

Consistent Tests for Conditional Treatment Effects

Yu-Chin Hsu*

Department of Economics
University of Missouri at Columbia

First version: May 23, 2011

This version: March 15, 2012

* Department of Economics, University of Missouri at Columbia, Columbia, MO 65211 U.S.A.; hsuy@missouri.edu.

Acknowledgements: I thank Jason Abrevaya, Richard Chiburis, Stephen G. Donald, Yingyao Hu, Cory Koedel, Robert P. Lieli, J. Isaac Miller, Peter Mueser, Xiaoxia Shi, Kyungchul Song and seminar participants at 20th Annual Meetings of the Midwest Econometrics Group for their insightful comments. All errors and omissions are my own responsibility.

Abstract

We construct a Kolmogorov-Smirnov test for the null hypothesis that the average treatment effect is non-negative conditional on every possible value of the covariates. Our test can be more informative than the traditional average treatment effect on the whole population. The null hypothesis can be characterized as a conditional moment inequality under the unconfoundedness assumption, and we employ the instrumental variable method to convert the conditional moment inequality into an infinite number of unconditional moment inequalities without information loss. A Kolmogorov-Smirnov test is constructed based on these unconditional moment inequalities. It is shown that our test can control the size uniformly over a broad set of data generating processes asymptotically, is consistent against fixed alternatives and is unbiased against some $N^{-1/2}$ local alternatives. Our test is also more powerful than Lee and Whang's (2009) against a broad set of $N^{-1/2}$ local alternatives when the local alternatives do not converge to the least favorable case. Monte-Carlo simulation results confirm our theoretical findings. Several interesting extensions of our test are discussed.

JEL classification: C01, C12, C21

Keywords: Hypothesis testing, treatment effects, test consistency, propensity score.

1 Introduction

This paper proposes a Kolmogorov-Smirnov (KS) test for the null hypothesis that the conditional average treatment effect (CATE) given a covariate value is nonnegative for every value in the support of the covariates. Our test can be more informative than the traditional average treatment effect on the whole population (ATE). For example, when the ATE is positive, it is possible that the CATE is negative for a subset of covariate values. In this case, a policy or a treatment that is beneficial on average for the whole population could make some people with certain characteristics worse off (on average). Therefore, our test can provide more information than the ATE to the policy maker. An acceptance of the null hypothesis implies that for all subpopulations defined by the covariate values, the expected welfare over the subpopulation will not decrease under the treatment. Also, an acceptance of our test implies that the ATE is non-negative. In addition, our test directly applies to cases where one is interested in a subpopulation defined by a subset of the support of the covariate values. This is useful when the policy maker is only interested in a specific subpopulation.

The null hypothesis can be characterized as a conditional moment inequality under the unconfoundedness assumption. We employ Andrews and Shi's (2011; AS hereafter) instrumental variable approach to transform the conditional moment inequality into an infinite number of unconditional moment inequalities without information loss. An inverse probability weighted estimator (IPW) similar to Hirano, Imbens and Ridder (2003; HIR hereafter) is used to estimate each of the unconditional moments. The KS test statistic is defined as the supremum of the estimated unconditional moments indexed by the instrument functions. A critical value is constructed based on a simulated process and the generalized moment selection (GMS) approach. The GMS method introduced by Andrews and Soares (2010) and AS is similar to the recentering method of Hansen (2005) and Donald and Hsu (2010), as well as the contact set method of Linton, Song and Whang (2010). These methods are used to obtain tests with better power than the ones using the least favorable configuration (LFC). We show that our test can control the size uniformly over a broad set of data generating processes (DGP) asymptotically, is consistent against any fixed alternatives and is unbiased against some $N^{-1/2}$ local alternatives.

Note that our problem is different from AS in that a preliminary consistent estimator for propensity score is needed in our case. Section 8 of AS provide high-level conditions for a test to have uniform size control when a preliminary consistent estimator is needed, but these conditions are difficult to verify. We contribute to the literature by providing

low-level sufficient conditions that are easier to verify in practice. It is important for a test involving inequality constraints to have uniform size control because pointwise asymptotics fail to capture the finite-sample properties of the test statistic due to the discontinuity in the limiting distribution of the test statistic, please see Andrews and Guggenberger (2009) and AS for more details.

Our test is more powerful than Lee and Whang’s (2009; LW hereafter) against a broad set of $N^{-1/2}$ local alternatives that do not converge to the least favorable case. LW’s test is the only alternative to our test for conditional treatment effects and their test statistic is a one-sided L_1 -type functional of the nonparametric kernel estimator of the conditional average treatment effects.¹ The local power advantage of our test suggests that our test is more likely to detect the violation of the null hypotheses in finite sample. We conduct Monte-Carlo simulations to study the finite sample performance of our test and LW’s and the results support our theoretical findings.

This paper is related to the treatment effect literature. For recent reviews of this literature, see Imbens (2004) and Imbens and Wooldridge (2009), among others. Most papers in the literature focus on the estimation and inference for the average treatment effects, and only a few papers construct tests for the average treatment effect conditional on covariates, e.g., LW and Crump, Hotz, Imbens and Mintnik (2008). We have discussed LW’s test. Crump, Hotz, Imbens and Mintnik (2008) construct nonparametric tests for two different null hypotheses: (1) that the average treatment effects conditional on the covariates for all values of covariates are equal to zero, and (2) that they are equal to a constant.² The null hypotheses in their study involve conditional moment equalities, which are different from ours. The methods developed in this paper extend directly to those null hypotheses, but a detailed comparison between our method and theirs is beyond the scope of the present study.

This paper is also related to the literature on conditional moment inequalities, e.g.,

¹Delgado and Escanciano’s (2011) and Lee, Song and Whang’s (2011) methods may be used as well. However, Delgado and Escanciano (2011) require the propensity score function to depend on a finite number of parameters, but LW and our paper allow for the form of the propensity score function to be completely unspecified. Also, if there are more than one continuous covariates, they can only test for the necessary conditions on the null hypothesis and their test may not be consistent in this case. On the other hand, Lee, Song and Whang (2011) require that the treatment assignment be random and independent of the covariates with the probability of assignment known. However, we only require that the treatment assignment is unconfounded, and allow for the treatment assignment to be dependent of the potential outcomes and the covariates, and the probability of assignment to be unknown. Please refer to their papers for more details.

²LW also consider the first null hypothesis.

Chernozhukov, Lee and Rosen (2008), Galichon and Henry (2009), Fan (2008), Kim (2008), Armstrong (2011a, 2011b) and AS. These papers construct confidence sets for the parameters defined by a set of conditional moment inequalities and/or equalities. The focus of our paper differs from these studies in that we are interested in testing a null hypothesis that is characterized by a moment inequality.

We consider several extensions of our tests. We first extend our results to test the null hypothesis that the conditional stochastic dominance relation between the potential outcomes holds for all values of the covariates, which is also considered by LW. The importance of the distributional treatment effect has been pointed out by Imbens and Wooldridge (2009). In the treatment effect literature, only a small number of papers discuss the stochastic dominance relation between two groups under the unconfoundedness assumption such as Donald and Hsu (2011) and Maier (2011). Furthermore, Abadie (2002) considers tests for the stochastic dominance relation when the treatment assignment is endogenous. These papers focus on unconditional stochastic dominance relation between the potential outcomes, but we focus on the conditional stochastic dominance relation. Second, we extend our results to cases where the conditioning set is a strict subset of the covariates and the unconfoundedness assumption does not hold if we condition on this subset of covariates. The only paper in the literature that discusses the conditional average treatment effects in this case is Abrevaya, Hsu and Lieli (2011), who propose a two-step kernel estimator for the conditional average treatment effect, but here we are interested in testing whether the conditional average treatment effect is uniformly non-negative conditional on this subset of covariates. Finally, we extend our test to cases where the treatment assignment is endogenous, as in the local average treatment effect setup of Imbens and Angrist (1994), Abadie, Angrist and Imbens (2002), Abadie (2002, 2003), Frölich (2007) and Donald, Hsu and Lieli (2011).

The rest of this paper is organized as follows. In Section 2, we introduce the model setup and show that the null hypothesis can be characterized as a conditional moment inequality. We employ AS's instrumental variable approach to transform the conditional moment inequality to a continuum of unconditional moment inequalities without information loss. We introduce an IPW estimator to estimate the unconditional moments. The test statistic and the decision rule are also introduced. Section 3 derives the properties of the estimated moments. Section 4 presents a simulation method to approximate the processes of the estimated moments, and it also presents the GMS method. Based on these, a critical value is constructed. Section 5 discusses the uniform size control, the consistency against fixed alternatives and the asymptotic local power against $N^{-1/2}$

local alternatives of our test. We introduce LW's test and make comparisons between our test and LW's in Section 6. Section 7 summarizes Monte-Carlo simulation results, and Section 8 discusses some extensions of our tests. Section 9 concludes. All mathematical proofs are deferred to the Appendix.

2 Test for Conditional Average Treatment Effects

2.1 Hypothesis Formulation

Let D be a dummy variable such that $D = 1$ if the individual receives treatment; otherwise, $D = 0$. Let X be a d_x -dimensional vector of covariates with $d_x \geq 1$ with a compact support \mathcal{X} . Define $Y(1)$ as the potential outcome for the individual under treatment and $Y(0)$ as that without treatment. We observe D , X and $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$. We have a random sample of size N generated from P .

Let $\mu_0(x) = E_P[Y(0)|X = x]$ and $\mu_1(x) = E_P[Y(1)|X = x]$ where E_P denotes the expectation under the distribution P . Let P_x denote the marginal distribution of P on \mathcal{X} . The null hypothesis of interest is that the conditional average treatment effect defined as $\mu_1(x) - \mu_0(x)$ is non-negative for each $x \in \mathcal{X}$. This can be formulated as

$$H_0 : \mu_1(x) - \mu_0(x) \geq 0, \quad \text{for all } x \in \mathcal{X}. \quad (2.1)$$

Note that in the null hypothesis, one can replace \mathcal{X} with any \mathcal{X}_s that is a subset of \mathcal{X} and $P(X \in \mathcal{X}_s) > 0$ and all the results below still hold.³ This type of null hypothesis is interesting especially when the policy maker is interested in the treatment effect among a specific subpopulation defined by a subset of covariate values.

We assume that the treatment assignment is unconfounded. The unconfoundedness assumption introduced by Rosenbaum and Rubin (1983) requires that treatment assignment be independent of the potential outcomes conditional on the observable covariates. Unconfoundedness assumption is also known as selection-on-observables, conditional independence, and ignorability in the literature. The formal definition of the unconfoundedness assumption is the following.

Assumption 2.1 (Unconfoundedness Assumption): $(Y(0), Y(1)) \perp D | X$.

³For the cases where $P(X \in \mathcal{X}_s) = 0$, a different approach is required. For example, if $\mathcal{X}_s = \{(x_1, \dots, x_{d_x}) : x_1 = a\}$ for some constant a and X_1 is a continuous random variable, then the approach of Andrews and Shi (2012) that involves nonparametric kernel estimation is needed. We leave this extension for future research.

Let $p(x) = P(D = 1|X = x) = E_P[D|X = x]$ denote the propensity score, the probability of getting treatment for an individual with covariates x , which is assumed to be bounded away from 0 and 1 on \mathcal{X} . Under Assumption 2.1, $\mu_0(x)$ and $\mu_1(x)$ are identified as

$$\mu_0(x) = E_P\left[\frac{(1-D)Y}{1-p(X)}\middle|X=x\right], \quad \mu_1(x) = E_P\left[\frac{DY}{p(X)}\middle|X=x\right].$$

Hence, under Assumption 2.1, (2.1) is equivalent to

$$H_0 : E_P\left[\left(\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)}\right)\middle|X=x\right] \geq 0, \quad \text{for all } x \in \mathcal{X} \quad (2.2)$$

The null hypothesis defined in (2.2) involves a conditional moment inequality.⁴ To extract all the information from (2.2), we adopt AS's instrumental variable approach to transform the conditional moment inequality into infinitely many unconditional ones without information loss. Define $\ell = (x, r) \in \mathbb{R}^{d_x} \times \mathbb{R}$ and $\mathcal{L} = \mathcal{X} \times [0, \bar{r}]$ where $\bar{r} > 0$.⁵ The set of instrument functions we consider is defined as

$$\begin{aligned} \mathcal{G}_{cube} &= \{g_\ell(X) = 1(X \in C_\ell) : C_\ell \in \mathcal{C}_{cube}\}, \\ \mathcal{C}_{cube} &= \left\{C_\ell = \prod_{j=1}^{d_x} [x_j - r, x_j + r] : \ell \in \mathcal{L}\right\}. \end{aligned}$$

AS show that the null hypotheses in (2.1) and (2.2) are equivalent to

$$H_0 : \nu(\ell) \equiv E_P\left[-g_\ell(X)\left(\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)}\right)\right] \leq 0, \quad \text{for all } \ell \in \mathcal{L}. \quad (2.3)$$

That is, we can transform the conditional moment inequality to a continuum of unconditional moment inequalities indexed by the instrument functions. Other choices of the set of instrument functions are available and please AS for more examples.

2.2 Estimation of $\nu(\ell)$

We estimate $\nu(\ell)$ by an IPW estimator similar to HIR

$$\hat{\nu}(\ell) = \frac{1}{N} \sum_{i=1}^N -g_\ell(X_i) \left(\frac{D_i Y_i}{\hat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} \right),$$

⁴We could have defined the null hypothesis as $H_0 : E_P\left[\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)}\middle|X\right] \geq 0$ a.s. in X . However, one of our regularity requires that $\mu_1(x)$ and $\mu_0(x)$ are continuous in $x \in \mathcal{X}$. Because of this condition, the a.s. version of our null hypothesis is identical to (2.2).

⁵We should have defined $\mathcal{L} = CH(\mathcal{X}) \times [0, \bar{r}]$, where $CH(A)$ denotes the convex hull of A , as in AS. However, the regularity conditions introduced below require that \mathcal{X} be a Cartesian product of compact and connected intervals, which implies that $CH(\mathcal{X}) = \mathcal{X}$.

where $\hat{p}(X_i)$ is a nonparametric estimator for $p(x)$. As in HIR, we use the Series Logit Estimator (SLE) to estimate $p(x)$ based on a power series. Let $\phi = (\phi_1, \dots, \phi_{d_x})' \in \mathbb{Z}_+^{d_x}$ be an d_x -dimensional vector of non-negative integers, where \mathbb{Z}_+ denotes the set of non-negative integers, and define the norm for ϕ as $|\phi| = \sum_{j=1}^{d_x} \phi_j$. Let $\{\phi(k)\}_{k=1}^\infty$ be a sequence including all distinct $\phi \in \mathbb{Z}_+^{d_x}$ such that $|\phi(k)|$ is non-decreasing in k and let $x^\phi = \prod_{j=1}^{d_x} x_j^{\phi_j}$. For any integer K , define $R^K(x) = (x^{\phi(1)}, \dots, x^{\phi(K)})'$ as a vector of power functions. Let $L(a) = \exp(a)/(1 + \exp(a))$ be the logistic cumulative distribution function (CDF). The SLE for $p(X_i)$ is defined as $\hat{p}(x) = L(R^K(x)' \hat{\pi}_K)$, where

$$\hat{\pi}_K = \arg \max_{\pi_K} \frac{1}{N} \sum_{i=1}^N D_i \cdot \log(L(R^K(X_i)' \pi_K)) + (1 - D_i) \cdot \log(1 - L(R^K(X_i)' \pi_K)).$$

Other nonparametric estimators can be used to estimate the propensity score function, e.g., local polynomial estimators in Ichimura and Linton (2005), but the estimated propensity score is not necessarily bounded away from 0 and 1 in finite samples, and proper trimming is required. However, trimming is not required for SLE because the estimated propensity score function is automatically bounded away from 0 and 1. Furthermore, one can also use the imputation estimator to estimate $\nu(\ell)$, as in Heckman, Ichimura and Todd (1997, 1998), Heckman, Ichimura, Smith and Todd (1998), and Hahn (1998):

$$\hat{\nu}(\ell) = \frac{1}{N} \sum_{i=1}^N -g_\ell(X_i)(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)),$$

where $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$ are nonparametric estimators for $\mu_0(x)$ and $\mu_1(x)$ for all $x \in \mathcal{X}$. We expect that under suitable assumptions, all the results discussed below still hold when one uses the imputation estimator.

2.3 Test Statistic and Decision Rule

The KS test statistic is defined as

$$\hat{S}_N = \sqrt{N} \sup_{\ell \in \mathcal{L}} \hat{\nu}(\ell). \tag{2.4}$$

In this paper, we focus on the non-standardized version of the test, but the results developed below can be extended to the standardized version of the test as in AS.⁶ Also, all results can be extended easily to Cramér-von Mises type tests.

⁶Ideally, the standardized version of our test statistics should be defined as $\hat{S}_N = \sqrt{N} \sup_{\ell \in \mathcal{L}} \hat{\nu}(\ell)/\hat{\sigma}(\ell)$ where $\hat{\sigma}^2(\ell)$ is an estimator for asymptotic variance of $\sqrt{N}(\hat{\nu}(\ell) - \nu(\ell))$. However, because $\hat{\sigma}^2(\ell)$ is not uniformly bounded away from 0, this will cause problem when $\hat{\nu}(\ell)$ is divided by $\hat{\sigma}(\ell)$, so we need to modify $\hat{\sigma}(\ell)$ and the standardized test statistic. For more details, refer to Section 3.1 of AS.

Given a simulated critical value c that will be defined later, the decision rule is the following:

$$\text{Reject } H_0 \text{ if } \widehat{S}_N > c. \quad (2.5)$$

3 Asymptotics of $\widehat{\nu}(\ell)$ and the Test Statistic

3.1 Assumptions

In addition to the unconfoundedness assumption, we assume the following regularity conditions which are slightly stronger than those in HIR given that we want to obtain uniformity results of our tests which are stronger than those in HIR. For a scalar function $h(x)$, define the partial derivatives $\partial^\phi h(x) = \partial^\phi h(x) / \partial x_1^{\phi_1} \cdots \partial x_{d_x}^{\phi_{d_x}}$ and $|h(x)|_s = \max_{|\phi| \leq s} \sup_{x \in \mathcal{X}} |\partial^\phi h(x)|$. Let P_x denote the marginal distribution of P on \mathcal{X} . We make the following assumptions.

Assumption 3.2 (Support of X): *The support of the d_x -dimensional covariates X is a Cartesian product of compact intervals, $\mathcal{X} = \prod_{j=1}^{d_x} [x_{\ell_j}, x_{u_j}]$.*

Assumption 3.3 *Let (Ω, F, \mathbb{P}) be the underlying probability space equipped with probability distribution \mathbb{P} . Let \mathcal{P} denote the collection of distributions P such that:*

- (i) $\{W_i = (Y_i, D_i, X_i) : i \geq 1\}$ are i.i.d. under P .
- (ii) P_x is absolutely continuous on \mathcal{X} w.r.t the Lebesgue measure with density $0 < \delta \leq f(x) \leq M < \infty$ for all $x \in \mathcal{X}$.
- (iii) $E_P[|Y(0)|^{2+\delta}] \leq M$ and $E_P[|Y(1)|^{2+\delta}] \leq M$.
- (iv) $p(x)$ is continuously differentiable of order $s \geq 7d_x$ with $|p(x)|_s \leq M$.
- (v) $p(x)$ is bounded away from 0 and 1: $\delta \leq p(x) \leq 1 - \delta$.
- (vi) $\mu_0(x)$ and $\mu_1(x)$ are continuously differentiable of order $s_1 \geq 1$ for all $x \in \mathcal{X}$ and $|\mu_0(x)|_1 \leq M$ and $|\mu_1(x)|_1 \leq M$.
- (vii) $\sup_{x \in \mathcal{X}} \text{Var}(Y(0)|X = x) \leq M$ and $\sup_{x \in \mathcal{X}} \text{Var}(Y(1)|X = x) \leq M$.
- (viii) M and δ are positive constants not dependent on P .

Note that the functions $p(x)$, $\mu_0(x)$ and $\mu_1(x)$ depend on the underlying P , and the set of \mathcal{P} depend on the M and δ , but for notational simplicity, we suppress their dependence.

Assumption 3.4 (Series Estimator): *The SLE of $p(x)$ uses a power series with $K = N^\vartheta$ for some $d_x/4(s - d_x) < \vartheta < 1/9$.*

Remarks on the Assumptions:

1. The requirement that $f(x)$ is uniformly bounded away from zero is important for our uniform asymptotic. As in the proof of Theorem 4 of Newey (1997), this implies that the minimum eigenvalue of $E_P[R^K(X)R^K(X)'] \geq C$ for all $P \in \mathcal{P}$ where C is some positive number and $P^K(x)$ is obtained by replacing the individual powers in each term by the Jacobi polynomials of the same order such that $\int_{\mathcal{X}} R^K(x)R^K(x)'dx = I_K$ where I_K denotes the $K \times K$ identity matrix. See Appendix A for more details. This property is similar to Assumption G (ii) of Andrews (1991) and this is important for us to obtain the uniform version of Lemma 1 and Lemma 2 of HIR.
2. Assumption 3.3 (ii) requires that all of the covariates are continuous. However, at the expense of additional notation, we can deal with the case where X has both continuous and discrete components.
3. Assumption 3.4 restricts the growth rate of the number of approximating functions to be included in the series approximation to the propensity score function. Assumption 3.3 (iv) and (v) ensure the existence of a ϑ satisfying the conditions in Assumption 3.4.
4. The moment requirement on $Y(0)$ and $Y(1)$ in Assumption 3.3(iii) is stronger than Assumption 3 of HIR which only requires the second moments of $Y(0)$ and $Y(1)$ to be bounded. We need stronger moment conditions on $Y(0)$ and $Y(1)$ because we are interested in uniform asymptotics which in general requires stronger conditions than pointwise asymptotics.
5. Assumption 3.3(vii) is needed for the estimators for $\mu_0(x)$ and $\mu_1(x)$ to be consistent uniformly over $x \in \mathcal{X}$ and $P \in \mathcal{P}$. This condition is not specified in HIR, but they use this condition implicitly in their Theorem 2.

3.2 Asymptotics of $\hat{\nu}(\ell)$

Following Linton, Song and Whang (2010), we define two notations: $O_{\mathcal{P}}(1)$ and $o_{\mathcal{P}}(1)$. Let Z_N be a sequence of random variables and we say that Z_N is $O_{\mathcal{P}}(1)$ uniformly in $P \in \mathcal{P}$ if for any $\epsilon > 0$, there exists $M > 0$ and $N_{\epsilon} > 0$ such that $\sup_{P \in \mathcal{P}} P(|Z_N| < M) < \epsilon$ for all $N \geq N_{\epsilon}$ and we denote it as $Z_N = O_{\mathcal{P}}(1)$. We say that Z_N is $o_{\mathcal{P}}(1)$ uniformly in $P \in \mathcal{P}$ if for any $\epsilon > 0$, such that $\sup_{P \in \mathcal{P}} P(|Z_N| > \epsilon) \rightarrow 0$ and we denote it as $Z_N = o_{\mathcal{P}}(1)$.

Lemma 3.1 *Suppose that Assumptions 2.1, 3.2, 3.3 and 3.4 hold. Then*

$$\sup_{\ell \in \mathcal{L}} \left| \sqrt{N}(\hat{\nu}(\ell) - \nu(\ell)) - \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi_{\ell}(W_i) \right| = o_{\mathcal{P}}(1),$$

where

$$\psi_{\ell}(W) = g_{\ell}(X) \left(\frac{(1-D)Y}{1-p(X)} - \frac{DY}{p(X)} + (D-p(X)) \left(\frac{\mu_0(X)}{1-p(X)} + \frac{\mu_1(X)}{p(X)} \right) \right) - \nu(\ell)$$

where $W \equiv \{Y, D, X\}$.

Lemma 3.1 shows that the estimator $\sqrt{N}(\hat{\nu}(\ell) - \nu(\ell))$ has an asymptotic linear representation uniformly over $\ell \in \mathcal{L}$ and uniformly over $P \in \mathcal{P}$. To show Lemma 3.1, we need to obtain uniform versions of Lemma 1 and Lemma 2 of HIR regarding the SLE. We summarize this set of results in Appendix A which might be of separate interest.⁷

Let $h_2(\cdot, \cdot)$ be a covariance kernel on $\mathcal{L} \times \mathcal{L}$. Let \mathcal{H}_2 be the collection of all possible covariance kernel function on $\mathcal{L} \times \mathcal{L}$. For any pair of $h_2^{(1)}$ and $h_2^{(2)}$, we define the distance between them as

$$d(h_2^{(1)}, h_2^{(2)}) = \sup_{\ell_1, \ell_2 \in \mathcal{L}} |h_2^{(1)}(\ell_1, \ell_2) - h_2^{(2)}(\ell_1, \ell_2)|. \quad (3.1)$$

Define $h_{2,P}(\ell_1, \ell_2) = E_P[\psi_{\ell_1}(W)\psi_{\ell_2}(W)] = Cov_P(\psi_{\ell_1}(W)\psi_{\ell_2}(W))$ which denotes the covariance kernel generated by $\psi_{\ell}(W)$ under distribution $P \in \mathcal{P}$ defined in Assumption 3.3. An implication of Lemma 3.1 is the following.

Lemma 3.2 *Suppose that Assumptions 2.1, 3.2, 3.3 and 3.4 hold. For any subsequence of a_N of N such that $\lim_{N \rightarrow \infty} d(h_{2,P_{a_N}}, h_2^*) = 0$ for some $h_2^* \in \mathcal{H}_2$, we have*

$$\sqrt{a_N}(\hat{\nu}_{a_N}(\cdot) - \nu_{a_N}(\cdot)) \Rightarrow \Psi_{h_2^*}(\cdot),$$

where $\Psi_{h_2^*}(\cdot)$ denotes a mean zero Gaussian process with covariance kernel h_2^* .

Note that Lemma 3.2 implies the Assumption EP (b) of AS.⁸ Another implication of Lemma 3.2 is that under any fixed $P \in \mathcal{P}$, $\sqrt{N}(\hat{\nu}(\cdot) - \nu(\cdot)) \Rightarrow \Psi_{h_{2,P}}(\cdot)$, i.e., Assumption EP (a) of AS holds. As noted by AS, the Assumption EP of AS is the key condition for us to obtain the uniformity of tests when preliminary consistent estimator is required. Assumption EP of AS is a high-level assumption, so we contribute to the literature by providing the low-level sufficient conditions for Assumption EP of AS.

⁷Our proof can be generalized to obtain the uniform version of the results in Andrews (1991) and Newey (1997), and we leave this extension for future research.

⁸In our proof, we will not need Assumption EP (b)(ii) of AS.

4 GMS Critical Value

In this section, we introduce the GMS critical value for our tests. We introduce the simulated stochastic process $\Psi^u(\cdot)$ and AS's GMS method. We defined the simulated critical value for our test using the simulated process and the GMS.

4.1 Simulated Process

Let U_1, U_2, \dots be i.i.d. random variables with mean equal to zero and variance equal to one that are independent of the sequence $\mathcal{W} = \{W_1, W_2, \dots\}$. For all $\ell \in \mathcal{L}$, we define the simulated stochastic processes $\Psi^u(\ell)$ as

$$\Psi^u(\ell) = \frac{1}{\sqrt{N}} \sum_{i=1}^N U_i \left(g_\ell(X_i) \left(\frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} - \frac{D_i Y_i}{\hat{p}(X_i)} + (D_i - \hat{p}(X_i)) \left(\frac{\hat{\mu}_0(X_i)}{1-\hat{p}(X_i)} + \frac{\hat{\mu}_1(X_i)}{\hat{p}(X_i)} \right) \right) - \hat{\nu}(\ell) \right), \quad (4.1)$$

where $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$ are the series estimators for $\mu_0(x)$ and $\mu_1(x)$:

$$\begin{aligned} \hat{\mu}_0(x) &= \left(\sum_{i=1}^N \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} R^K(X_i) \right)' \left(\sum_{i=1}^N R^K(X_i) R^K(X_i)' \right)^{-1} R^K(x), \\ \hat{\mu}_1(x) &= \left(\sum_{i=1}^N \frac{D_i Y_i}{\hat{p}(X_i)} R^K(X_i) \right)' \left(\sum_{i=1}^N R^K(X_i) R^K(X_i)' \right)^{-1} R^K(x). \end{aligned} \quad (4.2)$$

In Appendix A, we show that $\sup_{x \in \mathcal{X}} |\hat{\mu}_0(x) - \mu_0(x)| = o_{\mathcal{P}}(1)$ and $\sup_{x \in \mathcal{X}} |\hat{\mu}_1(x) - \mu_1(x)| = o_{\mathcal{P}}(1)$.

The following lemma summarizes the property of our simulated processes which might be of separate interest.

Lemma 4.3 *Suppose that Assumptions 2.1, 3.2, 3.3 and 3.4 hold. For any subsequence of a_N of N such that $\lim_{N \rightarrow \infty} d(h_{2, P_{a_N}}, h_2^*) = 0$ for some $h_2^* \in \mathcal{H}_2$, we have $\Psi_{a_N}^u(\ell) \xrightarrow{\mathbb{P}} \Psi_{h_2^*}(\ell)$.⁹*

Lemma 4.3 shows that for any subsequences P_{a_N} of P_N such that $\lim_{N \rightarrow \infty} d(h_{2, P_{a_N}}, h_2^*) = 0$, then we can approximate the limiting process of $\sqrt{a_N}(\hat{\nu}_{a_N}(\cdot) - \nu_{a_N}(\cdot))$ well. It is obvious that if $P_N = P$ for all N , then we can approximate the limiting process of $\sqrt{N}(\hat{\nu}(\cdot) - \nu(\cdot))$ well.

⁹Let X_N be a stochastic process defined on the probability space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2, \mathcal{P}_1 \times \mathcal{P}_2)$ and $X_N(\omega_1)$ the process of X_N conditional on ω_1 and P_1^* the outer probability. Then $X_N \xrightarrow{a.s.} X$ if $P_1^*(\{\omega_1 | X_N(\omega_1) \Rightarrow X\}) = 1$. $X_N \xrightarrow{\mathbb{P}} X$ if for any subsequence h_N of N , there is a further subsequence ℓ_N such that $X_{\ell_N} \xrightarrow{a.s.} X$. In our case, we have $(\Omega_1, \mathcal{A}_1, \mathcal{P}_1) = (\Omega, \mathcal{F}, \mathbb{P})$ and $(\Omega_2, \mathcal{A}_2, \mathcal{P}_2)$ is the probability space for U^∞ .

4.2 Generalized Moment Selection

As most papers in the moment inequality literature, our paper uses the generalized moment selection method to construct the critical value. The GMS method is introduced by Andrews and Soares (2010) and AS. It is similar to the recentering method in Hansen (2005) and Donald and Hsu (2010), as well as the contact set approach in Linton, Song and Whang (2010). By doing this, one can construct a more powerful test without resorting to the LFC.

We define the GMS function or the recentering function for our case. Let B_N be a sequences of positive numbers. Define $\phi_N(\ell)$ as

$$\phi_N(\ell) = -B_N \cdot 1(\sqrt{N}\hat{\nu}(\ell) < a_N), \quad (4.3)$$

where a_N is a sequence of negative numbers. We will pick B_N in the way such that $\phi_N(\ell) \leq \nu(\ell)$ uniformly over $\ell \in \mathcal{L}$ with probability approaching one. This property allows us to obtain a test with correct uniform asymptotic size as in AS.

Assumption 4.5 *Assume that:*

1. Let a_N be a sequence of negative numbers satisfying $\lim_{N \rightarrow \infty} a_N = -\infty$ and $\lim_{N \rightarrow \infty} a_N/\sqrt{N} = 0$.
2. Let B_N be a sequence of positive numbers satisfying that B_N is non-decreasing, $\lim_{N \rightarrow \infty} B_N = \infty$ and $\lim_{N \rightarrow \infty} B_N/a_N = 0$.

We suggest $a_N = -(0.3 \ln(N))^{1/2}$ and $B_N = (0.4 \ln(N)/\ln \ln(N))^{1/2}$ according to AS.¹⁰ As we will see below, the generalized moment selection function is added to the simulated process before we take the supremum. By doing this, we can approximate the null distribution without resorting to the LFC so as to improve the power of our test.

4.3 Simulated Critical Value

For a significance level $\alpha < 1/2$, we define the critical value \hat{c}_η as

$$\hat{c}_\eta = \sup \left\{ q \mid P^* \left(\sup_{\ell \in \mathcal{L}} \Psi^u(\ell) + \phi_N(\ell) \leq q \right) \leq 1 - \alpha + \eta \right\} + \eta,$$

¹⁰If one is only interested in the pointwise asymptotic, one can define the GMS function as $\phi_N(\ell) = \sqrt{N}\hat{\nu}(\ell) \cdot 1(\sqrt{N}\hat{\nu}(\ell) < a_N)$ and this can further improve the power of our test and control the pointwise size of our test asymptotically.

where $\eta > 0$ is an arbitrarily small positive number, e.g., 10^{-6} . Note that \hat{c}_η is the $(1 - \alpha + \eta)$ -th quantile of the simulated null distribution plus η . AS call the constant η an infinitesimal uniformity factor that is used to avoid the problems that arise due to the presence of the infinite-dimensional nuisance parameter $\nu(\ell)$ and to eliminate the need for complicated and difficult-to-verify uniform continuity and strictly-increasing conditions on the large sample distribution functions of the test statistic.

5 Size and Power Properties

In this section, we show that our test can control the size uniformly over a set of DGPs asymptotically and is consistent against fixed alternatives. We also discuss the local power properties of our test and show that our test is asymptotically unbiased against some $N^{-1/2}$ local alternatives.

5.1 Uniform Size Control

Assumption 5.6 *Let \mathcal{P}_0 be the subset of \mathcal{P} that satisfies Assumption 3.3 such that the null hypothesis in (2.1) holds under P if $P \in \mathcal{P}_0$.*

The following theorem summarizes the uniform size of our test. Let $\lambda(\cdot)$ denotes the Lebesgue measure.

Theorem 5.1 *Suppose Assumptions 2.1, 3.2, 3.3, 3.4, 4.5 and 5.6 hold and $\alpha < 1/2$. We reject H_0 when $\hat{S}_N > \hat{c}_\eta$. Then, for every compact subset of $\mathcal{H}_{2,cpt}$ of \mathcal{H}_2*

$$(a) \limsup_{N \rightarrow \infty} \sup_{\{P \in \mathcal{P}_0: h_{2,P} \in \mathcal{H}_{2,cpt}\}} P(\hat{S}_N > \hat{c}_\eta) \leq \alpha.$$

(b) *if there exists $P_c \in \mathcal{P}_0$ and $h_{2,P_c} \in \mathcal{H}_{2,cpt}$ such that $\lambda(\{x \in \mathcal{X} : \mu_1(x) = \mu_0(x)\}) > 0$ under P_c , then $\lim_{\eta \rightarrow 0} \limsup_{N \rightarrow \infty} \sup_{\{P \in \mathcal{P}_0: h_{2,P} \in \mathcal{H}_{2,cpt}\}} P(\hat{S}_N > \hat{c}_\eta) = \alpha$.*

Theorem 5.1(a) shows that our test have correct uniform asymptotic size over a compact sets of covariance kernels which is similar to Theorem 2(a) of AS. Theorem 5.1(b) shows that our test is at most infinitesimally conservative asymptotically when there exists at least one P_c such that $\lambda(\{x \in \mathcal{X} : \mu_1(x) = \mu_0(x)\}) > 0$ under P_c . Theorem 5.1 is similar to Theorem 2 of AS except that we consider the case where preliminary consistent estimator for the propensity score function is needed. AS only provide the high-level assumptions for our Theorem 5.1 and we make contribution to the literature by providing low-level conditions that are easier to verify.

5.2 Power against Fixed Alternatives

We present the power of our test against fixed alternatives. Let $P_1 \in \mathcal{P}$ such that $\mu_1(x^*) - \mu_0(x^*) < 0$ for some $x^* \in \mathcal{X}$ under P_1 . Then by the continuity of $\mu_1(x) - \mu_0(x)$, there exists a neighborhood $\mathcal{N}(x^*)$ around x^* such that $\mu_1(x) - \mu_0(x) < 0$ for all $x \in \mathcal{N}(x^*)$ and $\lambda(\mathcal{N}(x^*)) > 0$. This implies that there exists $\ell^* \in \mathcal{L}$ such that $\nu(\ell^*) > 0$. The following theorem shows the consistency of our test.

Theorem 5.2 *Suppose Assumptions 2.1, 3.2, 3.3, 3.4, 4.5 hold and $\alpha < 1/2$. Let $P_1 \in \mathcal{P}$ such that $\mu_1(x^*) - \mu_0(x^*) < 0$ for some $x^* \in \mathcal{X}$ under P_1 . We reject H_0 when $\widehat{S}_N > \widehat{c}_\eta$. Then, $\lim_{N \rightarrow \infty} P(\widehat{S}_N > \widehat{c}_\eta) = 1$.*

Theorem 5.2 follows from the fact that the test statistic \widehat{S}_N diverges to positive infinity under P_1 and the critical value \widehat{c}_η is bounded in probability.

5.3 Local Asymptotic Power

We show that our test is unbiased against some $N^{-1/2}$ local alternatives. Define $A \setminus B \equiv \{x : x \in A \text{ but } x \notin B\}$ for any two subsets A and B . We consider a sequence of $P_N \in \mathcal{P} \setminus \mathcal{P}_0$ such that $\mu_{1,N}(x) - \mu_{0,N}(x) = \mu_1(x) - \mu_0(x) + \delta(x)/\sqrt{N}$ under P_N and $\mu_1(x) - \mu_0(x) \geq 0$ for all $x \in \mathcal{X}$ so that H_0 holds. Define $\mathcal{X}^o \equiv \{x : \mu_1(x) - \mu_0(x) = 0\}$. We impose the following conditions on the local alternatives we consider.

Assumption 5.7 *Let $\{P_N \in \mathcal{P} \setminus \mathcal{P}_0 : N \geq 1\}$ satisfy:*

1. $\mu_{1,N}(x) - \mu_{0,N}(x) = \mu_1(x) - \mu_0(x) + \delta(x)/\sqrt{N}$ under P_N .
2. $\mu_1(x) - \mu_0(x) \geq 0$ for all $x \in \mathcal{X}$.
3. $\lambda(\mathcal{X}^o) > 0$.
4. $\delta(x) \leq 0$ if $x \in \mathcal{X}^o$.
5. $\lambda(\delta^- \cap \mathcal{X}^o) > 0$ where $\delta^- \equiv \{x : \delta(x) < 0\}$.
6. $\lim_{N \rightarrow \infty} d(h_{2,P_N}, h_2^*) = 0$ for some $h_2^* \in \mathcal{H}_2$

The following Theorem summarizes the limiting distribution of the test statistic and limit of the critical value under the local alternatives that satisfy Assumption 5.7.

Theorem 5.3 *Suppose Assumptions 2.1, 3.2, 3.3, 3.4, 4.5 hold and $\alpha < 1/2$. Under $\{P_N : N \geq 1\}$ that satisfies Assumption 5.7, Then, $\lim_{\eta \rightarrow 0} \lim_{N \rightarrow \infty} P(\widehat{S}_N > \widehat{c}_\eta) \geq \alpha$.*

Theorem 5.3 shows that the asymptotic local power of our test is greater than or equal to α when η tends to zero, i.e., our test is unbiased against those local alternatives. It is well known that tests involving inequalities are only unbiased against some $N^{-1/2}$ local alternatives. Note that if the deviation from the null is allowed to be negative on \mathcal{X}^o , our test may be biased. A simple example regarding this can be found in Donald and Hsu's (2010) Example 4.6.

6 Comparisons with Lee and Whang (2009)

This section compares our tests and LW's, which is only alternative to our test for conditional treatment effect. We first summarize LW's test. We show that that LW's test, which is constructed based on a user-chosen strict subset of \mathcal{X} , may be inconsistent if the violation of the null is outside of the user-chosen subset. Furthermore, we show that under a broad set of $N^{-1/2}$ local alternatives that do not converge to the least favorable case, our test is more powerful than LW's.

6.1 Lee and Whang's Test

Define the kernel estimators

$$\hat{\mu}_0(x) = \frac{\sum_{\{i:D_i=0\}} Y_i K_h(x - X_i)}{\sum_{\{i:D_i=0\}} K_h(x - X_i)}, \quad \hat{\mu}_1(x) = \frac{\sum_{\{i:D_i=1\}} Y_i K_h(x - X_i)}{\sum_{\{i:D_i=1\}} K_h(x - X_i)},$$

where $K_h(x - X_i) = h^{-d} K((x - X_i)/h)$, $K(\cdot)$ is a kernel function and h is the bandwidth.¹¹ The test statistic of LW's is a one-sided version of L_1 -type functionals of $\hat{\mu}_0(x) - \hat{\mu}_1(x)$, which is defined as

$$\hat{T} = \int_{\mathcal{X}} \sqrt{N} \max\{\hat{\mu}_0(x) - \hat{\mu}_1(x), 0\} w(x) dx,$$

where $w(x) \geq 0$ is a weight function with support \mathcal{W}_x that is assumed to be a strict subset of \mathcal{X} so as to avoid the boundary problem of kernel estimators. In the least favorable case of the null hypothesis, LW shows that

$$\frac{\hat{T} - a_N}{\sigma_N} \xrightarrow{D} \mathcal{N}(0, 1). \tag{6.1}$$

The exact definitions of a_N and σ_N^2 are given in the Appendix. Let \hat{a}_N and $\hat{\sigma}_N^2$ be estimators for a_N and σ_N^2 , which are also defined in the Appendix. It is shown that the

¹¹Note that in LW's test $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$ denote the kernel estimators for $\mu_0(x)$ and $\mu_1(x)$. But in our test, $\hat{\mu}_0(x)$ and $\hat{\mu}_1(x)$ denote the series estimators for $\mu_0(x)$ and $\mu_1(x)$. Hopefully, there is no confusion caused by this abuse of notation.

asymptotic normality of \widehat{T} in (6.1) still holds with \widehat{a}_N and $\widehat{\sigma}_N$ in place of a_N and σ_N respectively. Hence, define the standardized test statistic as

$$\widehat{S}_{LW} = \frac{\widehat{T} - \widehat{a}_N}{\widehat{\sigma}_N},$$

and the rejection rule of LW's test is

$$\text{Reject } H_0 \text{ if } \widehat{S}_{LW} > z_{1-\alpha},$$

where $z_{1-\alpha}$ is the $(1-\alpha)$ -th quantile of the standard normal. Theorems 4.1 and 4.2 of LW show that their test can control the size well and is consistent against a fixed alternative if $\lambda(\{x \in \mathcal{W}_x : \mu_1(x) - \mu_0(x) < 0\}) > 0$.

6.2 Advantages of Our Test over LW's

1. The first advantage of our test over LW's is that our test is consistent against all fixed alternatives, but LW's test is only consistent if $\lambda(\{x \in \mathcal{W}_x : \mu_1(x) - \mu_0(x) < 0\}) > 0$ because they need to restrict their attention to the subset \mathcal{W}_x to avoid the boundary problem of the kernel estimators $\widehat{\mu}_0(x)$ and $\widehat{\mu}_1(x)$. Therefore, LW's test may not have power when the violation of the null is outside of \mathcal{W}_x . To restore the consistency of LW's test, one can allow the \mathcal{W}_x to expand to \mathcal{X} when N tends to infinity, but the theory for this result is not trivial and remains unresolved.¹²
2. If the violation of the null is within \mathcal{W}_x , our test and LW's are both consistent, and which test is more powerful depends on the underlying fixed alternatives. However, we can show that under a broad set of local alternatives, our test is more powerful than LW's. To show this, we modify our Assumption 5.7. Define $\mathcal{W}_x^o \equiv \{x \in \mathcal{W}_x : \mu_1(x) - \mu_0(x) = 0\}$.

Assumption 6.8 *In addition to the conditions in Assumption 5.7, we assume the following conditions:*

1. $\lambda(\mathcal{W}_x^o) > 0$.
2. $\lambda(\delta^- \cap \mathcal{W}_x^o) > 0$.
3. $\lambda(\mathcal{W}_x \setminus \mathcal{W}_x^o) > 0$.

¹²It may be possible to extend LW's theory to cases where $\mathcal{W}_x = \mathcal{X}$ under our assumption that the density of $f(x)$ is bounded away from 0, but this also requires some work.

The last condition requires that the measure of $\mathcal{W}_x \setminus \mathcal{W}_x^o \equiv \{x : \mu_1(x) - \mu_0(x) > 0\}$ is strictly positive, which implies that the local alternatives will not converge to the least favorable case. The following theorem shows that LW's test has no power against the local alternatives satisfying Assumption 6.8. Note that the local alternatives defined in Assumption 6.8 is a subset of those defined in Assumption 5.7, so our test is still unbiased under Assumption 6.8. Therefore, our test is more powerful than LW's against those local alternatives satisfying Assumption 6.8.

Theorem 6.4 *Suppose Assumptions 2.1, 3.2, 3.3, 3.4, 4.5 hold and $\alpha < 1/2$. Under $\{P_N : N \geq 1\}$ that satisfies Assumptions 5.7 and 6.8, $\lim_{N \rightarrow \infty} P(\widehat{S}_{LW} > z_{1-\alpha}) = 0$.*

Theorem 6.4 shows that LW's test has no power against those local alternatives that do not converge to the least favorable case. In the proof of Theorem 6.4, we show that \widehat{S}_{LW} will converge to negative infinity if the local alternatives do not converge to the least favorable case in the limit. Therefore, the local power of LW's test is $\lim_{N \rightarrow \infty} P(\widehat{S}_{LW} > Z_{1-\alpha}) = 0$. Again, our test remains unbiased in this case, so our test is more powerful than LW's against a broad set of local alternatives that do not converge to the least favorable case.¹³

7 Monte-Carlo Simulations

In this section, we conduct small-scale Monte-Carlo simulations to illustrate the finite sample performance of our test and LW's.

Example 7.1 *Let the data generating process (DGP) be:*

$$\begin{aligned} X &= U_x, & D &= 1(U_t < 0.3 + 0.4X), \\ Y(0) &= 2(X - 0.5) + \mathcal{N}_0, \\ Y(1) &= 2(X - 0.5) + \mathcal{N}_1, \\ Y &= DY(1) + (1 - D)Y(0), \end{aligned}$$

where U_x and U_t are uniform distributions over $[0, 1]$ and \mathcal{N}_0 and \mathcal{N}_1 are standard normals. U_x , U_t , \mathcal{N}_0 and \mathcal{N}_1 are independent.

¹³In Section 5.1 of LW, they propose a more powerful test based on the contact set approach. In fact, with a suitable modification of assumption 6.8, we can show that the result in Theorem 6.4 holds for LW's more powerful test.

In Example 7.1, we have $\mu_1(x) - \mu_0(x) = 0$ for all $x \in \mathcal{X}$, which is the least favorable case of the null hypothesis. We use this example to illustrate the size properties of our test and LW's when the null hypothesis is the least favorable case.

Example 7.2 *Let the DGP be the same as in Example 7.1 except that*

$$Y(0) = 2(X - 0.5) \cdot 1(X < 0.5) + \mathcal{N}_0, \quad Y(1) = \mathcal{N}_1.$$

In Example 7.2, we have $\mu_1(x) - \mu_0(x) \geq 0$ for all $x \in \mathcal{X}$ and the strictly inequality holds when $x < 0.5$. We use this example to illustrate the size properties of our test and LW's when the null hypothesis is not the least favorable case.

Example 7.3 *Let the DGP be the same as in Example 7.1 except that $Y(1) = \mathcal{N}_1$.*

In Example 7.3, we have $\mu_1(x) - \mu_0(x) \geq 0$ when $x \leq 0.5$ and $\mu_1(x) - \mu_0(x) < 0$ when $x > 0.5$, i.e., the null hypothesis is violated. We use this example to show the power of our test and LW's against fixed alternatives.

Example 7.4 *Let the DGP be the same as in Example 7.1 except that*

$$Y(0) = 2(X - 0.5) \cdot 1(X < 0.5) + N^{-1/2} + \mathcal{N}_0, \quad Y(1) = \mathcal{N}_1.$$

We use Example 7.4 to demonstrate the local power of our test and LW's test. Note that the DGP in Example 7.4 converges to the DGP in Example 7.2, which is not the least favorable case of the null hypothesis.

Following Section 3.5 of AS, we approximate \widehat{S}_N by a finite number of instrument functions. The intervals we consider here are those with lengths r^{-1} for $r = 1, \dots, r_1$ where $r_1 = 8$ so we use a total of 36 intervals.^{14,15} We set $a_N = -(0.3 \ln(N))^{1/2}$ and $B_N = (0.4 \ln(N) / \ln \ln(N))^{1/2}$ according to AS.¹⁶ When approximating the critical value, we let U_i 's be independent standard normal random variables, i.e., $U_i \sim \mathcal{N}(0, 1)$ and the simulated test statistic is approximated by the same set of instrument functions as well.

¹⁴For example, if $r = 3$, we use instrument functions $1(0 \leq X \leq 1/3)$, $1(1/3 \leq X \leq 2/3)$ and $1(2/3 \leq X \leq 1)$. Therefore, if $r_1 = 8$, we have a total of 36 instrument functions.

¹⁵Our simulation results are not sensitive to the choice of r_1 .

¹⁶In general, the smaller the a_N and the larger the B_N , the more conservative and the less powerful our test in finite samples.

For each simulation, we approximate the critical value by 1,000 repetitions and we pick $\eta = 10^{-6}$.¹⁷ We consider three different sample sizes: 300, 500 and 1,000. When $N = 300$, the propensity score function is estimated by the SLE with power series: 1 and X . When $N = 500$ and 1,000, we use 1, X and X^2 .¹⁸

When implementing LW's test, we set the $\mathcal{W}_x = [0.05, 0.95]$ and we use the uniform weight function, with $w(x) = 1$ for all $x \in \mathcal{W}_x$ and $w(x) = 0$ otherwise. The kernel function is $K(u) = 1.5(1 - (2u)^2) \cdot 1(|u| \leq 0.5)$ with bandwidth $h = C_h \hat{\sigma}_X N^{-2/7}$, where $C_h \in \{2, 3\}$ and $\hat{\sigma}_X$ is the sample standard deviation of X as suggested by LW. We use the Reimann sum to approximate the integral in the expression of \hat{T} , \hat{a}_N and $\hat{\sigma}_N^2$ based on 500 gridpoints evenly distributed in $[0.05, 0.95]$.¹⁹

The rejection rates are calculated based on 5,000 simulation repetitions and are summarized in Table 1-4. Table 1 presents the size of our test and LW's for Example 7.1, where the null hypothesis is in the least favorable case. Both tests are a little bit oversized, but the size distortions of our test which are at most 0.66 percentage points are smaller than theirs which are between 1.18 and 3.1 percentage points.²⁰

The simulation results for Example 7.2 are summarized in Table 2. In Table 2, we find that our test is much less conservative than LW's. Actually, by the same argument for Theorem 6.4, we can show that the size of LW's test is equal to zero asymptotically when the null hypothesis is not in the least favorable case. Simulation results in Table 2 support this result because the size of LW's test decreases to zero when the sample size increases for both $C_N = 2$ and $C_N = 3$ cases. On the other hand, the size of our test increases when the sample size increases, but is below 5% level.

Table 3 summarizes the simulation results for Example 7.3, which demonstrates the power of our test and LW's against fixed alternatives. In this case, our test is more powerful than LW's because the rejection rates of our test are all larger than theirs for all sample sizes we consider. Note that when $N = 300$, the difference in power between ours and theirs can be as large as 32.24 percentage points.

Table 4 presents simulation results for Example 7.4. We use this example to illustrate Theorems 5.3 and 6.4 concerning the asymptotic local power of our test and LW's. Note

¹⁷We get similar results when we set $\eta = 0$.

¹⁸The simulation results are not sensitive to the choices of the power series.

¹⁹The simulation results are not sensitive to the numbers of gridpoints we choose among 300, 500 and 1,000.

²⁰When $N = 1000$, the size of our test is 5.22% which is large than 5% when $N = 500$. We consider this as sampling error from the simulation. Note that if the true size is 5%, then the standard errors for the rejection rates based on 5,000 repetitions is $\sqrt{0.05 \cdot 0.95/5000} \approx 0.31\%$. As a result, two standard error is 0.62%.

that the DGP in Example 7.4 satisfies Assumption 6.8. Sections 5 and 6 show that the asymptotic local power of our test will be greater than 5%, but that of LW's is zero. The simulation results support our finding because the simulated rejection rate of our test is 4.44% when $N = 300$ and increases to 7.32% when $N = 1,000$, and those of LW's are 2.22% and 2.28% with $N = 300$ and decrease to 0.68% and 0.76%, respectively, when N is 1,000.

8 Extensions

We present several extensions of our test in this section. We first extend our test to test for the condition stochastic dominance relation between the treatment and control groups. Second, we discuss how to test for the null hypothesis that the conditional treatment effect in non-negative conditional on X_a , which is a strict subset of X . Specifically, we focus on cases where the unconfoundedness assumption will not hold when we condition on X_a only. Third, we extend our tests to the cases where the unconfoundedness assumption does not hold, but there is a binary instrumental variable available, as in the traditional local average treatment effect setup.

8.1 Condition Stochastic Dominance Treatment Effects

We can extend our method to test the conditional stochastic dominance relation between treatment and control groups, which is also considered in LW. To test the conditional stochastic dominance relation, we first replace Y with $1(Y \leq y)$ for all $y \in \mathcal{Y}$ where \mathcal{Y} is the union of the compact supports of $Y(0)$ and $Y(1)$. To be more specific, let $F_0(y|x)$ and $F_1(y|x)$ denote the conditional CDFs of $Y(0)$ and $Y(1)$ given $X = x$, the null hypothesis regarding the conditional stochastic dominance treatment effects is defined as

$$H_0^{sd} : F_1(y|x) - F_0(y|x) \leq 0 \text{ for all } y \in \mathcal{Y} \text{ and for all } x \in \mathcal{X},$$

which is equivalent to

$$H_0^{sd} : \nu(\ell, y) \equiv E_P \left[g_\ell(X) \left(\frac{D \cdot 1(Y \leq y)}{p(X)} - \frac{(1-D) \cdot 1(Y \leq y)}{1-p(X)} \right) \right] \leq 0$$

for all for all $y \in \mathcal{Y}$ and $\ell \in \mathcal{L}$. (8.1)

We estimate $\nu(\ell, y)$ by

$$\hat{\nu}(\ell, y) = \frac{1}{N} \sum_{i=1}^N g_\ell(X_i) \left(\frac{D_i \cdot 1(Y_i \leq y)}{\hat{p}(X_i)} - \frac{(1-D_i) \cdot 1(Y_i \leq y)}{1-\hat{p}(X_i)} \right).$$

Given that $\widehat{F}_0(y|x)$ and $\widehat{F}_1(y|x)$ are monotonically increasing and uniformly consistent estimators²¹ for $F_0(y|x)$ and $F_1(y|x)$ as in Donald and Hsu (2011), then the simulated processes is defined as

$$\Psi^u(\ell, y) = \frac{1}{\sqrt{N}} \sum_{i=1}^N U_i \left(g_\ell(X_i) \left(\frac{D_i \cdot 1(Y_i \leq y)}{\widehat{p}(X_i)} - \frac{(1 - D_i) \cdot 1(Y_i \leq y)}{1 - \widehat{p}(X_i)} + (D_i - \widehat{p}(X_i)) \left(\frac{\widehat{F}_1(y|X_i)}{\widehat{p}(X_i)} + \frac{\widehat{F}_0(y|X_i)}{1 - \widehat{p}(X_i)} \right) \right) - \widehat{v}(\ell, y) \right). \quad (8.2)$$

The test statistic is defined as

$$\widehat{S}_N^{sd} = \sqrt{N} \sup_{\ell \in \mathcal{L}, y \in \mathcal{Y}} \widehat{v}(\ell, y).$$

Given the generalized moment selection function $\phi_N(\ell, y) = -B_N 1(\sqrt{N} \widehat{v}(\ell, y) < a_N)$, the critical value \widehat{c}_η^{sd} is defined as the $(1 - \alpha + \eta)$ -th quantile of $\sup_{\ell \in \mathcal{L}, y \in \mathcal{Y}} (\Psi^u(\ell, y) + \phi_N(\ell, y))$ plus η . The rejection rule is

$$\text{Reject } H_0^{sd} \text{ if } \widehat{S}_N^{sd} > \widehat{c}_\eta^{sd}.$$

Under the similar conditions in Donald and Hsu (2011) and our assumptions, the test for the conditional stochastic dominance treatment effect shares the same prosperities with our test regarding the conditional average treatment effect. The advantages of our method over LW's in this case are similar to the conditional average treatment effect case. Hence, we omit the formal presentation of these results. The extension of our test to the higher order stochastic dominance relation case is straightforward. For example, for j -th order stochastic dominance with $j \geq 2$, we just need to replace $1(Y \leq y)$ with $1(Y \leq y) \cdot (Y - y)^{j-1} / (j - 1)!$.

8.2 Conditional on a Strict Subset X_a

We extend our results to the case when the conditioning set is X_a , a strict subset X , and the unconfoundedness assumption does not hold if we only condition on X_a .²² In the treatment effect literature, most papers focus on either the treatment effects over the whole population or the conditional treatment effects conditional on the whole set of covariates X such that the unconfoundedness assumption holds. For a researcher,

²¹ $\widehat{F}_0(y|x)$ is monotonically increasing if $\widehat{F}_0(y_1|x) \geq \widehat{F}_0(y_2|x)$ for all $y_1 \geq y_2$ and for all $x \in \mathcal{X}$. $\widehat{F}_0(y|x)$ is uniformly consistent for $F_0(y|x)$ if $\sup_{y \in \mathcal{Y}, x \in \mathcal{X}} |\widehat{F}_0(y|x) - F_0(y|x)| = o_p(1)$.

²²If the unconfoundedness assumption holds when we condition only on X_a , then previous methods still work and the theory is valid when we replace X with X_a .

requiring a policy that has a uniformly positive effect on all subgroups defined by X may be too strict in some cases. The present extension will be useful when the researcher is interested in a policy that has a uniformly positive effect on all subgroups defined by X_a . Let \mathcal{X}_a be the support of X_a . The null hypothesis is

$$H_0^a : E_P[Y(1) - Y(0)|X_a = x_a] \geq 0 \text{ for all } x_a \in \mathcal{X}_a. \quad (8.3)$$

Let d_a denote the dimension of X_a and $d_a < d_x$. Define $\ell_a = (x_a, r) \in \mathbb{R}^{d_a} \times \mathbb{R}$ and $\mathcal{L}_a = \mathcal{X}_a \times [0, \bar{r}]$ with $\bar{r} > 0$. The set of instrument functions is defined as

$$\begin{aligned} \mathcal{G}_{cube}^a &= \{g_{\ell_a}(X_a) = 1(X_a \in C_{\ell_a}) : C_{\ell_a} \in \mathcal{C}_{cube}^a\}, \\ \mathcal{C}_{cube}^a &= \left\{C_{\ell_a} = \prod_{j=1}^{d_a} [x_j - r, x_j + r] : \ell_a \in \mathcal{L}_a\right\} \end{aligned}$$

The null hypothesis defined in (8.3) is equivalent to

$$H_0^a : \nu(\ell_a) \equiv E_P \left[-g_{\ell_a}(X_a) \left(\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)} \right) \right] \leq 0, \text{ for all } \ell_a \in \mathcal{L}_a.$$

By replacing $g_{\ell}(X)$ with $g_{\ell_a}(X_a)$, all results of our test can be easily extended to this case. However, it is not trivial to extend LW's test to this case. The main reason is that $\hat{\tau}(x_a)$ defined as

$$\widehat{CATE}(x_a) \equiv \frac{\sum_{i=1}^N (1 - D_i) Y_i K_h(x_a - X_{ai})}{\sum_{i=1}^N (1 - D_i) K_h(x_a - X_{ai})} - \frac{\sum_{i=1}^N D_i Y_i K_h(x_a - X_{ai})}{\sum_{i=1}^N D_i K_h(x_a - X_{ai})}$$

is not a consistent estimator for $CATE(x_a) = E[Y(0) - Y(1)|X_a = x_a]$ because the treatment status is not unconfounded conditional on X_a . For LW's test to work, a two-step estimator for $CATE(x_a)$ as in Abrevaya, Hsu and Lieli (2011) is needed and the extension of the theory on LW's test to this case is not trivial.

8.3 Tests without Unconfoundedness

Finally, like LW's test, our method can be applied to the case where the unconfoundedness assumption does not hold, but there is a binary instrument (Z) available. Under the conditional local average treatment effect (LATE) setup as in Abadie, Angrist and Imbens (2002), Abadie (2002, 2003), Frölich (2007) and Donald, Hsu and Lieli (2011), the LATE is defined as $LATE(x) \equiv E[Y(1) - Y(0)|X = x, \mathcal{C}]$ where \mathcal{C} denotes the group of compliers and the null hypothesis of interest is defined as

$$H_0^{late} : LATE(x) \geq 0 \text{ for all } x \in \mathcal{X}. \quad (8.4)$$

It is well-known that $LATE(x)$ is identified by

$$LATE(x) = E\left[\frac{ZY}{q(X)} - \frac{(1-Z)Y}{1-q(X)} \middle| X = x\right] / E\left[\frac{ZD}{q(X)} - \frac{(1-Z)D}{1-q(X)} \middle| X = x\right],$$

where $q(x) = P(Z = 1|X = x)$. Given that the denominator is assumed to be strictly positive for all $x \in \mathcal{X}$, the null hypothesis in (8.4) is equivalent to

$$H_0^{late} : E_P\left[\frac{ZY}{q(X)} - \frac{(1-Z)Y}{1-q(X)} \middle| X = x\right] \geq 0, \quad \text{for all } x \in \mathcal{X} \quad (8.5)$$

which is identical to (2.2) after we replace D and $p(x)$ with Z and $q(x)$, respectively. In other words, if we treat Z as the primary treatment status, then the test for (8.5) is equivalent to (2.2). A similar argument applies to the stochastic dominance case and to the case where the conditioning set X_a is a strict subset of X .

9 Conclusion

We propose a KS test for the null hypothesis that the conditional average treatment effect is uniformly non-negative over all subgroups defined by the covariates. Our test can control the size uniformly over a set of DGPs asymptotically, is consistent against any fixed alternative and is unbiased against some local alternatives. Our test is more powerful than LW's against a broad set of local alternatives. Monte-Carlo simulations confirm our theoretical findings. Several extensions of our test are presented as well.

APPENDIX

Let C be a generic constant that varies in different cases.

A Uniform Asymptotics for Series Logit Estimator

We extend HIR's pointwise asymptotics of the SLE to uniform asymptotics over a set of DGPs. Following HIR, we consider power series and choose the following sequence of powers of x :

$$\mathcal{R} \equiv \{1, x_1, \dots, x_{d_x}, x_1^2, x_1x_2, x_1x_3, \dots, x_2x_3, x_2x_4, \dots\}, \quad (\text{A.1})$$

and let $r^K(x)$ denotes the first K elements of \mathcal{R} . Let A_K be a $K \times K$ nonsingular matrix, then it is true that the the function spaces generated by $r^K(x)$ and $A_K r^K(x)$ are identical. Define $R^K(x) = A_K r^K(x)$ and we pick A_K such that $\int_{\mathcal{X}} R^K(x) R^K(x)' dx = I_K$. For a matrix A , let the matrix norm be $\|A\| = \sqrt{\text{tr}(A'A)}$ where $\text{tr}(B)$ denotes the trace of a square matrix B . Define

$$\zeta(K) = \sup_{x \in \mathcal{X}} \|R^K(x)\|. \quad (\text{A.2})$$

Newey (1997) shows $\zeta(K) \leq CK$ for some $C > 0$.

Let $L(z) = \exp(z)/(1 + \exp(z))$ be the logistic cdf and it is true that $L^{-1}(p) = \ln(p/(1-p))$ and $L'(z) = L(z)(1-L(z))$. For each K , we define the pseudo true propensity score $p_K^*(x) = L(R^K(x)' \pi_K^*)$ under distribution P where

$$\pi_K^* = \arg \max_{\pi} E_P \left[p(X) \ln (L(R^K(X)' \pi)) + (1-p(X)) \ln (1 - L(R^K(X)' \pi)) \right]. \quad (\text{A.3})$$

Note that $p_K^*(x)$ and π_K^* depends on P and we suppress the dependence when there is no confusion.

We make the assumptions on the set of distribution we consider \mathcal{P}_s which are weaker than Assumption 3.3.

Assumption A.1 *Let \mathcal{P}_s denote the collection of distributions P such that:*

- (i) $\{(D_i, X_i) : i \geq 1\}$ are i.i.d. under P .
- (ii) P_x is absolutely continuous on \mathcal{X} w.r.t the Lebesgue measure with density $0 < \delta \leq f(x) \leq M < \infty$ for all $x \in \mathcal{X}$.
- (iii) $p(x)$ is continuously differentiable of order $s \geq 4d_x$ with $|p(x)|_s \leq M$.
- (iv) $p(x)$ is bounded away from 0 and 1: $\delta \leq p(x) \leq 1 - \delta$.
- (v) $\mu_0(x)$ and $\mu_1(x)$ are continuously differentiable of order $s_1 \geq 1$ for all $x \in \mathcal{X}$ and $|\mu_0(x)|_1 \leq M$ and $|\mu_1(x)|_1 \leq M$.
- (vi) $\sup_{x \in \mathcal{X}} \text{Var}(Y(0)|X=x) \leq M$ and $\sup_{x \in \mathcal{X}} \text{Var}(Y(1)|X=x) \leq M$.
- (vii) M and δ are positive constants not dependent on P .

The main difference between Assumption 3.3 and Assumption A.1 is that in the Assumption A.1 only requires that $p(x)$ is continuously differentiable of order $s \geq 4d_x$ instead of $s \geq 7d_x$ as in Assumptions 3.3. As a result, \mathcal{P} is a subset of \mathcal{P}_s .

The following lemmas are the uniform versions of Lemma 1 and Lemma 2 of HIR.

Lemma A.1 *Suppose Assumptions 3.2 and A.1 hold. Then*

(a) *for each $P \in \mathcal{P}_s$, there exists $\pi_K(P)$ such that*

$$\sup_{P \in \mathcal{P}_s} \sup_{x \in \mathcal{X}} \left| \frac{p(x)}{1-p(x)} - R^K(x)' \pi_K(P) \right| < CK^{-s/d_x}. \quad (\text{A.4})$$

(b) $\sup_{P \in \mathcal{P}_s} \|\pi_K(P) - \pi_K^*(P)\| = O(K^{-s/(2d_x)})$, where $\pi_K(P)$ is defined in (A.4).

(c) $\sup_{P \in \mathcal{P}_s} \sup_{x \in \mathcal{X}} |p(x) - p_K^*(x)| = O(K^{-s/(2d_x)} \zeta(K))$.

Lemma A.2 *Suppose Assumptions 3.2 and A.1 hold. In addition, suppose that $K(N)$ is a sequence of values of K satisfying $\lim_{N \rightarrow \infty} K(N) = \infty$ and $\lim_{N \rightarrow \infty} \zeta(K(N))^4/N = 0$. Then,*

$$\|\hat{\pi}_{K(N)} - \hat{\pi}_{K(N)}^*\| = O_{\mathcal{P}_s} \left(\sqrt{\frac{K(N)}{N}} \right). \quad (\text{A.5})$$

Lemma A.3 *Under the same conditions in Lemma A.2, then*

$$\sup_{x \in \mathcal{X}} |\hat{p}(x) - p(x)| = O_{\mathcal{P}_s} \left(N[\sqrt{K/N} + K^{-s/(2d_x)}] \right). \quad (\text{A.6})$$

Lemma A.4 *Under the same conditions in Lemma A.2, then*

$$\sup_{x \in \mathcal{X}} |\hat{\mu}_1(x) - \mu_1(x)| = o_{\mathcal{P}_s}(1), \quad \sup_{x \in \mathcal{X}} |\hat{\mu}_0(x) - \mu_0(x)| = o_{\mathcal{P}_s}(1). \quad (\text{A.7})$$

Proof for Lemma A.1: For (a), note that $|\ln(p(x)/(1-p(x)))|_s \leq C$ and C does not depend on P because of Assumption A.1 (iii) and (iv) and the property of $\ln(\cdot)$. Hence, (a) holds by Theorem 8 of Lorentz (1986). By the same argument for (29) of HIR, we have the uniform version of it as

$$\sup_{P \in \mathcal{P}_s} \sup_{x \in \mathcal{X}} |p(x) - L(R^K(x)' \pi_K)| < CK^{-s/d_x}, \quad (\text{A.8})$$

because $L'(z)$ is bounded by $1/4$ and $p(x) = L(\ln(p(x)/(1-p(x))))$. Similarly, we define

$$\begin{aligned} Q^*(\pi) &= E_P \left[p(X) \ln [L(R^K(X)' \pi)] + (1-p(X)) (1 - \ln [L(R^K(X)' \pi)]) \right], \\ Q_K(\pi) &= E_P \left[L(R^K(X)' \pi_K) \ln [L(R^K(X)' \pi)] + (1 - L(R^K(X)' \pi_K)) (1 - \ln [L(R^K(X)' \pi)]) \right], \end{aligned} \quad (\text{A.9})$$

where we again suppress the dependence of $Q^*(\pi)$, $Q_K(\pi)$, π_K on P . By definition, $\pi_K^* = \arg \max_{\pi} Q^*(\pi)$ and $\pi_K = \arg \max_{\pi} Q(\pi)$. Let $\eta = \inf_{P \in \mathcal{P}_s} \inf_{x \in \mathcal{X}} p(x)(1-p(x))$ and by Assumption A.1(iv), $\eta > 0$. Define Π_K as

$$\Pi_K = \left\{ \pi \in \mathbb{R}^K : \inf_{x \in \mathcal{X}} \ln [L(R^K(X)' \pi)] \left(1 - \ln [L(R^K(X)' \pi)] \right) \geq \eta/2 \right\}. \quad (\text{A.10})$$

Because of (A.8), for K large, we have $\pi_K(P) \in \Pi_K$ for all $P \in \mathcal{P}_s$. Note that because we choose $\int_{\mathcal{X}} R^K(x)R^K(x)'dx = I_K$ and by Assumption A.1(ii), it is true that

$$\lambda_{\min}\left(E_P[L'(R^K(X)'\pi)R^K(x)R^K(x)']\right) \geq \delta \cdot \eta/2 \quad (\text{A.11})$$

uniformly over $P \in \mathcal{P}_s$ where $\lambda_{\min}(A)$ denotes the minimum eigenvalue of a matrix A . Then following HIR, it is true that for some fixed $C_1 > 0$

$$\sup_{P \in \mathcal{P}_s} \sup_{\pi \in \Pi_K} \left| Q^*(\pi) - Q_K(\pi) \right| \leq C_1 K^{-s/d_x}. \quad (\text{A.12})$$

Let $C_2 > 0$ and $C_2 = \sqrt{4C_2/(\delta\eta)}$ for C_1 in (A.12), then when K is large enough, we have for all $P \in \mathcal{P}_s$, $\tilde{\Pi}_K(P) \subset \Pi_K$ where

$$\tilde{\Pi}_K(P) = \{\pi \in \mathbb{R}^K : \|\pi - \pi_K(P)\|\}. \quad (\text{A.13})$$

Next, by the same argument in HIR, we have the local maximum of $Q^*(\pi)(P)$ is in the interior of $\tilde{\Pi}_K(P)$ for all $P \in \mathcal{P}_s$. This shows Lemma A.1(b) which implies that

$$\begin{aligned} \sup_{P \in \mathcal{P}_s} \sup_{x \in \mathcal{X}} \left| L(R^K(x)'\pi_K) - L(R^K(x)'\pi_K^*) \right| &\leq \|R^K(x)\| \|\pi_K - \pi_K^*\| \\ &= O(\zeta(K)K^{-s/(2d_x)}). \end{aligned} \quad (\text{A.14})$$

By triangle inequality, (A.8) and (A.14) together imply Lemma A.1(c). \square

Proof for Lemma A.2: Let $S_K(P) = E_P[R^K(X)R^K(X)']$ whose smallest eigenvalue is uniformly bounded away from 0 by Assumption A.1(ii). It is true that

$$\tilde{\zeta}(K) = \sup_{x \in \mathcal{X}} \left| S_K(P)^{-1/2} R^K(x) \right| \leq C\zeta(K) \quad (\text{A.15})$$

uniformly over $P \in \mathcal{P}_s$ and $E_P[S_K(P)^{-1/2}R^K(x)R^K(x)'S_K(P)^{-1/2}] = I_K$ for all $P \in \mathcal{P}_s$. Hence, by replacing $R^K(x)$ with $S_K(P)^{-1/2}R^K(x)$ for each $P \in \mathcal{P}_s$, it is without loss of generality to assume that $S_K(P) = I_K$ for all $P \in \mathcal{P}_s$. Similar to (A.1) of Newey (1997), we have

$$\|\hat{S}_K(P) - I\| = O_{\mathcal{P}_s}\left(\zeta(K)\sqrt{\frac{K}{N}}\right), \quad (\text{A.16})$$

where

$$\hat{S}_K(P) = \frac{1}{N} \sum_{i=1}^N R^K(X_i)R^K(X_i)'$$

Note that the rest of the argument of the proof of Lemma 2 of HIR holds uniformly over $P \in \mathcal{P}_s$ under Assumption A.1, so Lemma A.2 follows. \square

Proof for Lemma A.3: Lemma A.1 and Lemma A.2 together imply Lemma A.3. \square

Proof for Lemma A.4: We show the $\mu_1(x)$ case and the proof for $\mu_0(x)$ is similar. Define

$$\begin{aligned}\Phi_K &= \frac{1}{N} \sum_{i=1}^N \frac{D_i Y_i}{\hat{p}(X_i)} R^K(X_i), \quad \widehat{\Phi}_K = \frac{1}{N} \sum_{i=1}^N \frac{D_i Y_i}{\widehat{\hat{p}}(X_i)} R^K(X_i), \\ \widehat{S}_K &= \frac{1}{N} \sum_{i=1}^N R^K(X_i) R^K(X_i)'\end{aligned}$$

then it is true that $\hat{\mu}_1(x) = \widehat{\Phi}'_K \widehat{S}_K^{-1} R^K(x)$. By similar proof of Lemma A.3, it is true that

$$\sup_{x \in \mathcal{X}} \left| \Phi'_K \widehat{S}_K^{-1} R^K(x) - \mu_1(x) \right| = O_{\mathcal{P}_s} \left(N \left[\sqrt{K/N} + K^{-s_1/(2d_x)} \right] \right) = o_{\mathcal{P}_s}(1). \quad (\text{A.17})$$

By the similar proof of Theorem 2 of HIR, it is true that

$$\sup_{x \in \mathcal{X}} \left| \Phi'_K \widehat{S}_K^{-1} R^K(x) - \hat{\mu}_1(x) \right| = o_{\mathcal{P}_s}(1). \quad (\text{A.18})$$

Then by triangle inequality, (A.17) and (A.18) imply Lemma A.4. \square

B Proof for the Theorems

Proof of Theorem 5.1: Our proof is similar to that of Theorem 6.3 of Donald and Hsu (2010). Let \mathcal{H}_1 denote the set of all functions from \mathcal{L} to $[-\infty, 0]$. Let $h = (h_1, h_2)$, where $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$, and define

$$T(h) = \sup_{\ell \in \mathcal{L}} (\Psi_{h_2}(\ell) + h_1(\ell)).$$

Define $c_0(h_1, h_2, 1 - \alpha)$ as the $(1 - \alpha)$ -th quantile of $T(h)$.

Similar to Lemma A2 of AS, we can show that for any $\xi > 0$,

$$\limsup_{N \rightarrow \infty} \sup_{\{P \in \mathcal{P}_0: h_{2,P} \in \mathcal{H}_{2,cpt}\}} P \left(\widehat{S}_N > c_0(h_{1,N}^P, h_{2,P}, 1 - \alpha) + \xi \right) \leq \alpha, \quad (\text{B.1})$$

where $h_{1,N}^P = \sqrt{N} \nu_P(\cdot)$ and $h_{1,N}^P$ belongs to \mathcal{H}_1 under $P \in \mathcal{P}_0$. Also, similar to Lemma A3 of AS, we can show that for all $\alpha < 1/2$

$$\limsup_{N \rightarrow \infty} \sup_{\{P \in \mathcal{P}_0: h_{2,P} \in \mathcal{H}_{2,cpt}\}} P \left(c_0(\phi_N, h_{2,P}, 1 - \alpha) < c_0(h_{1,N}^P, h_{2,P}, 1 - \alpha) \right) = 0. \quad (\text{B.2})$$

As a result, to complete the proof of Theorem 5.1, it suffices to show that for all $0 < \delta < \eta$

$$\limsup_{N \rightarrow \infty} \sup_{\{P \in \mathcal{P}_0: h_{2,P} \in \mathcal{H}_{2,cpt}\}} P \left(\hat{c}_\eta < c_0(\phi_N, h_{2,P}, 1 - \alpha) + \xi \right) = 0. \quad (\text{B.3})$$

Let $\{P_N \in \mathcal{P}_0 | N \geq 1\}$ be a sequence for which the probability in the statement of (B.3) evaluated at P_N differs from its supremum over $P \in \mathcal{P}_0$ by δ_N or less, where $\delta_N > 0$ and $\lim_{N \rightarrow \infty} \delta_N = 0$. By the definition of \limsup , such sequence always exists. Therefore, it is equivalent to show that for $0 < \xi < \eta$,

$$\lim_{N \rightarrow \infty} P \left(\hat{c}_{N,\eta} < c_0(\phi_N, h_{2,P}, 1 - \alpha) + \xi \right) = 0, \quad (\text{B.4})$$

where $\hat{c}_{N,\eta}$ denotes the critical value under P_N . To be more specific, we know that quantity on the left hand side exists, but we want to show that it is 0. Given that we restrict to a compact set $\mathcal{H}_{2,cpt}$, there exists a subsequence k_N of N such that $h_{2,P_{k_N}}$ converges to h_2^* for some $h_2^* \in \mathcal{H}_{2,cpt}$. By Lemma 4.3,

$$\Psi_{k_N}^u(\cdot) \xrightarrow{\mathbb{P}} \Psi_{h_2^*}(\cdot).$$

By the definition of $\xrightarrow{\mathbb{P}}$, there exists a further subsequence m_N of k_N such that

$$\Psi_{m_N}^u(\cdot) \xrightarrow{a.s.} \Psi_{h_2^*}(\cdot).$$

For any $\omega \in \{\omega \in \Omega | \Psi_{m_N}^u(\cdot)(\omega) \Rightarrow \Psi_{h_2^*}(\cdot)\} \equiv \Omega_1$, by the same argument for Theorem 1 of AS we can show that for any constant $a_{m_N} \in \mathbb{R}$ which may depends on h_1 and P and for any $0 < \xi_1$,

$$\begin{aligned} \limsup_{N \rightarrow \infty} \sup_{h_1 \in \mathcal{H}_1} P_U \left(\sup_{\ell \in \mathcal{L}} (\Psi_{m_N}^u(\ell)(\omega) + h_1) \leq a_{\ell_N} \right) \\ - P \left(\sup_{\ell \in \mathcal{L}} (\Psi_{h_2^*}(\ell) + h_1) \leq a_{\ell_N} + \xi_1 \right) \leq 0. \end{aligned} \quad (\text{B.5})$$

(B.5) is similar to (12.28) in AS. By (B.5) and by the similar argument for Lemma A5 of AS, we have that for all $0 < \xi < \xi_1 < \eta$,

$$\liminf_{N \rightarrow \infty} \hat{c}_{N,\eta}(\omega) \geq c_0(\phi_{m_N}, h_{2,P_{m_N}}, 1 - \alpha) + \xi_1. \quad (\text{B.6})$$

Therefore, for any $\omega \in \Omega_1$, (B.6) holds. Given that $P(\Omega_1) = 1$, we have that for all $0 < \xi < \xi_1 < \eta$

$$P \left(\left\{ \omega | \liminf_{N \rightarrow \infty} \hat{c}_\eta(\omega) \geq c_0(\phi_{m_N}, h_{2,P_{m_N}}, 1 - \alpha) + \xi_1 \right\} \right) = 1,$$

which implies that

$$\lim_{N \rightarrow \infty} P(\hat{c}_{N,\eta} < c_0(\phi_{m_N}, h_{2,P_{m_N}}, 1 - \alpha) + \delta) = 0. \quad (\text{B.7})$$

Note that for any convergent sequence A_N , if there exists a subsequence A_{ℓ_N} converging to A , then A_N converges to A as well. Therefore, (B.7) is sufficient for (B.4). Theorem 5.1(a) is shown by combining (B.1), (B.2) and (B.3).

To show Theorem 5.1(b), note that if under P_c $\lambda(\{x \in \mathcal{X} : \mu_1(x) = \mu_0(x)\}) > 0$, then it is true that there exists an open ball around some point $x^* \in \mathcal{X}_0 \equiv \{x \in \mathcal{X} : \mu_1(x) = \mu_0(x)\}$, $\mathcal{N}(x^*)$, such that $\mathcal{N}(x^*) \subseteq \mathcal{X}_0$. This implies that $\lambda(\mathcal{L}_0) > 0$ under P_c where $\mathcal{L}_0 \equiv \{\ell : \nu(\ell) = 0\}$. Then by the same proof based on the pointwise asymptotics as in the proof of Proposition 1 of Barrett and Donald (2003) and Lemma 1 of Donald and Hsu (2010), we have $\widehat{S}_N \xrightarrow{d} \sup_{\ell \in \mathcal{L}_0} \Psi_{h_{2,P_c}}(\ell)$ whose CDF $H(a)$ is continuous and strictly increasing $a \in (0, \infty)$ and $H(0) > 1/2$.

By the same proof for Theorem 2(b) of AS, it is true that $\hat{c}_\eta \rightarrow c(1-\alpha+\eta)+\eta$ where $c(1-\alpha+\eta)$ denotes the $(1-\alpha+\eta)$ -th quantile of $\sup_{\ell \in \mathcal{L}_0} \Psi_{h_{2,P_c}}(\ell)$. Because, $H(a)$ is continuous at $c(1-\alpha)$, we have $\lim_{\eta \rightarrow 0} c(1-\alpha+\eta)+\eta = c(1-\alpha)$. This suffices to show that $\lim_{N \rightarrow \infty} P(\widehat{S}_N > \hat{c}_\eta) = \alpha$ under P_c and Theorem 5.1(b) holds. \square

Proof of Theorem 5.2: Under any fixed alternative, there exists ℓ^* such that $\nu(\ell^*) > 0$. so $\widehat{S}_N/\sqrt{N} \geq \nu(\ell^*)$ in probability that implies that \widehat{S}_N will diverge to positive infinity in probability. Also, the \widehat{c}_η is bounded in probability, so $\lim_{N \rightarrow \infty} P(\widehat{S}_N > \widehat{c}_\eta) = 1$. \square

Proof of Theorem 5.3: Define $\mathcal{L}^\circ \equiv \{\ell : E[-g_\ell(X)(\mu_1(X) - \mu_0(X))] = 0\}$, $\mathcal{L}^{++} = \{\ell : E[-g_\ell(X)(\mu_{1,N}(X) - \mu_{0,N}(X))] > 0 \text{ eventually}\}$ and $d(\ell) = E[-\delta(X)g_\ell(X)]$. Under Assumption 5.7, it is straightforward to show that $\mathcal{L}^{++} \subseteq \mathcal{L}^\circ$ with $\lambda(\mathcal{L}^{++}) > 0$. In addition, $d(\ell) \geq 0$ when $\ell \in \mathcal{L}^\circ$ and $d(\ell) > 0$ when $\ell \in \mathcal{L}^{++}$.

It can be shown that $\widehat{S}_N \xrightarrow{D} \sup_{\ell \in \mathcal{L}^\circ} (\Psi_{h_2^*}(\ell) + d(\ell))$ and $\widehat{c}_\eta \xrightarrow{P} c_\eta + \eta$ where c_η is the $(1 - \alpha + \eta)$ -th quantile of $\sup_{\ell \in \mathcal{L}^\circ} \Psi_{h_2^*}(\ell)$. Then limit of the local power is

$$P(\sup_{\ell \in \mathcal{L}^\circ} (\Psi_{h_2^*}(\ell) + d(\ell)) \geq c_\eta + \eta).$$

As a result, by the continuity of the $\sup_{\ell \in \mathcal{L}^\circ} \Psi_{h_2^*}(\ell) + d(\ell)$ and $\sup_{\ell \in \mathcal{L}^\circ} \Psi_{h_2^*}(\ell)$,

$$\lim_{\eta \rightarrow 0} P(\sup_{\ell \in \mathcal{L}^\circ} (\Psi_{h_2^*}(\ell) + d(\ell)) \geq c_\eta + \eta) = P(\sup_{\ell \in \mathcal{L}^\circ} (\Psi_{h_2^*}(\ell) + d(\ell)) \geq c),$$

where c is the $(1 - \alpha)$ -th quantile of $\sup_{\ell \in \mathcal{L}^\circ} \Psi_{h_2^*}(\ell)$.

By assumption, $d(\ell)$ are non-negative if $\ell \in \mathcal{L}^\circ$, so $\sup_{\ell \in \mathcal{L}^\circ} (\Psi(\ell) + d(\ell))$ first order stochastically dominates $\sup_{\ell \in \mathcal{L}^\circ} \Psi(\ell)$ and it follows that

$$P(\sup_{\ell \in \mathcal{L}^\circ} (\Psi(\ell) + d(\ell)) \geq c) \geq \alpha,$$

which completes the proof. \square

Before showing Theorem 6.4, we present some notations related to LW's test. Let

$$K_*(t) = \int K(u)K(u+t)du, \quad \rho_0(t) = \frac{K_*(t)}{K_*(0)}.$$

For any x , we define

$$\begin{aligned} \gamma(x) &= \frac{E[Y^2|X=x, D=1] - E[Y|X=x, D=1]^2}{p(x)f(x)} \\ &\quad + \frac{E[Y^2|X=x, D=0] - E[Y|X=x, D=0]^2}{(1-p(x))f(x)}, \\ \rho_2(x, t) &= \gamma(x)K_*(t). \end{aligned}$$

For any measurable subset $B \subseteq \mathcal{X}$, we define

$$\begin{aligned} a_N(B) &= \frac{h^{-k/2}}{\sqrt{2\pi}} \int_B \sqrt{\rho_2(x, 0)} w(x) dx, \\ \sigma_N^2(B) &= \int_{-1}^1 F(\rho_0(t)) dt \int_B \rho_2(x, 0) w^2(x) dx, \\ F(\rho) &= \text{Cov}(\max\{\sqrt{1-\rho}Z_1 + \rho Z_2\}, \max\{Z_2, 0\}), \end{aligned}$$

where \mathbb{Z}_1 and \mathbb{Z}_2 are mutually independent standard normal distributions. Define

$$\begin{aligned}\hat{\gamma}_1(x) &= \sum_{i:D_i=1} \frac{Y_i^2 K_h(x - X_i)}{N \hat{\varrho}_1^2(x)} + \sum_{i:D_i=0} \frac{Y_i^2 K_h(x - X_i)}{N \hat{\varrho}_0^2(x)}, \\ \hat{\gamma}_2(x) &= \sum_{i:D_i=1} \sum_{j:D_j=1} \frac{Y_i Y_j K_h(x - X_i) K_h(x - X_j)}{N^2 \hat{\varrho}_1^3(x)} + \sum_{i:D_i=0} \sum_{j:D_j=0} \frac{Y_i Y_j K_h(x - X_i) K_h(x - X_j)}{N^2 \hat{\varrho}_0^3(x)}, \\ \hat{\varrho}_1(x) &= \frac{\sum_{i:D_i=1} K_h(x - X_i)}{N}, \quad \hat{\varrho}_0(x) = \frac{\sum_{i:D_i=0} K_h(x - X_i)}{N}.\end{aligned}$$

Hence, $\gamma(x)$, $\rho_2(x, t)$, $a_N(B)$ and $\sigma_N^2(B)$ are estimated by

$$\begin{aligned}\hat{\gamma}(x) &= \hat{\gamma}_1(x) - \hat{\gamma}_2(x), \quad \hat{\rho}_2(x, t) = \hat{\gamma}(x) K_*(t), \\ \hat{a}_N(B) &\equiv \frac{h^{-k/2}}{\sqrt{2\pi}} \int_B \sqrt{\hat{\rho}_2(x, 0)} w(x) dx, \quad \hat{\sigma}_N^2(B) \equiv \int_{-1}^1 F(\rho_0(t)) dt \int_B \hat{\rho}_2(x, 0) w^2(x) dx.\end{aligned}$$

We also define

$$\tilde{a}_N(B) = \int_B E[\max\{\delta(x) + h^{-k/2} \sqrt{\rho_2(x, 0)} \mathbb{Z}, 0\}] w(x) dx$$

where \mathbb{Z} is standard normal.

Proof of Theorem 6.4: First, under Assumption 6.8 and following LW's argument, we can show that

$$\frac{\hat{T} - \tilde{a}_N(\mathcal{W}_x^o)}{\sigma_N(\mathcal{W}_x^o)} \xrightarrow{D} \mathcal{N}(0, 1) \tag{B.8}$$

By Lemma A.13 of LW, we have $\hat{a}_N(\mathcal{W}_x^o) = a_N(\mathcal{W}_x^o) + o_p(1)$ and by the definition of a_N and $\tilde{a}_N(\mathcal{W}_x^o)$, we have

$$a_N - a_N(\mathcal{W}_x^o) = h^{-k/2} \int_{\mathcal{W}_x \setminus \mathcal{W}_x^o} \sqrt{\rho_2(x, 0)} w(x) dx \rightarrow \infty. \tag{B.9}$$

Note that $\lambda(\mathcal{W}_x \setminus \mathcal{W}_x^o) > 0$ and $\rho_2(x, 0) > 0$, so $\int_{\mathcal{W}_x \setminus \mathcal{W}_x^o} \sqrt{\rho_2(x, 0)} w(x) dx$ is strictly positive. Also, $h^{-k/2} \rightarrow \infty$, so $a_N - a_N(\mathcal{W}_x^o) \rightarrow \infty$. That is, $a_N - a_N(\mathcal{W}_x^o)$ diverges to infinity at rate $h^{-k/2}$.

On the other hand, we have

$$\begin{aligned}& h^{k/2} (\tilde{a}_N(\mathcal{W}_x^o) - a_N(\mathcal{W}_x^o)) \\ &= \int_{\mathcal{W}_x^o} (E[\max\{h^{k/2} \delta(x) + \sqrt{\rho_2(x, 0)} \mathbb{Z}, 0\}] - \sqrt{\frac{\rho_2(x, 0)}{2\pi}}) w(x) dx \\ &= \int_{\mathcal{W}_x^o} (E[\max\{h^{k/2} \delta(x) + \sqrt{\rho_2(x, 0)} \mathbb{Z}, 0\}] - E[\max\{\sqrt{\rho_2(x, 0)} \mathbb{Z}, 0\}]) w(x) dx \rightarrow 0.\end{aligned} \tag{B.10}$$

Because $h^{k/2} \delta(x) \rightarrow 0$, the difference between the two expectations in the integral in the last two lines will converge to zero. By dominated convergence theorem, the last line holds. (B.10) implies that $\tilde{a}_N(\mathcal{W}_x^o) - a_N(\mathcal{W}_x^o)$ is $o(h^{-k/2})$. By (B.9) and (B.10), we have

$$a_N - \tilde{a}_N(\mathcal{W}_x^o) = (a_N - a_N(\mathcal{W}_x^o)) + (a_N(\mathcal{W}_x^o) - \tilde{a}_N(\mathcal{W}_x^o)) \rightarrow \infty.$$

By (B.8), we have $\widehat{T} - \tilde{a}_N(\mathcal{W}_x^o)$ is $O_p(1)$. As a result,

$$\widehat{T} - \hat{a}_N = (\widehat{T} - \tilde{a}_N(\mathcal{W}_x^o)) - (\tilde{a}_N(\mathcal{W}_x^o) - a_N) + o_p(1) \rightarrow -\infty,$$

which implies that

$$\widehat{S}_{LW} = \frac{\widehat{T} - \hat{a}_N}{\hat{\sigma}_0} \rightarrow -\infty.$$

As a result, $P(\widehat{S}_{LW} > Z_{1-\alpha}) \rightarrow 0$, which completes our proof. \square

C Proof for the Lemmas

Proof of Lemma 3.1: Note that under our assumptions, it is true that all the bounds in the addendum of HIR hold uniformly over $\ell \in \mathcal{L}$ and $P \in \mathcal{P}$ in our case and this is sufficient to show that

$$\sup_{\ell \in \mathcal{L}} \left| \sqrt{N}(\hat{\nu}(\ell) - \nu(\ell)) - \frac{1}{\sqrt{N}} \sum_{i=1}^N \psi_{\ell}(W_i) \right| = o_p(1).$$

Hence, Lemma 3.1 follows. \square

Proof of Lemma 3.2: For notational simplicity, we show it for sequence N and all the results go through with a_N in place of N . Let

$$\begin{aligned} f_{N,i}(\ell) &= \frac{1}{\sqrt{N}} \psi_{N,\ell}(W_i), \\ F_{N,i} &= \frac{1}{\sqrt{N}} \left| \frac{(1-D)Y}{1-p(X)} - \frac{DY}{p(X)} + (D-p(X)) \left(\frac{\mu_0(X)}{1-p(X)} + \frac{\mu_1(X)}{p(X)} \right) \right| + \frac{2M}{\sqrt{N}}. \end{aligned} \quad (\text{C.1})$$

The $2M$ comes from the fact that $|\nu(\ell)| \leq 2M$ by Assumption 3.3(vi). By Lemma 3.1, it is equivalent to show that

$$\sum_{i=1}^N f_{N,i}(\cdot) \Rightarrow \Psi_{h_2^*}(\cdot). \quad (\text{C.2})$$

We use Theorem 10.6 (functional central limit theorem) of Pollard (1990) to show (C.2). First, note that \mathcal{G}_{cube} is a Vapnik-Chervonenkis class of functions and by Lemma E1 of AS, then the triangular array of random processes $\{f_{N,i}(\ell)(\omega) : \ell \in \mathcal{L}, i \leq N, N \geq 1\}$ is manageable with respect to $\{F_{N,i}(\omega) : i \leq N, N \geq 1\}$ where $f_{N,i}(\ell)(\omega)$ denotes the realization of $f_{N,i}(\ell)$ at ω . Similar notation applies to other functions in the rest of the paper. Hence, (i) of Theorem 10.6 of Pollard (1990) holds. Let $\chi_N(\ell) \equiv \sum_{i=1}^N \psi_{N,\ell}(W_i)/\sqrt{N}$. Then $E_P[\chi_N(\ell_1)\chi_N(\ell_2)] = h_{2,P_N}(\ell_1, \ell_2)$ which converges to $h_2^*(\ell_1, \ell_2)$ for all $\ell_1, \ell_2 \in \mathcal{L}$, because $\lim_{N \rightarrow \infty} d(h_{2,P_{a_N}}, h_2^*) = 0$ by assumption. Therefore, (ii) of Theorem 10.6 of Pollard (1990) holds. Because $\lim_{N \rightarrow \infty} d(h_{2,P_{a_N}}, h_2^*) = 0$, (v) also holds. By Assumption 3.3, it is true that $E_{P_N}[|\sqrt{N}F_{n,i}|^{2+\delta}] < C$ uniformly and by (16.21)

of AS with $\eta = 1$ in their argument, (iii) follows. Last, by the same argument of (16.39) of AS, it is true that for any $\epsilon > 0$,

$$\sum_{i=1}^N E_{P_N} [F_{n,i}^2 \cdot 1(F_{n,i} > \epsilon)] \rightarrow 0.$$

Therefore, by Theorem 10.6 of Pollard (1990) and Lemma 3.1, we have $\sqrt{N}(\hat{\nu}_N(\cdot) - \nu_N(\cdot)) \Rightarrow \Psi_{h_2^*}(\cdot)$. \square

Proof of Lemma 4.3: We show it for sequence N and all the results go through with a_N in place of N . Rewrite

$$\begin{aligned} \Psi_N^u(\ell) &= \sum_{i=1}^N \frac{U_i}{\sqrt{N}} \psi_{N,\ell}(W_i) \\ &+ \sum_{i=1}^N \frac{U_i}{\sqrt{N}} \left(g_\ell(X_i) \left(\frac{D_i Y_i}{p(X_i)} - \frac{D_i Y_i}{1 - \hat{p}(X_i)} \right) \right) \\ &+ \sum_{i=1}^N \frac{U_i}{\sqrt{N}} \left(g_\ell(X_i) \left(\frac{(1 - D_i) Y_i}{1 - \hat{p}(X_i)} - \frac{(1 - D_i) Y_i}{1 - p(X_i)} \right) \right) \\ &+ \sum_{i=1}^N \frac{U_i}{\sqrt{N}} \left(g_\ell(X_i) (p(X_i) - \hat{p}(X_i)) \left(\frac{\hat{\mu}_0(X_i)}{1 - \hat{p}(X_i)} + \frac{\hat{\mu}_1(X_i)}{\hat{p}(X_i)} \right) \right) \\ &+ \sum_{i=1}^N \frac{U_i}{\sqrt{N}} \left(g_\ell(X_i) (D_i - p(X_i)) \left(\frac{\hat{\mu}_0(X_i)}{1 - \hat{p}(X_i)} + \frac{\hat{\mu}_1(X_i)}{\hat{p}(X_i)} - \frac{\mu_0(X_i)}{1 - p(X_i)} - \frac{\mu_1(X_i)}{p(X_i)} \right) \right) \\ &+ \sum_{i=1}^N \frac{U_i}{\sqrt{N}} (\hat{\nu}(\ell) - \nu(\ell)). \end{aligned} \tag{C.3}$$

The proof can be done by showing that $\sum_{i=1}^N U_i \psi_{N,\ell}(W_i) / \sqrt{N} \xrightarrow{\mathbb{P}} \Psi_{h_2^*}(\cdot)$ and other terms are $o_p(1)$ uniformly over $\ell \in \mathcal{L}$. To show $\sum_{i=1}^N U_i \psi_{N,\ell}(W_i) / \sqrt{N} \xrightarrow{\mathbb{P}} \Psi_{h_2^*}(\cdot)$, we need to show that for any subsequence a_N of N , there is a further subsequence of h_N of a_N such that $\sum_{i=1}^{h_N} U_i \psi_{h_N,\ell}(W_i) / \sqrt{h_N} \xrightarrow{a.s.} \Psi_{h_2^*}(\cdot)$. First, under Assumption 3.3, it is true that

$$\sup_{P \in \mathcal{P}} \sup_{\ell_1, \ell_2 \in \mathcal{L}} E_P [|\psi_{\ell_1}(W_i) \psi_{\ell_2}(W_i)|^{1+\delta/2}] < C$$

which implies that

$$\sup_{\ell_1, \ell_2 \in \mathcal{L}} \left| \frac{1}{N} \sum_{i=1}^N \psi_{N,\ell_1}(W_i) \psi_{N,\ell_2}(W_i) - h_2^*(\ell_1, \ell_2) \right| \xrightarrow{P} 0.$$

Also, let $F_{N,i}$ be defined in (C.1) and by law of large number, it is true that $\sum_{i=1}^N F_{N,i}^2 \xrightarrow{P} C_1$ where

$$C_1 = \lim_{N \rightarrow \infty} E_{P_N} \left[\left\{ \left| \frac{(1-D)Y}{1-p(X)} - \frac{DY}{p(X)} + (D-p(X)) \left(\frac{\mu_0(X)}{1-p(X)} + \frac{\mu_1(X)}{p(X)} \right) \right| + \frac{2M}{\sqrt{N}} \right\}^2 \right].$$

The limit of the right-hand side exists because of $d(h_{2,P_N}, h_2^*) \rightarrow 0$. Next, define the $\varrho_N(\epsilon) = \sum_{i=1}^N F_{N,i}^2 \cdot 1(F_{N,i} > \epsilon)$ for $\epsilon \in (0, r]$ for some $r > 0$. Then it is true that $\{\sum_{i=1}^N F_{N,i}^2 \cdot 1(F_{N,i} > \epsilon) : \epsilon \in$

$(0, r], i \leq N, N \geq 1\}$ is measurable with respect to $\{F_{N,i}^2 : i \leq N, N \geq 1\}$. Then by Lemma E2 of AS or Theorem 8.2 of Pollard (1990), it is true that

$$\sup_{\epsilon \in (0, r]} \varrho_N(\epsilon) \xrightarrow{P} 0. \quad (\text{C.4})$$

As a result, for any subsequence a_N of N , there exists further subsequence of h_N of a_N such that

$$\begin{aligned} \sup_{\ell_1, \ell_2 \in \mathcal{L}} \left| \frac{1}{h_N} \sum_{i=1}^{h_N} \psi_{h_N, \ell_1}(W_i) \psi_{h_N, \ell_2}(W_i) - h_2^*(\ell_1, \ell_2) \right| &\xrightarrow{a.s.} 0. \\ \sum_{i=1}^{h_N} F_{h_N, i}^2 &\xrightarrow{a.s.} C_1, \quad \sup_{\epsilon \in (0, r]} \varrho_{h_N}(\epsilon) \xrightarrow{a.s.} 0. \end{aligned}$$

Let

$$\begin{aligned} \Omega_h \equiv \left\{ \omega \in \Omega : \sup_{\ell_1, \ell_2 \in \mathcal{L}} \left| \frac{1}{h_N} \sum_{i=1}^{h_N} \psi_{h_N, \ell_1}(W_i) \psi_{h_N, \ell_2}(W_i) - h_2^*(\ell_1, \ell_2) \right|(\omega) \rightarrow 0 \right. \\ \left. \sum_{i=1}^{h_N} F_{h_N, i}^2(\omega) \rightarrow C_1, \quad \sup_{\epsilon \in (0, r]} \varrho_{h_N}(\epsilon)(\omega) \rightarrow 0 \right\} \end{aligned} \quad (\text{C.5})$$

It is obvious that $P(\Omega_h) = 1$. We claim that for all $\omega \in \Omega_h$, $\sum_{i=1}^{h_N} U_i \psi_{h_N, \ell}(W_i)(\omega) / \sqrt{h_N} \Rightarrow \Psi_{h_2^*}(\cdot)$. Let $(\Omega_u, F_u, \mathbb{P}_u)$ be the probability space of U^∞ . Conditional on any $\omega \in \Omega_h$, define

$$\begin{aligned} g_{h_N, i}(\ell | \omega) &= \frac{U_i}{\sqrt{h_N}} \psi_{h_N, \ell}(W_i)(\omega), \quad (\text{C.6}) \\ G_{h_N, i}(\omega) &= \frac{U_i}{\sqrt{h_N}} \left| \frac{(1-D)Y}{1-p(X)} - \frac{DY}{p(X)} + (D-p(X)) \left(\frac{\mu_0(X)}{1-p(X)} + \frac{\mu_1(X)}{p(X)} \right) \right|(\omega) + \frac{2M}{\sqrt{N}} \end{aligned}$$

It is true that the triangular array $\{g_{h_N, i}(\omega_u, \ell | \omega) : \ell \in \mathcal{L}, i \leq h_N, h_N \geq 1\}$ is manageable with respect to $G_{h_N, i}(\omega_u, \ell | \omega) : i \leq h_N, h_N \geq 1\}$ by the similar argument in Lemma 3.2. Hence, (i) of Theorem 10.6 of Pollard (1990) holds. Define $\chi_{h_N}^u(\ell) = \sum_{i=1}^{h_N} g_{h_N, i}(\omega_u, \ell | \omega)$. Then $E_U[\chi_{h_N}^u(\ell_1) \chi_{h_N}^u(\ell_2)] = \sum_{i=1}^{h_N} \psi_{h_N, \ell_1}(W_i | \omega) \psi_{h_N, \ell_2}(W_i | \omega) / h_N$ where E_U denotes the expectation taken over U_i 's. By (C.5), it is true that $E_U[\chi_{h_N}^u(\ell_1) \chi_{h_N}^u(\ell_2)] \rightarrow h_2^*(\ell_1, \ell_2)$ uniformly over $\ell_1, \ell_2 \in \mathcal{L}$. Hence, (ii) and (v) of Theorem 10.6 of Pollard (1990) hold. We have

$$\sum_{i=1}^{h_N} E_U[G_{h_N, i}^2(\ell | \omega)] = \sum_{i=1}^{h_N} F_{h_N, i}^2(\omega) \rightarrow C_1,$$

by (C.5). This is sufficient for (iii) of Theorem 10.6 of Pollard (1990). For any $\epsilon^* > 0$,

$$\sum_{i=1}^{h_N} E_U[G_{h_N, i}^2(\ell | \omega) \cdot 1(G_{h_N, i}^2 > \epsilon^*)] = \varrho_{h_N}(\epsilon^*)(\omega) \rightarrow 0 \quad (\text{C.7})$$

by (C.5) and the fact that $\varrho_{h_N}(\epsilon^*)(\omega)$ is non-increasing on $\epsilon^* \in (0, \infty)$. Therefore, for all $\omega \in \Omega_h$ with $P(\Omega_h) = 1$, we have $\Psi_{h_N}^u(\cdot)(\omega) \Rightarrow \Psi_{h_2^*}(\cdot)$ by Theorem 10.6 of Pollard (1990). That is, for any subsequence a_N of N , there exists a further subsequence h_N of a_N such that $\Psi_{h_N}^u(\cdot) \xrightarrow{a.s.} \Psi_{h_2^*}(\cdot)$ which shows $\Psi_N^u(\cdot) \xrightarrow{P} \Psi_{h_2^*}(\cdot)$.

Similar argument applies to the second term to the fifth term of (C.3). The only difference is that the limiting covariate kernel of those processes are 0, i.e., $h_2(\ell_1, \ell_2) = 0$ for all $\ell_1, \ell_2 \in \mathcal{L}$. These imply that these processes are $o_p(1)$ uniform over $\ell \in \mathcal{L}$.

For the last term in (C.3),

$$P\left(\sup_{\ell \in \mathcal{L}} \left| (\hat{\nu}_N(\ell) - \nu_N(\ell)) \frac{1}{\sqrt{N}} \sum_{i=1}^N U_i \right| > \epsilon\right) \leq \frac{E[\sup_{\ell \in \mathcal{L}} |\hat{\nu}_N(\ell) - \nu_N(\ell)|]^2}{\epsilon^2} \rightarrow 0.$$

The first inequality follows from the Chebyshev's inequality and $E[\sum_{i=1}^N U_i / \sqrt{N}]^2 = 1$. By dominated convergence theorem, we also have

$$E\left[\sup_{\ell \in \mathcal{L}} |\hat{\nu}_N(\ell) - \nu_N(\ell)|\right]^2 \rightarrow 0,$$

which is sufficient to show that the last term is $o_p(1)$ uniformly over $\ell \in \mathcal{L}$.

Lemma 4.3 is implied by these results. \square

References

- Abadie, A. (2002). “Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models,” *Journal of the American Statistical Association*, **97**, 284–292.
- Abadie, A. (2003). “Semiparametric Instrument Variable Estimation of Treatment Response Models,” *Journal of Econometrics*, **113**, 231–263.
- Abadie, A., J. Angrist and G. W. Imbens (2002). “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings,” *Econometrica*, **70**, 91–117.
- Abrevaya, J., Hsu, Y.-C. and R. P. Lieli (2011). “Nonparametric Estimation of Conditional Average Treatment Effects,” Working Paper.
- Andrews, D. W. K. (1991) “Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models,” *Econometrica*, **59**, 307–345.
- Andrews, D. W. K., and X. Shi (2011). “Inference Based on Conditional Moment Inequalities,” Working Paper.
- Andrews, D. W. K., and X. Shi (2012). “Nonparametric Inference Based on Conditional Moment Inequalities,” Working Paper.
- Andrews, D. W. K., and G. Soares (2010). “Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection,” *Econometrica*, **78**, 119–157.
- Armstrong T. B. (2011a). “Asymptotically Exact Inference in Conditional Moment Inequality Models,” Working Paper.
- Armstrong T. B. (2011b). “Weighted KS Statistics for Inference on Conditional Moment Inequalities,” Working Paper.
- Barrett, G. F., and S. G. Donald (2003). “Consistent Tests for Stochastic Dominance,” *Econometrica*, **71**, 71–104.
- Chernozhukov, V., S. Lee and A. Rosen (2008): “Inference with Intersection Bounds,” Working Paper.
- Crump, R. K., V. J. Hotz, G. W. Imbens and O. A. Mitnik (2008) “Nonparametric Tests for Treatment Effect Heterogeneity,” *Review of Economics and Statistics*, **90**, 389–405.
- Delgado, M. A. , and J. C. Escanciano (2011). “Conditional Stochastic Dominance Testing,” Working Paper.
- Donald, S. G., and Y.-C. Hsu (2010). “Improving the Power of Tests of Stochastic Dominance,” Working Paper.
- Donald, S. G., and Y.-C. Hsu (2011). “Estimation and Inference for Distribution Functions and Quantile Functions in Treatment Effect Models,” Working Paper.

- Donald, S. G., Y.-C. Hsu and R. P. Lieli (2011). “Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT,” Working Paper.
- Fan, T. (2008). “Confidence Sets for Parameters Defined by Conditional Moment Inequalities/Equalities,” Working Paper.
- Frölich, M. (2007). “Nonparametric IV Estimation of Local Average Treatment Effects with Covariates,” *Journal of Econometrics*, **139**, 35–75.
- Galichon, A., and M. Henry (2009). “Set Identification in Models with Multiple Equilibria,” Working Paper.
- Hahn, J. (1988). “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, **66**, 315–331.
- Hansen, P. R. (2005). “A Test for Superior Predictive Ability,” *Journal of Business and Economic Statistics*, **23**, 365–380.
- Heckman, J., H. Ichimura, J. Smith and P. Todd (1998). “Characterizing Selection Bias Using Experimental Data,” *Econometrica*, **66**, 1017–1098.
- Heckman, J., H. Ichimura and P. Todd (1997). “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program,” *Review of Economic Studies*, **64**, 605–654.
- Heckman, J., H. Ichimura and P. Todd (1998). “Matching As An Econometric Evaluations Estimator,” *Review of Economic Studies*, **65**, 261–294.
- Hirano, K., G. W. Imbens and G. Ridder (2003). “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, **71**, 1161–1189.
- Ichimura, H., and O. Linton (2005). “Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators,” in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* ed. by Andrews, D.W.K. and Stock, J., Cambridge University Press.
- Imbens, G. W. (2004). “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” *Review of Economics and Statistics*, **86**, 4–29.
- Imbens, G. W., and J. D. Angrist (1994). “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, **62**, 467–475.
- Imbens, G. W., and J. W. Wooldridge (2009). “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, **47**, 5–86.
- Kim, K. (2008). “Set Estimation and Inference with Models Characterized by Conditional Moment Inequalities,” Working Paper.
- Lee, S., K. Song and J.-Y. Whang (2011). “Testing Functional Inequalities,” Working Paper.

- Lee, S., and J.-Y. Whang (2009). “Nonparametric Tests of Conditional Treatment Effects,” Cowles Foundation Discussion Papers 1740, Cowles Foundation, Yale University.
- Linton, O., K. Song and Y.-J. Whang (2010). “An improved Bootstrap Test of Stochastic Dominance,” *Journal of Econometrics*, **154**, 186–202.
- Newey, W. (1997). “Convergence Rates and Asymptotic Normality for Series Estimators,” *Journal of Econometrics*, **79**, 147–168.
- Pollard, D. (1990). *Empirical Processes: Theory and Application*, CBMS Conference Series in Probability and Statistics, Vol. 2. Hayward, CA: Institute of Mathematical Statistics.
- Rosenbaum, P., and D. Rubin (1983): “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, **70**, 41–55.
- Tsirel’son, V. S. (1975): “The Density of the Distribution of the Maximum of a Gaussian Process,” *Theory of Probability and its Application*, **16**, 847–856.
- van der Vaart, A. W., and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: With Application to Statistics*, New York: Springer-Verlag.

Table 1: Size Properties of Example 7.1.

	$N=300$	500	1000
Hsu's test	0.0566	0.0500	0.0522
LW's test with $C_h=2$	0.0810	0.0680	0.0672
LW's test with $C_h=3$	0.0662	0.0660	0.0618

Significance level is set at 5%.

Table 2: Size Properties of Example 7.2.

	$N=300$	500	1000
Hsu's test	0.0268	0.0354	0.0390
LW's test with $C_h=2$	0.0092	0.0060	0.0012
LW's test with $C_h=3$	0.0102	0.0060	0.0032

Significance level is set at 5%.

Table 3: Power Properties of Example 7.3.

	$N=300$	500	1000
Hsu's test	0.8274	0.9776	1.0000
LW's test with $C_h=2$	0.5050	0.7270	0.9642
LW's test with $C_h=3$	0.5636	0.7868	0.9806

Significance level is set at 5%.

Table 4: Local Power Properties of Example 7.4.

	$N=300$	500	1000
Hsu's	0.0444	0.0598	0.0732
LW's test with $C_h=2$	0.0228	0.0114	0.0068
LW's test with $C_h=3$	0.0222	0.0164	0.0076

Significance level is set at 5%.