_i(0)).$$

The causal parameter $\dot{\lambda}_0$ can be interpreted as the treatment effect on the variance of the potential outcomes. To see how the variance difference of the potential outcomes can be used to detect heterogeneity, let $\tau_i = Y_i(1) - Y_i(0)$ denote the individual treatment effect. We have

$$Y_i(1) = \tau_i + Y_i(0). \tag{18}$$

If the individual treatment effect is constant, then the variances of the potential outcomes are the same, i.e., $Var(Y_i(1)) = Var(Y_i(0))$. The equality fails to hold in the presence of heterogeneity, thus making $\dot{\lambda}_0$ a measure of treatment effect heterogeneity.

We propose an identification strategy by $Var(Y_i(d)) = \mathbb{E}[Y_i(d)^2] - \mathbb{E}[Y_i(d)]^2$. Formally, let

$$\begin{aligned}
\psi^1(O_i) &= \frac{D_i(Y_i - \mu_0(1, X_i))}{e_0(X_i)} + \mu_0(1, X_i) \\
\psi^0(O_i) &= \frac{(1 - D_i)(Y_i - \mu_0(0, X_i))}{1 - e_0(X_i)} + \mu_0(0, X_i) \\
\phi^1(O_i) &= \frac{D_i(Y_i^2 - \mu_0^2(1, X_i))}{e_0(X_i)} + \mu_0^2(1, X_i) \\
\phi^0(O_i) &= \frac{(1 - D_i)(Y_i^2 - \mu_0^2(0, X_i))}{1 - e_0(X_i)} + \mu_0^2(0, X_i)
\end{aligned}$$

where $\mu_0^2(d,x) = \mathbb{E}[Y_i^2(d)|X_i = x]$ for $d = 0, 1$. Note that $\psi(O_i) = \psi^1(O_i) - \psi^0(O_i)$. Then the identification of $\dot{\lambda}_0$ is given by the following proposition.

Proposition 7 (Identification of the Variance Difference of Potential Outcomes)**.** *Under Assumption 3 and 4, the variance difference of the potential outcomes is identified by*

$$\lambda_0 = \left(\mathbb{E}[\phi^1(O_i)] - \mathbb{E}[\psi^1(O_i)]^2\right) - \left(\mathbb{E}[\phi^0(O_i)] - \mathbb{E}[\psi^0(O_i)]^2\right) = \dot{\lambda}_0.$$

And the hypothesis in Equation (6) can be translated into

$$\begin{aligned} H_0 &: \lambda_0 = 0, \\ H_a &: \lambda_0 \neq 0. \end{aligned} \tag{19}$$

It is important to note that $\lambda_0 = 0$ serves as a sufficient but not necessary condition for the null hypothesis that $\tau(x)$ is constant for all x. This distinction arises because λ_0 captures broader distributional treatment effects beyond the heterogeneity of conditional average effects $\tau(x)$. There may exist cases where these distributional effects aggregate to a constant for each value of x, resulting in $\lambda_0 \neq 0$ despite the null hypothesis of constant $\tau(x)$ holding true. Essentially, $\lambda_0 = 0$ is equivalent to the null hypothesis that τ_i is constant for all i. While the test in Equation (19) can be used to investigate distributional effects, it needs to be treated with caution for examining heterogeneity in conditional average effects.

Proposition 8 (Influence Function for the Variance Difference of Potential Outcomes)**.** *The IF of λ_0 is given by*

$$\begin{aligned} IF_\lambda(O_i) = \phi^1(O_i) - \mathbb{E}[\phi^1(O_i)] - 2\mathbb{E}[\psi^1(O_i)]\left(\psi^1(O_i) - \mathbb{E}[\psi^1(O_i)]\right) \\ - \left(\phi^0(O_i) - \mathbb{E}[\phi^0(O_i)] - 2\mathbb{E}[\psi^0(O_i)]\left(\psi^0(O_i) - \mathbb{E}[\psi^0(O_i)]\right)\right) \end{aligned} \tag{20}$$

Based on the IF, the estimating equation estimator of λ_0 is given by

$$\begin{aligned} \hat{\lambda} = \frac{1}{n}\sum_{i=1}^n \left(\hat{\phi}^1(O_i) + \left(\frac{1}{n}\sum_{i=1}^n \hat{\psi}^1(O_i)\right)^2 - \hat{\psi}^1(O_i)\cdot\frac{2}{n}\sum_{i=1}^n \hat{\psi}^1(O_i)\right) \\ - \frac{1}{n}\sum_{i=1}^n \left(\hat{\phi}^0(O_i) + \left(\frac{1}{n}\sum_{i=1}^n \hat{\psi}^0(O_i)\right)^2 - \hat{\psi}^0(O_i)\cdot\frac{2}{n}\sum_{i=1}^n \hat{\psi}^0(O_i)\right) \end{aligned} \tag{21}$$

where $\hat{\phi}^d$ and $\hat{\psi}^d$ are estimators of ϕ^d and ψ^d, analogous to the plug-in estimator $\hat{\psi}$ in Equation (10).

Based on the IF, we can derive the asymptotic distribution of $\hat{\lambda}$ and construct a Z-test similar to Proposition 6 and Theorem 2.

Proposition 9 (Asymptotic Distribution of $\hat{\lambda}$)**.** *Assume that there exists constants* $\xi, K \in (0, \infty)$ *such that* $\hat{e}(X_i) \in (\xi, 1-\xi)$ *almost surely,* $\mathbb{E}[(Y_i - \hat{\mu})^2 | X_i] < K$, $\mathbb{E}[(Y_i^2 - \hat{\mu}^2)^2 | X_i] < K$, $\|e_0 - \hat{e}\|_P \|\mu_0 - \hat{\mu}\|_P = o_P(n^{-1/2})$, *and* $\|e_0 - \hat{e}\|_P \|\mu_0^2 - \hat{\mu}^2\|_P = o_P(n^{-1/2})$. *Suppose one of the two conditions holds:*

(i) Donsker condition: the quantities $\hat{\phi}^d(O_i)$, $\hat{\psi}^d(O_i)$, *and* $\hat{\psi}^{d,2}(O_i)$ *for* $d = 0, 1$ *fall within a P-Donsker class with probability approaching 1.*

(ii) Cross-fitting: The sample used to estimate $\hat{e}(x)$, $\hat{\mu}(x)$, $\hat{\mu}^2(x)$ *is independent of the sample used to construct* $\hat{\lambda}$.

Then $\hat{\lambda}$ *is a regular asymptotically linear estimator of* λ_0 *with IF in Equation (20). Hence,*

$$\sqrt{n}(\hat{\lambda} - \lambda_0) \xrightarrow{d} N(0, V_\lambda)$$

with $V_\lambda = \mathbb{E}[IF_\lambda(O_i)^2]$.

Theorem 3 (Potential Outcome Variance Test)**.** *If Assumption 3, Assumption 4, and the conditions in Proposition 9 hold, under* $H_0 : \lambda_0 = 0$, *the Z-statistic*

$$Z_\lambda = \frac{\hat{\lambda}}{\sqrt{\hat{V}_\lambda / n}} \xrightarrow{d} N(0, 1)$$

where $\hat{V}_\lambda$ *is a consistent estimator of* V_λ.

Unlike the one-sided CVT, the POVT is a two-sided test. The reason is as follows. By decomposing $Var(Y_i(1))$ according to Equation (18), we have

$$Var(Y_i(1)) = Var(Y_i(0)) + Var(\tau_i) + 2Cov(\tau_i, Y_i(0)).$$

This decomposition yields a variance difference of $\dot{\lambda}_0 = Var(\tau_i) + 2Cov(\tau_i, Y_i(0))$.

Without additional assumptions on $Cov(\tau_i, Y_i(0))$, $\dot{\lambda}_0$ can take both positive and negative values under H_a.

An alternative quantity for testing the equality of variances is the ratio

$$Var(Y_i(1))/Var(Y_i(0)),$$

which was previously considered by Ding et al. (2016) in experiments. Proposition 7 suggests that this quantity is identified using our strategy, allowing an extension to observational data under Assumption 3. However, the F-test for the variance ratio requires the marginal distribution of potential outcomes to be normal, an assumption that is difficult to verify and impractical in empirical applications. Based on Equation (18), another possible way to test heterogeneity is by the variance of the individual treatment effect $Var(Y_i(1) - Y_i(0))$. This quantity cannot be identified without imposing assumptions on the joint distribution of $(Y_i(1), Y_i(0))$. Therefore, we focus on $\dot{\lambda}_0$ as it is identified under standard assumptions in the treatment effect literature.

A final note is that, since both CPT and POVT are two-sided tests of equality, they can be combined to form a joint Wald test. According to Proposition 3 and 9, each estimator of β and λ is asymptotically normal based on the IFs. One can stack IF_β and IF_λ to obtain the variance-covariance matrix of $\hat{\beta}$ and $\hat{\lambda}$, and the construction of the joint Wald test statistic follows naturally. This joint testing approach enhances statistical power compared to individual tests. As discussed previously, the CATE projection test may fail to reject the null hypothesis in certain cases, combining it with POVT can address this limitation. On the other hand, combining the CPT benefits POVT, as variance, being a second-order statistic, typically requires larger samples to achieve asymptotic properties. Thus, a joint test will leverage the complementary strengths of each test to provide more robust detection of treatment effect heterogeneity. This idea of using a joint test to improve statistical power was previously considered by Kleibergen (2005) in the context of GMM identification tests.

4 Simulation

In this section, we conduct Monte Carlo simulations to evaluate the size and power of the proposed tests in finite samples. For each simulation exercise, we generate 1000 datasets with sample size $n = 200, 400, 800, 1600$, and report the rejection proportions at 5% significance level. The data generating process is specified as follows:

- Low-dimensional ($p = 10$) correlated covariates, non-sparse $\mu(\cdot, x)$, $e(x)$ and $\tau(x)$

  1. Constant CATE model:

  $$X_i \sim N(0, \Sigma), \quad \Sigma_{ii} = 1,\ \Sigma_{ij, i\neq j} \sim \text{Uniform}(0.1, 0.3)$$
  $$D_i | X_i = x \sim \text{Bern}(e(x)), \quad e(x) = \frac{1}{1 + \exp(-x'\alpha)},\ \alpha = (0.1, \ldots, 0.1)_{10\times 1}$$
  $$Y_i | D_i = d, X_i = x \sim N(x'\beta + 2d, 1/2), \quad \beta = (0.5, \ldots, 0.5)_{10\times 1}$$

  2. Linear outcome model: treatment and covariates are distributed as above. The outcome is defined by

  $$Y_i | D_i = d, X_i = x \sim N(x'(\beta + d \cdot \gamma_1 + (1-d)\cdot \gamma_0), 1/2),$$
  $$-\gamma_0 = \gamma_1 = (1, \ldots, 1)_{10\times 1}$$

  3. Kinked outcome model: treatment and covariates are distributed as above. The outcome is defined by

  $$Y_i | D_i = d, X_i = x \sim N((1-d)\cdot x'\gamma + d \cdot \text{diag}(\mathcal{I}(x > 0))\gamma, 1/2),$$
  $$\gamma = (2, \ldots, 2)_{10\times 1}$$

  4. Nonlinear outcome model: treatment and covariates are distributed as above. The outcome is defined by

  $$Y_i | D_i = d, X_i = x \sim N(\exp(x'\beta) + d \cdot x'\gamma, 1/2),$$
  $$\gamma = (2, \ldots, 2)_{10\times 1}$$

- High-dimensional ($p = 100$) uncorrelated covariates, sparse $\mu(\cdot, x)$, $e(x)$ and $\tau(x)$

    1. Constant CATE model: Models similar to the low-dimensional case, but we create sparsity by setting the parameters differently:

$$\begin{aligned} &\Sigma \text{ is identity matrix,} \\ &\alpha = ((0.2, \ldots, 0.2)_{10\times 1}, (0, \ldots, 0)_{90\times 1}), \\ &\beta = ((2, \ldots, 2)_{20\times 1}, (0, \ldots, 0)_{80\times 1}) \end{aligned}$$

    2. Linear outcome model: Models similar to the low-dimensional case, with parameters defined as above, except that:

$$-\gamma_0 = \gamma_1 = ((5, \ldots, 5)_{50\times 1}, (0, \ldots, 0)_{50\times 1})$$

    3. Kinked outcome model: Models similar to the low-dimensional case, with parameters defined as in constant CATE model, except that:

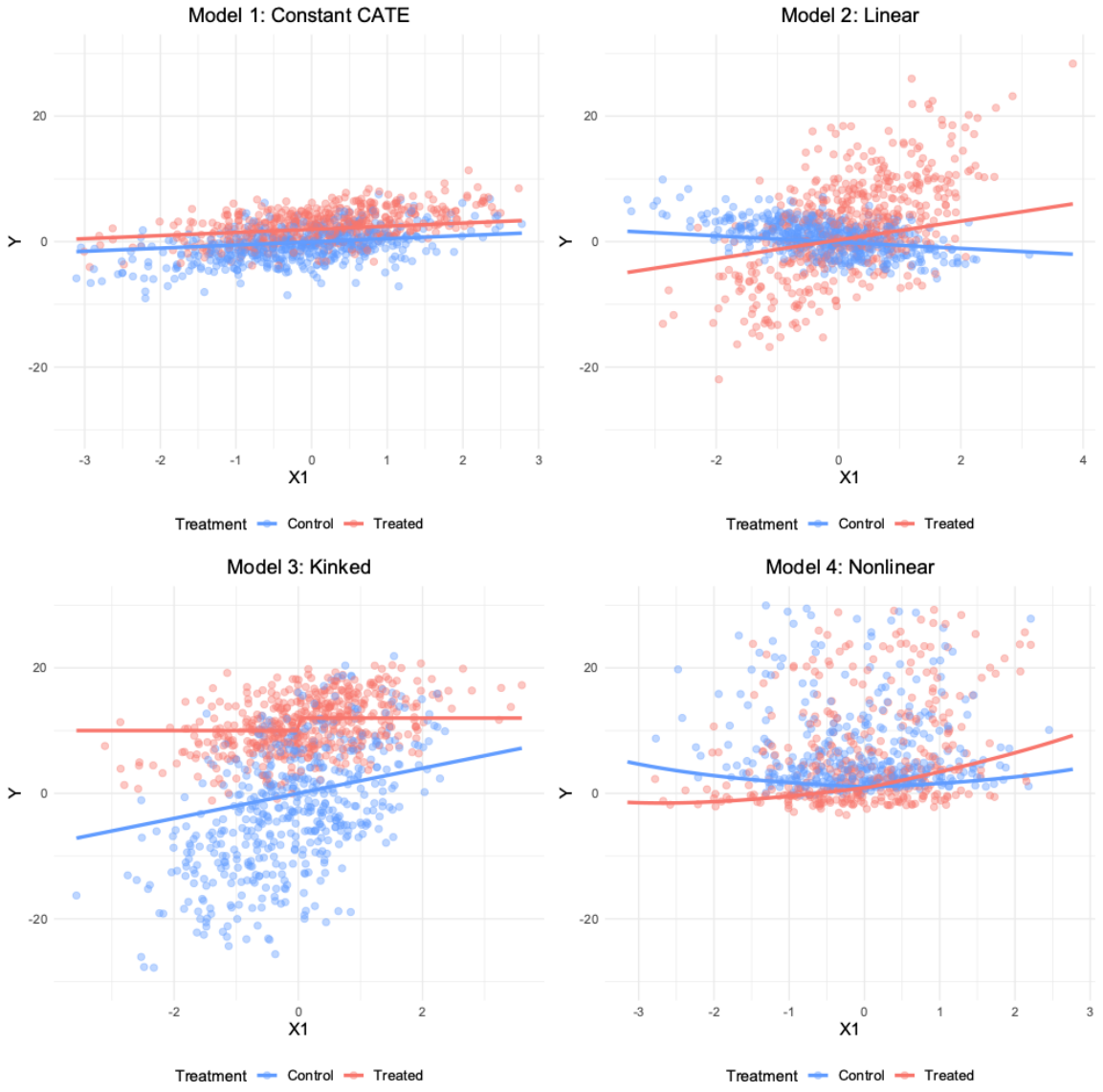
$$\gamma = ((10, \ldots, 10)_{50\times 1}, (0, \ldots, 0)_{50\times 1})$$

    4. Nonlinear outcome model: Models similar to the low-dimensional case, with parameters defined as in constant CATE model, except that:

$$\begin{aligned} &\beta = ((1, \ldots, 1)_{20\times 1}, (0, \ldots, 0)_{80\times 1}), \\ &\gamma = ((5, \ldots, 5)_{50\times 1}, (0, \ldots, 0)_{50\times 1}) \end{aligned}$$

Figure 1 provides visualizations of the data generating processes through the scatter plots of Y_i against X_{1i}, alongside the true relationships between potential outcomes and X_{1i}. The models are designed to reflect different patterns of treatment effects. The constant CATE model falls under the null of Equation (6), while the other three models fall under the alternative. The linear and nonlinear models are specifically designed to have a zero ATE, where the conventional ATE-targeted approaches such as OLS or IPW would fail to detect the existence of treatment effects.

Figure 1: Sketches of the Data Generating Processes



We consider two machine learning algorithms for estimating the nuisance parameters: Lasso and XGBoost[3]. The cross-fitting estimators analogous to Equation (12) are used to estimate β_0, θ_0 and λ_0, and we implement the tests proposed in Theorems 1, 2 and 3.

[3]We use the `glmnet` R package for Lasso and the `xgboost` package for XGBoost. For both algorithms, 5-fold cross-validation is used to select the optimal tuning parameters.

Table 1: Simulated Rejection Rate

		Lasso			XGBoost		
	n	CPT	CVT	POVT	CPT	CVT	POVT
Low Dimension							
Model 1	200	2.9	0	3.6	3.4	0	1.1
	400	2.5	0	3.9	4.8	0	2.8
	800	2.5	0	4.5	2.1	0	3.9
	1600	2.5	0	5.0	2.0	0	4.7
Model 2	200	100	75.1	99.9	100	99.5	82.8
	400	100	83.6	100	100	100	98.9
	800	100	83.6	100	100	100	100
	1600	100	92.2	100	100	100	100
Model 3	200	96.5	66.5	100	99.1	51.0	77.4
	400	99.9	74.2	100	100	96.1	98.6
	800	100	78.0	100	100	100	100
	1600	100	81.8	100	100	100	100
Model 4	200	99.8	84.3	97.2	98.0	68.7	78.7
	400	99.9	89.5	98.9	99.8	89.9	96.1
	800	100	92.9	100	100	96.9	99.9
	1600	100	95.4	100	100	99.0	99.9
High Dimension							
Model 1	200	3.3	0.1	3.0	3.2	0	3.0
	400	2.4	0	3.8	2.9	0	3.2
	800	2.3	0	4.1	2.8	0	3.9
	1600	2.1	0	4.2	2.8	0	4.4
Model 2	200	88.6	43.3	91.1	89.4	69.7	85.9
	400	98.3	59.8	94.2	98.6	85.5	92.4
	800	100	64.7	97.2	100	98.3	99.9
	1600	100	69.0	100	100	100	99.9
Model 3	200	77.2	40.5	88.6	84.3	55.6	89.3
	400	89.6	47.0	97.9	94.9	61.8	96.1
	800	99.9	54.2	100	100	67.2	99.9
	1600	100	59.1	100	100	71.2	100
Model 4	200	91.2	75.4	89.8	90.3	60.9	67.8
	400	100	78.8	96.6	96.7	77.9	80.6
	800	100	81.0	99.5	100	83.2	98.5
	1600	100	84.9	100	100	87.4	100

Empirical rejection proportions, in percentage points, at 5% significance level based on 1000 simulations.

The empirical rejection proportions of the tests are presented in Table 1. The POVT maintains rejection rates approximating the nominal 5% level under the null hypothesis when the sample size increases, whereas both CPT and CVT exhibit size distortions. In terms of power, CPT and POVT reach a satisfactory rejection rate close to 100% at $n = 400$ across all four models. In contrast, CVT requires larger sample sizes to attain comparable power and exhibits a particular weakness in detecting heterogeneity within the kinked outcome model. While all tests perform marginally better in the low-dimensional setting, this difference is not substantial. Finally, we observe that the performance of the tests also depends on the choice of machine learning algorithms, particularly for CVT, where XGBoost-based implementations consistently outperform their Lasso-based counterparts and the difference is especially pronounced in low-dimensional settings. Given that no single test demonstrates uniform superiority across all data generating processes and sample sizes examined, we recommend researchers implement all three tests concurrently and form conclusions based on their collective evidence.

5 Application

In this section, we demonstrate the application of the proposed tests to the NSW job training program data. In this program, participants were randomly assigned to either a job training program or a control group, and the treatment effect on future earnings can be estimated by directly comparing outcomes of the treated and control groups. In order to evaluate the validity of econometric estimators of treatment effects, LaLonde (1986) compared the treated individuals from the experiment to control groups drawn from two survey datasets: the Panel Study of Income Dynamics (PSID) and Current Population Survey (CPS). The resulting datasets have been extensively analyzed in the influential works by Dehejia and Wahba (1999); Smith and Todd (2005); Angrist and Pischke (2009); Słoczyński (2022), among others. In the context of CATE hypothesis testing, the dataset was analyzed by Hsu (2017) and Dai et al. (2023), who focused specifically on the heterogeneity with respect to the age of the individuals. Using the proposed tests, we will examine the heterogeneity with respect to all available covariates.

The dataset we use is NSW-CPS, which contains 185 treated units from the experiment and 15992 control units from the CPS. The outcome Y_i is the earnings in 1978, and the treatment D_i is a binary indicator of whether the individual received the job training. We consider the same set of covariates as those in column 4 of Table 3.3.3 in Angrist and Pischke (2009), which includes age, age squared, education, dummy variables for black and Hispanic, marital status, dummy indicator for high school degree, and pre-treatment earnings in 1974 and 1975. For this set of covariates X_i, we test for $H_0 : \tau(x) = c$ for all covariate values x. For nuisance parameter estimation, we employ XGBoost, as well as a parametric approach with OLS for outcome regression and logistic regression for propensity score. Since OLS and logistic regression fall into the P-Donsker class trivially, we implement the estimators in Equation (11), (17) and (21) for CPT, CVT and POVT, respectively.

The test statistics and their corresponding p-values are presented in Table 2. All tests reject the null of constant CATE at 1% significance level, providing strong evidence for the presence of heterogeneous treatment effects. Table 3 presents parameter estimates from both parametric and XGBoost approaches. While there are some differences in the magnitude of the estimates between the parametric and XGBoost approaches, the signs of the estimates remain consistent. These estimates can offer additional insights into the heterogeneity. For example, parametric estimates of $\hat{\beta}$ suggest that marital status and earnings in 1975 are potential drivers of the heterogeneity; the negative estimate of $\hat{\lambda}$ indicates that the treatment group exhibits smaller variance of earnings compared to the control group, suggesting that the job training program is effective in reducing the individual earnings gaps. These findings complement the conventional ATE-focused analyses.

Table 2: Test Statistics and P-values

	Parametric		XGBoost	
Test	Stat.	p-value	Stat.	p-value
CPT	7722.17	< 0.01	4806.72	< 0.01
CVT	9.55	< 0.01	10.78	< 0.01
POVT	-2.91	< 0.01	-10.82	< 0.01

Table 3: Parameter Estimates and Standard Errors

		Parametric		XGBoost	
Parameter		Est.	S.E.	Est.	S.E.
	age	0.01	0.12	0.07	0.31
	age^2	0.00	0.00	-0.00	0.00
	$educ$	0.18	0.11	-0.15	0.31
β	$black$	1.18	2.34	5.24	7.00
	$hispanic$	-0.48	0.82	-0.16	2.45
	$married$	1.57	0.66	0.70	1.61
	$nodegree$	-1.82	0.60	-0.97	1.35
	$re74$	-0.09	0.18	-0.05	0.55
	$re75$	-0.50	0.16	-0.66	0.46
θ		23.70	2.48	51.91	4.82
λ		-9.72	3.34	-68.18	6.30

6 Conclusion

This paper develops three hypothesis tests for detecting heterogeneous treatment effects in observational studies, bridging the gap between average treatment effect inference and heterogeneity analysis. Our tests build on distinct but complementary measures: the CATE Projection Test examines linear projections of treatment effects on covariates, the CATE Variance Test examines variation in conditional effects, and the Potential Outcome Variance Test investigates the variance change in potential outcomes. By developing influence functions for these parameters, we enable valid inference under both classical nonparametric and modern machine learning frameworks, with Monte Carlo simulations demonstrating good size control and power against various data generating processes, including high-dimensional settings.

The empirical application to the NSW job training program reveals significant heterogeneity in treatment effects, highlighting the importance of moving beyond average effects in policy evaluation. Our work contributes to the growing literature on heterogeneous treatment effects by providing practitioners with practical tools to determine when more detailed analysis of effect heterogeneity is warranted. Since the building block of our tests is the influence function of the CATE, it is possible to extend the tests to other causal parameters, such as heterogeneity in the treatment effect on

the treated or quantile treatment effects, and we refer to Chernozhukov et al. (2018) and Firpo (2007) for the corresponding influence functions. We leave this for future research.

References

Abadie, A. (2002): "Bootstrap tests for distributional treatment effects in instrumental variable models," *Journal of the American statistical Association*, 97, 284–292.

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015): "Estimating conditional average treatment effects," *Journal of Business & Economic Statistics*, 33, 485–505.

Angrist, J. D. and J.-S. Pischke (2009): *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Athey, S. and G. Imbens (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

Athey, S. and G. W. Imbens (2017): "The state of applied econometrics: Causality and policy evaluation," *Journal of Economic perspectives*, 31, 3–32.

Bickel, P. J., C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov (1993): *Efficient and adaptive estimation for semiparametric models*, vol. 4, Springer.

Chang, M., S. Lee, and Y.-J. Whang (2015): "Nonparametric tests of conditional treatment effects with an application to single-sex schooling on academic achievements," *The Econometrics Journal*, 18, 307–346.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," .

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2008): "Nonparametric tests for treatment effect heterogeneity," *The Review of Economics and Statistics*, 90, 389–405.

Dai, M., W. Shen, and H. S. Stern (2023): "Nonparametric tests for treatment effect heterogeneity in observational studies," *Canadian Journal of Statistics*, 51, 531–558.

Dehejia, R. H. and S. Wahba (1999): "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American statistical Association*, 94, 1053–1062.

Ding, P., A. Feller, and L. Miratrix (2016): "Randomization inference for treatment effect variation," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78, 655–671.

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

Firpo, S. (2007): "Efficient semiparametric estimation of quantile treatment effects," *Econometrica*, 75, 259–276.

Hampel, F. R. (1974): "The influence curve and its role in robust estimation," *Journal of the american statistical association*, 69, 383–393.

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts," *The Review of Economic Studies*, 64, 487–535.

Hsu, Y.-C. (2017): "Consistent tests for conditional treatment effects," *The econometrics journal*, 20, 1–22.

Kennedy, E. H., S. Balakrishnan, and M. G'sell (2020): "Sharp instruments for classifying compliers and generalizing causal effects," *The Annals of Statistics*, 2008–2030.

Kleibergen, F. (2005): "Testing parameters in GMM without assuming that they are identified," *Econometrica*, 73, 1103–1123.

Laan, M. J. and J. M. Robins (2003): *Unified methods for censored longitudinal data and causality*, Springer.

LaLonde, R. J. (1986): "Evaluating the econometric evaluations of training programs with experimental data," *The American economic review*, 604–620.

Nie, X. and S. Wager (2021): "Quasi-oracle estimation of heterogeneous treatment effects," *Biometrika*, 108, 299–319.

Pfanzagl, J. and W. Wefelmeyer (1985): "Contributions to a general asymptotic statistical theory," *Statistics & Risk Modeling*, 3, 379–388.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American statistical Association*, 89, 846–866.

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, 66, 688.

Sant'Anna, P. H. (2021): "Nonparametric tests for treatment effect heterogeneity with duration outcomes," *Journal of Business & Economic Statistics*, 39, 816–832.

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

Słoczyński, T. (2022): "Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights," *Review of Economics and Statistics*, 104, 501–509.

Smith, J. A. and P. E. Todd (2005): "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of econometrics*, 125, 305–353.

Van Der Laan, M. J. and D. Rubin (2006): "Targeted maximum likelihood learning," *The international journal of biostatistics*, 2.

Van der Vaart, A. (1998): "Functional delta method," *Asymptotic Statistics*, 291–303.

Wager, S. and S. Athey (2018): "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 113, 1228–1242.

Zimmert, M. and M. Lechner (2019): "Nonparametric estimation of causal heterogeneity under high-dimensional confounding," *arXiv preprint arXiv:1908.08779*.

A Appendix: Proofs

Proof of Proposition 1. By the law of iterated expectation, we have $\mathbb{E}[b(X_i)'\psi(O_i)] = \mathbb{E}[b(X_i)'\mathbb{E}[\psi(O_i)|X_i]]$. Under Assumption 3 and 4, $\mathbb{E}[\psi(O_i)|X_i] = \mathbb{E}[Y_i(1) - Y_i(0)|X_i]$, which follows from standard identification of augmented inverse propensity weighting estimators. □

Proof of Proposition 2. We follow a simple strategy provided by Hampel (1974) to derive the influence function. Consider β_0 as a statistical functional $\beta(P_0)$ which maps the data distribution to the parameter space. Let f_t denote the probability density function of distribution P_t and $\mathbb{I}_{\tilde{o}}(o)$ denote the Dirac delta function with respect to $\tilde{o}$, that is, the density which is a point mass at o and zero elsewhere. Consider a perturbed density of f_0 in the direction of a single observation $\tilde{o}$,

$$f_t(o) = t\mathbb{I}_{\tilde{o}}(o) + (1-t)f_0(o)$$

with $t \in [0,1]$. The IF of $\beta(P_0)$ given by

$$IF_\beta(O_i) = \left.\frac{d\beta(P_t)}{dt}\right|_{t=0}$$

We begin with deriving the IF of $\mathbb{E}[Y_i(1)]$. First, notice that

$$\frac{d}{dt}f_t(o) = \mathbb{I}_{\tilde{o}}(o) - f_0(o), \tag{A.1}$$

which implies the IF of expectation $\mathbb{E}[O_i]$ is $O_i - \mathbb{E}[O_i]$. Under Assumption 3 and Assumption 4,

$$\mathbb{E}[Y_i(1)] = \mathbb{E}\left[D_i\frac{Y_i - \mu_0(1,X_i)}{e_0(X_i)} + \mu_0(1,X_i)\right] = \mathbb{E}[\mathbb{E}[Y_i|D_i=1,X_i]].$$

Denote $\mathbb{E}_{P_t}[\cdot]$ as the expectation evaluated at P_t and $\mathbb{E}[\cdot] = \mathbb{E}_{P_0}[\cdot]$. The IF of $\mathbb{E}[Y_i(1)]$

is then

$$
\begin{aligned}
&\frac{d}{dt}\mathbb{E}_{P_t}[\mathbb{E}_{P_t}[Y_i|D_i=1,X_i]]\Big|_{t=0} \\
=&\frac{d}{dt}\int\int y\frac{f_t(y,1,x)f_t(x)}{f_t(1,x)}dydx\Big|_{t=0} \qquad (A.2)\\
=&\int\int y\left(\frac{f(x)}{f(1,x)}\frac{d}{dt}f_t(y,1,x)\Big|_{t=0}+\frac{f(y,1,x)}{f(1,x)}\frac{d}{dt}f_t(x)\Big|_{t=0}-\frac{f(y,1,x)f(x)}{f(1,x)^2}\frac{d}{dt}f_t(1,x)\Big|_{t=0}\right)dydx \\
=&\int\int y\frac{f(y,1,x)f(x)}{f(1,x)}\left(\frac{\mathbb{I}_{\tilde{y},\tilde{d},\tilde{x}}(y,1,x)}{f(y,1,x)}+\frac{\mathbb{I}_{\tilde{x}}(x)}{f(x)}-\frac{\mathbb{I}_{\tilde{d},\tilde{x}}(1,x)}{f(1,x)}-1\right)dydx \\
=&\frac{\mathbb{I}_{\tilde{d}}(1)}{e_0(\tilde{x})}(\tilde{y}-\mu_0(1,\tilde{x}))+\mu_0(1,\tilde{x})-\mathbb{E}[Y_i(1)].
\end{aligned}
$$

Similarly, we can replace $d=1$ with $d=0$ and $e_0(x)$ with $1-e_0(x)$ to obtain the IF of $\mathbb{E}[Y_i(0)]$. The IF of $\mathbb{E}[Y_i(1)-Y_i(0)]$ is then the difference of the two IFs, which is $\psi(O_i)-\mathbb{E}[\tau(X_i)]$. Then we have

$$
\begin{aligned}
IF_\beta(O_i) &= \frac{d\beta(P_t)}{dt}\Big|_{t=0} \\
&= \frac{d}{dt}\left(\mathbb{E}_{P_t}[b(X_i)'b(X_i)]^{-1}\mathbb{E}_{P_t}[b(X_i)'\psi(O_i)]\right)\Big|_{t=0} \\
&= \frac{d}{dt}\left(\mathbb{E}_{P_t}[b(X_i)'b(X_i)]^{-1}\right)\Big|_{t=0}\mathbb{E}_{P_t}[b(X_i)'\psi(O_i)]+\mathbb{E}_{P_t}[b(X_i)'b(X_i)]^{-1}\frac{d}{dt}\mathbb{E}_{P_t}[b(X_i)'\psi(O_i)]\Big|_{t=0} \\
&= \mathbb{E}[b(X_i)'b(X_i)]^{-1}\left(b(X_i)'\psi(O_i)-\mathbb{E}[b(X_i)'\psi(O_i)]\right) \\
&\qquad -\mathbb{E}[b(X_i)'b(X_i)]^{-1}\left(b(X_i)'b(X_i)-\mathbb{E}[b(X_i)'b(X_i)]\right)\mathbb{E}[b(X_i)b(X_i)']^{-1}\mathbb{E}[b(X_i)'\psi(O_i)] \\
&= \mathbb{E}[b(X_i)'b(X_i)]^{-1}b(X_i)'\left(\psi(O_i)-b(X_i)\beta_0\right)
\end{aligned}
$$

where the fourth equation follows from the chain rule and the fact that the IF of $\mathbb{E}[b(X_i)'\psi(O_i)]$ is $b(X_i)'\psi(O_i)-\mathbb{E}[b(X_i)'\psi(O_i)]$. □

Proof of Proposition 4.

$$
Var(\tau(X_i)) = Var(\mathbb{E}[Y_i(1)-Y_i(0)|X_i]) = Var(\mu_0(1,X_i)-\mu_0(0,X_i))
$$

where the second equation holds under Assumption 3 and 4. □

Proof of Proposition 5. Let $\mu_t(d,x) = \mathbb{E}_{P_t}[Y_i|D_i = d, X_i = x]$ be the conditional expectation function evaluated at P_t. Accordingly, $\nu_t(x) = \mu_t(1,x) - \mu_t(0,x)$, and $e_t(x) = \mathbb{E}_{P_t}[D_i|X_i = x]$. Then

$$\begin{aligned}
IF_\theta(O_i) &= \frac{d}{dt}\mathbb{E}_{P_t}[(\nu_t(X_i) - \mathbb{E}_{P_t}[\nu_t(X_i)])^2]\big|_{t=0} \\
&= \mathbb{E}\left[2(\nu_t(X_i) - \mathbb{E}[\nu_t(X_i)])\frac{d}{dt}(\nu_t(X_i) - \mathbb{E}_{P_t}[\nu_t(X_i)])\big|_{t=0}\right] \\
&\quad + \frac{d}{dt}\mathbb{E}_{P_t}[(\nu_0(X_i) - \mathbb{E}[\nu_0(X_i)])^2]\big|_{t=0} \\
&= \mathbb{E}\left[2(\nu_t(X_i) - \mathbb{E}[\nu_t(X_i)])\mathbb{E}\left[\frac{d}{dt}(\nu_t(X_i))\right]\Bigg|_{t=0}\right] \\
&\quad + (\nu_0(X_i) - \mathbb{E}[\nu_0(X_i)])^2 - \theta_0
\end{aligned}$$

where the last equation follows from the influence function of expectations, i.e., $\frac{d}{dt}\mathbb{E}_{P_t}[A_i]|_{t=0} = A_i - \mathbb{E}[A_i]$ for any random variable A_i. Now we continue with the first term, which can be written as an integration form

$$\frac{d}{dt}\int\int 2(\nu_0(x) - \mathbb{E}[\nu_0(X_i)]) \cdot \left(y\frac{f_t(y,1,x)f_t(x)}{f_t(1,x)} - y\frac{f_t(y,0,x)f_t(x)}{f_t(0,x)}\right) dydx\Bigg|_{t=0}.$$

Note that this term is similar to Equation (A.2). Following the same steps, we have

$$2(\nu_0(\tilde{x}) - \mathbb{E}[\nu_0(X_i)])\left(\frac{\mathbb{I}_{\tilde{d}}(1)}{e_0(\tilde{x})}(\tilde{y} - \mu_0(1,\tilde{x})) - \frac{\mathbb{I}_{\tilde{d}}(0)}{1 - e_0(\tilde{x})}(\tilde{y} - \mu_0(0,\tilde{x}))\right).$$

Replacing variables with tildes with generic forms,

$$\begin{aligned}
&2(\nu_0(X_i) - \mathbb{E}[\nu_0(X_i)])\left(\frac{D_i}{e_0(X_i)}(Y_i - \mu_0(1,X_i)) - \frac{(1-D_i)}{1-e_0(X_i)}(Y_i - \mu_0(0,X_i))\right) \\
&= 2(\nu_0(X_i) - \mathbb{E}[\nu_0(X_i)])(\psi(O_i) - \nu_0(X_i))
\end{aligned}$$

The IF of θ_0 is then given by

$$\begin{aligned}
IF_\theta(O_i) &= 2(\nu_0(X_i) - \mathbb{E}[\nu_0(X_i)])(\psi(O_i) - \nu_0(X_i)) + (\nu_0(X_i) - \mathbb{E}[\nu_0(X_i)])^2 - \theta_0 \\
&= (\psi(O_i) - \tau_0)^2 - (\psi(O_i) - \nu_0(X_i))^2 - \theta_0
\end{aligned}$$

□

Proof of Proposition 7. Under Assumption 3 and 4, $\mathbb{E}[Y_i(d)^2] = \mathbb{E}[\phi^d(O_i)]$, which follows the same identification procedure as $\mathbb{E}[Y_i(d)]$. Therefore,

$$\dot{\lambda}_0 = \left(\mathbb{E}[\phi^1(O_i)] - \mathbb{E}[\psi^1(O_i)]^2\right) - \left(\mathbb{E}[\phi^0(O_i)] - \mathbb{E}[\psi^0(O_i)]^2\right).$$

□

Proof of Proposition 8. Let $\bar{\phi}^d = \mathbb{E}[\phi^d(O_i)]$ and $\bar{\psi}^d = \mathbb{E}[\psi^d(O_i)]$. In this proof, we use the IFs of $\bar{\psi}^d$ and $\bar{\phi}^d$ as building blocks and derive the IF of $\dot{\lambda}_0$ by applying the chain rule of influence function. First note that the IF of $\bar{\psi}^d$ is $\psi^d(O_i) - \bar{\psi}^d$, which has been proved in the literature of augmented inverse propensity weighting estimators. Following a similar procedure, it is easy to show that the IF of $\bar{\phi}^d$ is $\phi^d(O_i) - \bar{\phi}^d$.

Consider the mapping $g(\bar{\psi}^1, \bar{\phi}^1) = \bar{\phi}^1 - \bar{\psi}^{1,2}$. We have

$$\begin{aligned} IF_g(\bar{\psi}^1, \bar{\phi}^1) &= \nabla g \cdot (IF_{\bar{\psi}^1}(O_i), IF_{\bar{\phi}^1}(O_i))' \\ &= \phi^1(O_i) - \bar{\phi}^1 - 2\bar{\psi}^1(\psi^1(O_i) - \bar{\psi}^1) \end{aligned}$$

where the first equation follows from the chain rule of influence function. The second equation holds by plugging in $\nabla g = (-2\bar{\psi}^1, 1)$ and the IFs of $\bar{\psi}^1$ and $\bar{\phi}^1$. Similarly, we can obtain the IF when $d = 0$. Combining the results, we have

$$\begin{aligned} IF_\lambda(O_i) = \phi^1(O_i) - \mathbb{E}[\phi^1(O_i)] - 2\mathbb{E}[\psi^1(O_i)](\psi^1(O_i) - \mathbb{E}[\psi^1(O_i)]) \\ - \left(\phi^0(O_i) - \mathbb{E}[\phi^0(O_i)] - 2\mathbb{E}[\psi^0(O_i)](\psi^0(O_i) - \mathbb{E}[\psi^0(O_i)])\right) \end{aligned}$$

□

Proof of Propositions 3, 6, and 9. Here we provide a general proof for the asymptotic properties of $\hat{\beta}$, $\hat{\theta}$, and $\hat{\lambda}$ using a common empirical process notation. Let P_0 and P_n denote linear operators such that for some function $f(O_i)$, $P_0(f) = \mathbb{E}[f(O_i)] = \int f(O_i) dP_0$ and $P_n(f) = P_n(f(O_i)) = n^{-1}\sum_{i=1}^n f(O_i)$. Let $\hat{P}$ denote the plug-in estimator, e.g., $\psi(O_i, \hat{P}) = \hat{\psi}(O_i)$, $\nu(X_i, \hat{P}) = \hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)$.

For arbitrary statistical functional $\varphi(P_0)$ with estimating equation estimator $\hat{\varphi}$, we

have

$$\begin{aligned}\hat{\varphi} - \varphi_0 &= \varphi(\hat{P}) - \varphi(P_0) + P_n(IF_\varphi(O_i, \hat{P})) \\ &= (P_n - P_0)IF_\varphi(O_i, \hat{P}) + R_n \\ &= (P_n - P_0)IF_\varphi(O_i, P_0) + (P_n - P_0)(IF_\varphi(O_i, \hat{P}) - IF_\varphi(O_i, P_0)) + R_n \\ &= (P_n - P_0)IF_\varphi(O_i, P_0) + E_n + R_n\end{aligned}$$

where

$$\begin{aligned}E_n &= (P_n - P_0)(IF_\varphi(O_i, \hat{P}) - IF_\varphi(O_i, P_0)) \\ R_n &= \varphi(\hat{P}) - \varphi(P_0) + P_0(IF_\varphi(O_i, \hat{P}))\end{aligned}$$

The first term $(P_n - P_0)IF_\varphi(O_i, P_0)$ is a sample average of a fixed function, which behaves like a normally distributed random variable by central limit theorem, up to an error $o_P(n^{-1/2})$. The task is to show that the empirical process term E_n and the remainder term R_n are $o_P(n^{-1/2})$.

The empirical process term E_n:

The task is to prove

$$\|IF_\varphi(O_i, \hat{P}) - IF_\varphi(O_i, P_0)\|_P = o_P(1). \tag{A.3}$$

Then under Donsker condition and Lemma 19.24 of Van der Vaart (1998), or under cross-fitting condition and Lemma 2 of Kennedy et al. (2020), the above condition implies that E_n is $o_P(n^{-1/2})$.

For CPT:

$$
\begin{aligned}
E_n &= (P_n - P_0)(IF_\beta(O_i, \hat{P}) - IF_\beta(O_i, P_0)) \\
&= \mathbb{E}[b(X_i)'b(X_i)]^{-1}(P_n - P_0)(b(X_i)'b(X_i)(\beta(\hat{P}) - \beta(P_0))) && \text{(A.4)} \\
&\quad + \mathbb{E}[b(X_i)'b(X_i)]^{-1}(P_n - P_0)\Bigg(b(X_i)'[\\
&\quad + \left(\left(1 - \frac{D_i}{e_0(X_i)}\right)(\hat{\mu}(1, X_i) - \mu_0(1, X_i))\right) && \text{(A.5)} \\
&\quad + \frac{D_i(Y_i - \hat{\mu}(1, X_i))}{\hat{e}(X_i)e_0(X_i)}(e_0(X_i) - \hat{e}(X_i)) && \text{(A.6)} \\
&\quad - \left(\left(1 - \frac{1 - D_i}{1 - e_0(X_i)}\right)(\hat{\mu}(0, X_i) - \mu_0(0, X_i))\right) && \text{(A.7)} \\
&\quad - \frac{(1 - D_i)(Y_i - \hat{\mu}(0, X_i))}{(1 - \hat{e}(X_i))(1 - e_0(X_i))}(\hat{e}(X_i) - e_0(X_i))]\Bigg) && \text{(A.8)}
\end{aligned}
$$

The first term (A.4) is equal to $(\beta(\hat{P})-\beta(P_0))\mathbb{E}[b(X_i)'b(X_i)]^{-1}(P_n-P_0)(b(X_i)'b(X_i)) = o_P(n^{-1/2})$ by central limit theorem. Under the assumptions in Proposition 3, we also have the following conditions:

$$
\begin{aligned}
\|\text{(A.5)}\|_P &\leq \left(1 + \frac{1}{\xi}\right)\|\mu_0(1, X_i) - \hat{\mu}(1, X_i)\|_P = o_P(1) \\
\|\text{(A.6)}\|_P &\leq \frac{\sqrt{K}}{\xi^2}\|e_0(X_i) - \hat{e}(X_i)\|_P = o_P(1) \\
\|\text{(A.7)}\|_P &\leq \left(1 + \frac{1}{\xi}\right)\|\mu_0(0, X_i) - \hat{\mu}(0, X_i)\|_P = o_P(1) \\
\|\text{(A.8)}\|_P &\leq \frac{\sqrt{K}}{\xi^2}\|e_0(X_i) - \hat{e}(X_i)\|_P = o_P(1).
\end{aligned}
$$

Combined with the condition that $\|b(X_i)\|$ is bounded, the above conditions imply that Equation (A.3) holds, and thus $E_n = o_P(n^{-1/2})$.

For CVT:

$$
\begin{aligned}
E_n &= (P_n - P_0)(IF_\theta(O_i, \hat{P}) - IF_\theta(O_i, P_0)) \\
&= (P_n - P_0)(\theta_0 - \theta(\hat{P})) && \text{(A.9)} \\
&\quad + 2(P_n - P_0)((\hat{\psi} - \psi)(\hat{\nu} - \hat{\tau})) && \text{(A.10)} \\
&\quad + (P_n - P_0)((\psi - \hat{\tau})^2 - (\psi - \tau_0)^2) && \text{(A.11)} \\
&\quad - (P_n - P_0)((\psi - \hat{\nu})^2 - (\psi - \nu_0)^2) && \text{(A.12)}
\end{aligned}
$$

Here, for simplicity, we slightly abuse the notation to use $\hat{\tau}$ to denote $\tau(\hat{P})$, instead of the estimating equation estimator of τ_0. Note that the first term (A.9) is zero since $(P_n - P_0)(\theta_0 - \theta(\hat{P})) = (\theta_0 - \theta(\hat{P}))(P_n - P_0)[1] = 0$. Consider the quantity $\mathbb{E}[(\text{A.10})^2] = \mathbb{E}[(\hat{\psi} - \psi)^2(\hat{\nu} - \hat{\tau})^2] < K\mathbb{E}[(\hat{\psi} - \psi)^2]$. The second term $\mathbb{E}[(\hat{\psi} - \psi)^2]$ also appears in the derivation of ATE estimators in e.g., Theorem 5.1 of Chernozhukov et al. (2018), which is $o_P(1)$ similar to Equation (A.5)-(A.8), given that $\hat{e} \in (\xi, 1-\xi)$, $\mathbb{E}[(Y_i - \hat{\mu})^2|X_i] < K$ and $\|e_0 - \hat{e}\|\|\mu_0 - \hat{\mu}\| = o_P(n^{-1/2})$. For the fourth term (A.12), we have

$$
\begin{aligned}
&\mathbb{E}[((\psi - \hat{\nu})^2 - (\psi - \nu_0)^2)^2] \\
&= \mathbb{E}[\mathbb{E}[(2(\psi - \nu_0)(\nu_0 - \hat{\nu}) + (\nu_0 - \hat{\nu})^2)^2|X_i]] \\
&= \mathbb{E}[4Var(\psi|X_i)(\nu_0 - \hat{\nu})^2 + (\nu_0 - \hat{\nu})^4] \\
&< 4K\mathbb{E}[(\nu_0 - \hat{\nu})^2] + \mathbb{E}[(\nu_0 - \hat{\nu})^4]
\end{aligned}
$$

which is $o_P(1)$ given that $\|\nu_0 - \hat{\nu}\| = o_P(n^{-1/4})$. Similarly, the third term (A.11) is also $o_P(1)$. The above conditions imply that Equation (A.3) holds, and thus $E_n = o_P(n^{-1/2})$.

For POVT:

$$
\begin{aligned}
E_n &= (P_n - P_0)(IF_\lambda(O_i, \hat{P}) - IF_\lambda(O_i, P_0)) \\
&= (P_n - P_0)\left(\hat{\phi}^1 - \phi_0^1 - 2P_0(\hat{\psi}^1)\hat{\psi}^1 + 2P_0(\psi_0^1)\psi_0^1\right) && \text{(A.13)} \\
&\quad - P_0(\hat{\phi}^1 - \phi_0^1) + 2(P_0(\hat{\psi}^1) - P_0(\psi_0^1))(P_0(\hat{\psi}^1) + P_0(\psi_0^1)) && \text{(A.14)} \\
&\quad - (P_n - P_0)\left(\hat{\phi}^0 - \phi_0^0 - 2P_0(\hat{\psi}^0)\hat{\psi}^0 + 2P_0(\psi_0^0)\psi_0^0\right) && \text{(A.15)} \\
&\quad + P_0(\hat{\phi}^0 - \phi_0^0) - 2(P_0(\hat{\psi}^0) - P_0(\psi_0^0))(P_0(\hat{\psi}^0) + P_0(\psi_0^0)). && \text{(A.16)}
\end{aligned}
$$

First, note that the quantity $\|\hat{\phi}^1 - \phi_0^1\|_P$ is similar to (A.4) and (A.5) in the E_n of CPT, with ψ^1 replaced by ϕ^1. Following the same derivation as (A.4) and (A.5), it is straightforward to show that $\|\hat{\phi}^1 - \phi_0^1\|_P = o_P(1)$. For the quantity $\|P_0(\hat{\psi}^1)\hat{\psi}^1 - P_0(\psi_0^1)\psi_0^1\|_P$, we can write it as

$$
\begin{aligned}
\|P_0(\hat{\psi}^1)\hat{\psi}^1 - P_0(\psi_0^1)\psi_0^1\|_P &= \|P_0(\hat{\psi}^1 - \psi_0^1)\hat{\psi}^1 + P_0(\psi_0^1)(\hat{\psi}^1 - \psi_0^1)\|_P \\
&\leq \|P_0(\hat{\psi}^1 - \psi_0^1)\hat{\psi}^1\|_P + \|P_0(\psi_0^1)(\hat{\psi}^1 - \psi_0^1)\|_P \\
&\leq \|\hat{\psi}^1\|_P |P_0(\hat{\psi}^1 - \psi_0^1)| + |P_0(\psi_0^1)| \|\hat{\psi}^1 - \psi_0^1\|_P \\
&\leq \|\hat{\psi}^1\|_P \|\hat{\psi}^1 - \psi_0^1\|_P + |P_0(\psi_0^1)| \|\hat{\psi}^1 - \psi_0^1\|_P \\
&= o_P(1)
\end{aligned}
$$

where the first inequality follows from triangle inequality, the last inequality follows from Jensen's and Cauchy-Schwarz inequalities. The last equality holds as long as $\|\hat{\psi}^1\|_P$ is $O_P(1)$ and $\|\hat{\psi}^1 - \psi_0^1\|_P$ is $o_P(1)$, and the latter condition is already proved by (A.4) and (A.5).

The above conditions imply that the term (A.13) is $o_P(n^{-1/2})$. Next, we prove that the term (A.14) is $o_P(n^{-1/2})$.

The first term in (A.14) is $o_P(n^{-1/2})$ given that $\|\hat{\psi}^1 - \psi_0^1\|_P = o_P(n^{-1/2})$. For the

second term, we have

$$
\begin{aligned}
P_0(\hat{\phi}^1 - \phi_0^1) &= P_0\left(\frac{D_i}{\hat{e}(X_i)}(Y_i^2 - \hat{\mu}^2(1,X_i)) + \hat{\mu}^2(1,X_i) - \mu_0^2(1,X_i)\right) \\
&= P_0\left(\frac{D_i}{\hat{e}(X_i)}(Y_i^2 - \mu_0^2(1,X_i) + \mu_0^2(1,X_i) - \hat{\mu}^2(1,X_i)) + \hat{\mu}^2(1,X_i) - \mu_0^2(1,X_i)\right) \\
&= P_0\left(\left(\frac{D_i}{\hat{e}(X_i)} - 1\right)(\mu_0^2(1,X_i) - \hat{\mu}^2(1,X_i))\right) \\
&= P_0\left(\left(\frac{e_0(X_i)}{\hat{e}(X_i)} - 1\right)(\mu_0^2(1,X_i) - \hat{\mu}^2(1,X_i))\right) \\
&\leq \frac{1}{\xi} P_0((e_0(X_i) - \hat{e}(X_i))(\mu_0^2(1,X_i) - \hat{\mu}^2(1,X_i))) \\
&\leq \frac{1}{\xi}\|e_0 - \hat{e}\|_P \|\mu_0^2 - \hat{\mu}^2\|_P \\
&= o_P(n^{-1/2})
\end{aligned}
\tag{A.17}
$$

where the second equality follows from the fact that $D_i(Y_i^2 - \mu_0^2(1,X_i)) = 0$, third equality follows from the law of iterated expectations, the last inequality follows from the Cauchy-Schwarz inequality.

For the second term in (A.14), $P_0(\hat{\psi}^1 - \psi_0^1) = o_P(n^{-1/2})$ is similar to $P_0(\hat{\phi}^1 - \phi_0^1)$ with Y_i^2 replaced by Y_i, and μ_0^2 replaced by μ_0. Then the second term is $o_P(n^{-1/2})$ as long as $P_0(\hat{\psi}^1)$ is $O_P(1)$.

Similar bounds holds for $d = 0$, so (A.15) and (A.16) are also $o_P(n^{-1/2})$. All terms in E_n are $o_P(n^{-1/2})$, and thus $E_n = o_P(n^{-1/2})$.

The remainder term R_n:

For CPT:

$$
\begin{aligned}
R_n &= \beta(\hat{P}) - \beta(P_0) + P_0(IF_\beta(O_i, \hat{P})) \\
&= P_0(IF_\beta(O_i, \hat{P}) - \beta(P_0)) \\
&= \mathbb{E}[b(X_i)'b(X_i)]^{-1} P_0 \Bigg(b(X_i)' \left(\frac{e_0(X_i)}{\hat{e}(X_i)} - 1 \right) (\mu_0(1, X_i) - \hat{\mu}(1, X_i)) \\
&\quad - b(X_i)' \left(\frac{1 - e_0(X_i)}{1 - \hat{e}(X_i)} - 1 \right) (\mu_0(0, X_i) - \hat{\mu}(0, X_i)) \Bigg).
\end{aligned}
$$

The derivation is similar to Equation (A.17). Consider the quantity $\mathbb{E}[b(X_i)'b(X_i)]R_n$. If this quantity is $o_P(n^{-1/2})$, then R_n is $o_P(n^{-1/2})$ since $\mathbb{E}[b(X_i)'b(X_i)]$ is positive definite.

$$
\begin{aligned}
\|\mathbb{E}[b(X_i)'b(X_i)]R_n\| &\leq P_0 \left(\left\| b(X_i)' \left(\frac{e_0(X_i)}{\hat{e}(X_i)} - 1 \right) (\mu_0(1, X_i) - \hat{\mu}(1, X_i)) \right\| \right) \\
&\quad + P_0 \left(\left\| b(X_i)' \left(\frac{1 - e_0(X_i)}{1 - \hat{e}(X_i)} - 1 \right) (\mu_0(0, X_i) - \hat{\mu}(0, X_i))' \right\| \right) \\
&= P_0 \left(\|b(X_i)'\| \left| \left(\frac{e_0(X_i)}{\hat{e}(X_i)} - 1 \right) (\mu_0(1, X_i) - \hat{\mu}(1, X_i)) \right| \right) \\
&\quad + P_0 \left(\|b(X_i)'\| \left| \left(\frac{1 - e_0(X_i)}{1 - \hat{e}(X_i)} - 1 \right) (\mu_0(0, X_i) - \hat{\mu}(0, X_i)) \right| \right) \\
&\leq K P_0 \left(\left| \left(\frac{e_0(X_i)}{\hat{e}(X_i)} - 1 \right) (\mu_0(1, X_i) - \hat{\mu}(1, X_i)) \right| \right) \\
&\quad + K P_0 \left(\left| \left(\frac{1 - e_0(X_i)}{1 - \hat{e}(X_i)} - 1 \right) (\mu_0(0, X_i) - \hat{\mu}(0, X_i)) \right| \right) \\
&\leq \frac{K}{\xi} \|e_0 - \hat{e}\|_P \|\mu_0 - \hat{\mu}\|_P + \frac{K}{\xi} \|\hat{e} - e_0\|_P \|\mu_0 - \hat{\mu}\|_P \\
&= o_P(n^{-1/2})
\end{aligned}
$$

where the first inequality follows from triangle and Jensen's inequality. Then we have the desired result $R_n = o_P(n^{-1/2})$.

For CVT:

$$
\begin{aligned}
R_n &= \theta(\hat{P}) - \theta(P_0) + P_0(IF_\theta(O_i, \hat{P})) \\
&= P_0(\theta(\hat{P}) - \theta(P_0) + (\psi(O_i, \hat{P}) - \tau(\hat{P}))^2 - (\psi(O_i, \hat{P}) - \nu(X_i, \hat{P}))^2 - \theta(\hat{P})) \\
&= P_0((\psi(O_i, \hat{P}) - \tau(\hat{P}))^2 - (\psi(O_i, \hat{P}) - \nu(X_i, \hat{P}))^2 - (\nu(X_i, P_0) - \tau(P_0))) \\
&= P_0((\nu(X_i, \hat{P}) - \tau(\hat{P}))^2 - (\nu(X_i, P_0) - \tau(P_0))^2 + 2(\psi(O_i, \hat{P}) - \nu(X_i, \hat{P}))(\nu(X_i, \hat{P}) - \tau(\hat{P}))) \\
&= P_0((\tau(P_0) - \tau(\hat{P}))^2 - (\nu(X_i, P_0) - \nu(X_i, \hat{P}))^2 + 2(\nu(X_i, \hat{P}) - \nu(X_i, P_0))(\nu(X_i, \hat{P}) - \tau(\hat{P})) \\
&\quad + 2(\psi(O_i, \hat{P}) - \nu(X_i, \hat{P}))(\nu(X_i, \hat{P}) - \tau(\hat{P}))) \\
&= P_0((\tau(P_0) - \tau(\hat{P}))^2 - (\nu(X_i, P_0) - \nu(X_i, \hat{P}))^2 + 2(\psi(O_i, \hat{P}) - \nu(X_i, P_0))(\nu(X_i, \hat{P}) - \tau(\hat{P}))).
\end{aligned}
$$

The first term $P_0((\tau(P_0) - \tau(\hat{P}))^2 - (\nu(X_i, P_0) - \nu(X_i, \hat{P}))^2)$ is $o_P(n^{-1/2})$ given that both $\|\tau_0 - \tau(\hat{P})\|_P$ and $\|\nu_0 - \hat{\nu}\|_P$ are $o_P(n^{-1/4})$. For the second term, consider the quantity

$$
\begin{aligned}
& P_0((\psi(O_i, \hat{P}) - \nu(X_i, P_0))^2(\nu(X_i, \hat{P}) - \tau(\hat{P}))^2) \\
\leq\ & P_0((\psi(O_i, \hat{P}) - \nu(X_i, P_0))^2) P_0((\nu(X_i, \hat{P}) - \tau(\hat{P}))^2) \\
\leq\ & K P_0((\psi(O_i, \hat{P}) - \nu(X_i, P_0))^2)
\end{aligned}
$$

where the first inequality follows from Cauchy-Schwarz inequality, and the second inequality follows from the condition that $(\hat{\nu}(X_i) - \tau(\hat{P}))^2 < K$. The remaining term is similar to the R_n of CPT. Specifically,

$$
\begin{aligned}
(\psi(O_i, \hat{P}) - \nu(X_i, P_0))^2 = & \left(\left(\frac{e_0(X_i)}{\hat{e}(X_i)} - 1 \right) (\mu_0(1, X_i) - \hat{\mu}(1, X_i)) \right. \\
& \left. - \left(\frac{1 - e_0(X_i)}{1 - \hat{e}(X_i)} - 1 \right) (\mu_0(0, X_i) - \hat{\mu}(0, X_i)) \right)^2 \\
\leq\ & 2 \left(\frac{e_0(X_i)}{\hat{e}(X_i)} - 1 \right)^2 (\mu_0(1, X_i) - \hat{\mu}(1, X_i))^2 \\
& + 2 \left(\frac{1 - e_0(X_i)}{1 - \hat{e}(X_i)} - 1 \right)^2 (\mu_0(0, X_i) - \hat{\mu}(0, X_i))^2 \\
\leq\ & \frac{2}{\xi^2} (e_0(X_i) - \hat{e}(X_i))^2 ((\mu_0(1, X_i) - \hat{\mu}(1, X_i))^2) \\
& + \frac{2}{\xi^2} (e_0(X_i) - \hat{e}(X_i))^2 ((\mu_0(0, X_i) - \hat{\mu}(0, X_i))^2).
\end{aligned}
$$

Therefore, $P_0((\psi(O_i, \hat{P}) - \nu(X_i, P_0))^2) = o_P(n^{-1})$ by Cauchy-Schwarz inequality,

which implies $R_n = o_P(n^{-1/2})$.

For POVT:

$$R_n = \lambda(\hat{P}) - \lambda(P_0) + P_0(IF_\lambda(O_i, \hat{P}))$$

For $d = 1$, the remainder term is

$$R_n^1 = P_0(\hat{\phi}^1) - P_0(\phi^1) - (P_0(\hat{\psi}^1))^2 + (P_0(\psi^1))^2 \tag{A.18}$$
$$+ P_0(\hat{\phi}^1 - P_0(\hat{\phi}^1) - 2P_0(\hat{\psi}^1)(\hat{\psi}^1 - P_0(\hat{\psi}^1))) \tag{A.19}$$

Note that (A.19) is zero. The term

$$(\text{A.18}) = P_0(\hat{\phi}^1 - \phi^1) - (P_0(\hat{\psi}^1) + P_0(\psi^1))P_0(\hat{\psi}^1 - \psi^1) = o_P(n^{-1/2})$$

with the bounds of all terms already proved in previous derivations. Similar bounds holds for $d = 0$, and thus $R_n = o_P(n^{-1/2})$.

All proofs are complete.

□