# Testing Monotonicity of Mean Potential Outcomes in a Continuous Treatment with High-Dimensional Data

**Yu-Chin Hsu**[†]

Institute of Economics, Academia Sinica
Department of Finance, National Central University
Department of Economics, National Chengchi University and
CRETA, National Taiwan University

**Martin Huber**[*]

Department of Economics, University of Fribourg

**Ying-Ying Lee**[‡]

Department of Economics, University of California, Irvine

**Chu-An Liu**[§]

Institute of Economics, Academia Sinica

This version: July 2, 2023

[†] ychsu@econ.sinica.edu.tw, [*] martin.huber@unifr.ch, [‡] yingying.lee@uci.edu, [§] caliu@econ.sinica.edu.tw.

Acknowledgments: Yu-Chin Hsu gratefully acknowledges research support from the National Science and Technology Council of Taiwan (NSTC111-2628-H-001-001), the Academia Sinica Investigator Award of Academia Sinica, Taiwan (AS-IA-110-H01), and the Center for Research in Econometric Theory and Applications (107L9002) from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education of Taiwan. Chu-An Liu gratefully acknowledges research support from the Academia Sinica Career Development Award (AS-CDA-110-H02).

# Abstract

While most treatment evaluations focus on binary interventions, a growing literature also considers continuously distributed treatments. We propose a Cramér-von Mises-type test for testing whether the mean potential outcome given a specific treatment has a weakly monotonic relationship with the treatment dose under unconfoundedness. In a nonseparable structural model, applying our method amounts to testing the monotonicity of the average structural function in the continuous variable of interest. To flexibly control for a possibly high-dimensional set of covariates in our testing approach, we propose a double debiased machine learning estimator that accounts for covariates in a data-driven way. We show that the proposed test controls asymptotic size and is consistent against any fixed alternative. These theoretical findings are supported by Monte-Carlo simulations. As an empirical illustration, we apply our test to evaluate the Job Corps program and reject a weakly negative relationship between the treatment (hours in academic and vocational training) and labor market performance among relatively low treatment values.



**JEL classification:** C01, C12, C21

**Keywords:** Average dose response functions, average structural function, continuous treatment models, doubly robust, high dimension, hypothesis testing, machine learning, treatment monotonicity.

# 1 Introduction

Even though many studies on treatment or policy evaluation investigate the effects of binary or discrete interventions, a growing literature considers the assessment of continuously distributed treatments, e.g., hours spent in a training program whose effect on labor market performance is of interest (Flores et al., 2012; Kluve et al., 2012). Other examples include the efficacy of political advertisements on campaign contributions in Fong et al. (2018), nurse staffing on hospital readmissions penalties in Kennedy et al. (2017), among others. Most contributions like Imbens (2000), Hirano and Imbens (2004), Flores (2007), Flores et al. (2012), Galvao and Wang (2015), Lee (2018) and Colangelo and Lee (2022) focus on the identification and estimation of the average dose response function (ADF), which corresponds to the mean potential outcome as a function of the treatment dose. This permits assessing the average treatment effect (ATE) as the difference in the ADF assessed at two distinct treatment doses of interest, while Hirano and Imbens (2004), Flores et al. (2012), and Colangelo and Lee (2022) also consider the marginal effect of slightly increasing the treatment dose, which is the derivative of the ADF. Rather than considering the total effect of the treatment, Huber et al. (2020) suggest a causal mediation approach to disentangle the ATE into its direct effect and indirect effect operating through intermediate variables or mediators to assess the causal mechanisms of the treatment.

In this paper, we propose a method for testing whether the ADF has a weakly monotonic relationship with (i.e., is weakly increasing or decreasing in) the treatment dose under unconfoundedness, implying that a confounder of the treatment-outcome relation can be controlled for by observed covariates. Such a test appears interesting for verifying shape restrictions, e.g., whether increasing the treatment dose always has a non-negative effect, no matter what the baseline level of treatment is. Moreover, the treatment effect model is known to be equivalent to a nonseparable structural model of a nonseparable outcome with a general disturbance, as for instance Imbens and Newey (2009) and Lee (2018). In this case, the ADF corresponds to the average structural function in Blundell and Powell (2003). Therefore, our test can be applied to testing monotonicity of the average structural function in a nonseparable structural model under a conditional independence assumption. We also extend our test for the conditional ADF given a subset of the covariates.

To construct our test, we first transform the null hypothesis of a monotonic relationship to countably many moment inequalities based on the generalized instrumental function approach of Hsu et al. (2019) and Hsu and Shen (2020) that is a generalization of the instrumental function approach in Andrews and Shi (2013, 2014). We construct a Cramér-von Mises-type test statistic based on the estimated moments, which are shown to converge to a Gaussian process at the

regular root-$n$ rate. Importantly, by making use of moment inequalities, our method does not rely on the nonparametric estimation of the ADF or the partial effects, which would converge at slower nonparametric rates. To compute the critical values for our test, we apply a multiplier bootstrap method and the generalized moment selection (GMS) approach of Andrews and Shi (2013, 2014). We demonstrate that our test controls asymptotic size and is consistent against any fixed alternative.

To employ nonparametric or machine learning estimators in the presence of possibly high-dimensional nuisance parameters, we propose a double debiased machine learning (DML) estimator. Utilizing a doubly robust moment function based on a Neyman-type orthogonal score and cross-fitting, we give high-level conditions under which the nuisance function estimators do not affect the first-order large sample distribution of the DML estimator. Specifically, we give the high-level conditions on the standard mean-squared convergence rates of the first-step estimators, as the semiparametric models in Chernozhukov et al. (2018). The first-step estimators for the conditional expectation function and the conditional density can be kernel and series estimators, as well as modern ML methods, such as Lasso and deep neural networks. See Chernozhukov et al. (2018) and Athey and Imbens (2019) for potential ML methods, such as ridge, boosted trees, and various ensembles of these methods. As each ML method has its strengths and weaknesses depending on the data generating process and applications, it is desirable to flexibly employ various nuisance estimators. High-dimensional control variables are accommodated via the nuisance estimators; for example, Lasso allows the dimension of $X$ to grow with the sample size.

Our paper is related to a growing literature on testing monotonicity in regression problems such as Bowman et al. (1998), Ghosal et al. (2000), Gijbels et al. (2000), Hall and Heckman (2000), Dümbgen and Spokoiny (2001), Durot (2003), Baraud et al. (2005), Wang and Meyer (2011), Chetverikov (2019) and Hsu et al. (2019). The main difference between our GMS method and the previously suggested tests is that we rely on a two-step estimation procedure when computing the moments, with the first step consisting of estimating the generalized propensity score, i.e., the conditional density of a treatment dose given the covariates, and/or the conditional mean function. For this reason, it is necessary to take into account the behavior of the first step when we derive the limiting behavior of the estimated moment inequalities underlying our test. Rothenhäusler and Yu (2019) provide inference theory for a different *incremental causal effect* where the continuous treatment is slightly shifted across the whole population, while our average dose response function sets the continuous treatment at a given value across the whole population. A positive incremental causal effect does not imply that the average dose response function is increasing.

An alternative approach is to conduct uniform inference for the average dose response function $\mu(t)$ or the average partial effect $d\mu(t)/dt$ over a range of $t$; for example, linear functionals of

the conditional mean function using series in Belloni et al. (2015) and partitioning-based series in Cattaneo et al. (2020), binscatter for a partially linear model in Cattaneo et al. (2022), the Lasso estimator in Su et al. (2019) and DML estimator in Colangelo and Lee (2022) for a nonparametric nonseparable model. In particular, we compare our test with Su et al. (2019) in the simulation study. The supremum-type test is based on a uniform confidence band of $d\mu(t)/dt$, so its power can be driven by a large deviation of the null hypothesis at a specific point $t$. Our integral-type test generally has better power if the deviation of the null is more evenly spaced, e.g., when the data generating process with the violation of the null is the same for all $t$. Moreover, our test can have non-trivial local power against some $n^{-1/2}$ local alternatives as in Hsu et al. (2019), while the supremum-type test based on nonparametric estimators of $d\mu(t)/dt$ does not.

We investigate the finite sample behavior of the proposed test approach in a simulation study. As an empirical illustration, we apply our test to data from an experimental study on the Job Corps (see Schochet et al. (2001) and Schochet et al. (2008)) a program aimed at increasing the human capital of youths from disadvantaged backgrounds in the U.S. We consider hours in academic and vocational training in the first year of the program as the continuous treatment, and investigate its association with several labor market outcomes: weekly earnings in the fourth year, earnings and hours worked per week in quarter 16, and a binary employment indicator four years after assignment. For all outcomes, our test clearly rejects weakly negative monotonicity in the treatment when considering treatment doses between 40 and 3000 hours of training. In contrast, weakly positive monotonicity is not refuted at conventional levels of statistical significance. When, however, splitting the treatment range into three brackets of 40 to 1000, 1000 to 2000, and 2000 to 3000 hours, the test points to a violation of weakly negative monotonicity only in the lowest treatment bracket. In the remaining brackets, which have larger treatment values, we reject neither weakly positive nor weakly negative monotonicity. Our results are consistent with a concave ADF as, for instance, found in Flores et al. (2012), suggesting that the marginal effect of training on labor market performance is positive for relatively low treatment doses but decreases as hours of training increase. A potential explanation could be that participants attending more training in the first year might be induced to attain more education in the following years rather than to participate in the labor market.

The paper is organized as follows. Section 2 formulates the hypothesis of weak monotonicity to be tested. Section 3 proposes monotonicity tests under DML estimation. Section 4 presents a Monte-Carlo simulation and discusses how to choose the tuning parameters of the test in practice. Section 5 provides an empirical application to Job Corps data. Section 6 adapts the method to testing monotonicity with conditional mean potential outcomes given observed covariates. Section

7 concludes. The technical proofs are relegated to the Appendix.[1]

# 2 Monotonicity of Continuous Treatment Effect

We consider a nonparametric and model-free outcome equation $Y = g(T, X, \varepsilon)$. No functional form assumption is imposed on the unobserved disturbances $\varepsilon$, such as restrictions on dimensionality, monotonicity, or separability. Let $Y(t) = g(t, X, \varepsilon)$ denote the potential outcome corresponding to the level of treatment intensity $t \in \mathcal{T}$, where $\mathcal{T} = [a, b]$ with $-\infty < a < b < \infty$. Hirano and Imbens (2004) call $Y(t)$ the unit-level dose response function. Let $\mu(t) = E[Y(t)] = \int g(t, X, \varepsilon) dF_{X\varepsilon}$ for $t \in \mathcal{T}$ denote the expected value of the potential outcome, also known as the average dose response function (ADF) or the average structural function. In this paper, we test weather the ADF is weakly increasing in the treatment intensity within a specific range. We define the null hypothesis of interest as

$$H_0 : \ \mu(t_1) \geq \mu(t_2), \quad \text{for all } t_1 \geq t_2, \ \text{for } t_1, t_2 \in [t_\ell, t_u], \tag{2.1}$$

where $a \leq t_\ell < t_u \leq b$ so that $[t_\ell, t_u]$ is a convex and compact subset of $[a, b]$. Without loss of generality, we assume that $[t_\ell, t_u] = [0, 1]$.[2]

We apply the generalized instrumental function approach of Hsu et al. (2019) and Hsu and Shen (2020) to transform $H_0$ in (2.1) to countably many moment inequalities without loss of information. Specifically, suppose that $\mu(t)$ is a continuous function on $t = [0, 1]$ and $h(t)$ is a positive weighting function such that $\int_0^1 h(t)dt < \infty$. Then by Lemma 2.1 of Hsu and Shen (2020), $H_0$ in (2.1) is equivalent to

$$\frac{\int_{t_2}^{t_2+q^{-1}} \mu(s)h(s)ds}{\int_{t_2}^{t_2+q^{-1}} h(s)ds} - \frac{\int_{t_1}^{t_1+q^{-1}} \mu(s)h(s)ds}{\int_{t_1}^{t_1+q^{-1}} h(s)ds} \leq 0, \ \text{or} \tag{2.2}$$

$$\int_{t_2}^{t_2+q^{-1}} \mu(s)h(s)ds \cdot \int_{t_1}^{t_1+q^{-1}} h(s)ds - \int_{t_1}^{t_1+q^{-1}} \mu(s)h(s)ds \cdot \int_{t_2}^{t_2+q^{-1}} h(s)ds \leq 0 \tag{2.3}$$

for any $q = 2, \cdots$, and for any $t_1 \geq t_2$ such that $t_1, t_2 \in \{0, 1/q, 2/q, \cdots, 1 - 1/q\}$. Equations (2.2) and (2.3) hold by the fact that if a function is non-decreasing, then its weighted average over an interval will be non-decreasing as well when the interval moves to the right. In addition, following Hsu et al. (2019), Equations (2.2) and (2.3) contain the same information as the null hypothesis.

[1] An old version of this paper contains monotonicity tests under nonparametric and parametric estimations of the generalized treatment propensity score, which is available upon request.

[2] If $[t_\ell, t_u]$ is not $[0, 1]$, we can always apply an affine transformation $\phi$ on $t$ so that $\phi(t_\ell) = 0$ and $\phi(t_u) = 1$.

Therefore we transform $H_0$ in (2.1) to countably many moment inequalities based on which we will construct our test. Define

$$\mathcal{L} = \Big\{ \ell = (t_1, t_2, q^{-1}) : (t_1, t_2) \in [0,1]^2, t_1 > t_2, q = 2, 3, \cdots, \\ q \cdot (t_1, t_2) \in \{0, 1, 2, \cdots, q-1\}^2 \Big\}. \tag{2.4}$$

Choosing $h(t) = 1$ for simplicity, $H_0$ in (2.1) is equivalent to

$$H_0' : \ \nu(\ell) \equiv \nu_2(\ell) - \nu_1(\ell) \leq 0, \text{ for any } \ell = (t_1, t_2, q^{-1}) \in \mathcal{L}, \text{where} \tag{2.5}$$
$$\nu_j(\ell) \equiv \int_{t_j}^{t_j + q^{-1}} \mu(s) ds, \text{ for } j = 1, 2.$$

The set of the indicator functions of countable intervals $\mathcal{L}$ is also used in Hsu et al. (2019). This choice of $\mathcal{L}$ ensures that it is rich enough so there will be no loss of information when we transform $H_0$ to $H_0'$, and it is simple enough in order for a certain uniform central limit theory to apply.

The null hypothesis in (2.1) has a form that is similar to that in the literature on regression monotonicity; see for instance Hsu et al. (2019). However, the identification of $\mu(t)$ in our case is different from theirs. Next we discuss identification of $\nu(\ell)$.

**Assumption 2.1 (Unconfoundedness):** *$T$ and $\varepsilon$ are independent conditional on $X$.*

Assumption 2.1 is a commonly invoked identifying assumption based on observational data, also known as conditional independence, or selection on observables. It assumes that conditional on observables $X$, $T$ is as good as randomly assigned, or conditionally exogenous.

Since $\nu_j(\ell) \equiv \int_{t_j}^{t_j + q^{-1}} \mu(s) ds$ is a linear functional of $\mu$, its identification follows directly from the identification of $\mu$. The identification of $\mu$ has been established in the literature under Assumption 2.1; for example, in Colangelo and Lee (2022), for $t \in \mathcal{T}$, $\mu(t) = \int_{\mathcal{X}} E[Y|T = t, X] dF_X(X)$ that motivates the class of regression-based (or imputation) estimators. An alternative identifying moment function $\mu(t) = \lim_{h \to 0} E\left[K_h(T - t)Y / f_{T|X}(t|X)\right]$, where $K_h(T - t) \equiv k((T - t)/h)/h$ uses a suitable second-order symmetric kernel function $k(\cdot)$ with a bandwidth $h$, motivates the class of inverse probability weighting estimators.

We use an identifying moment function that is doubly robust or Neyman orthogonal, based on the Gateaux derivative of $\mu(t)$ derived in Colangelo and Lee (2022). We assume a sample $\{Z_i = (Y_i, T_i, X_i')'\}_{i=1}^n$, modeled as independent and identically distributed (i.i.d.) copies of $Z = (Y, T, X')'$, whose law is determined by the probability measure $P$ on $\mathcal{Z} \equiv \mathcal{Y} \times \mathcal{T} \times \mathcal{X}$. In Appendix

A, we give details of identifying $\nu_j(\ell) \equiv \int_{t_j}^{t_j+q^{-1}} \mu(s)ds = E[\phi_{j,q}(Z)]$, where the moment function

$$\phi_{j,q}(Z) \equiv \int_{t_j}^{t_j+q^{-1}} \gamma(s,X)ds + \frac{Y-\gamma(T,X)}{p(T,X)}\mathbf{1}(T \in [t_j, t_j+q^{-1}]) \tag{2.6}$$

with $\gamma(t,x) \equiv E[Y|T=t, X=x]$ and $p(t,x) \equiv f_{T|X}(t|x)$, for $j=1,2$.

# 3 DML Monotonicity Test

To deliver a reliable distributional approximation in practice, the double debiased ML (DML) method contains two key ingredients: a doubly robust moment function and cross-fitting. The doubly robust moment function reduces sensitivity in estimating $\nu(\ell)$ with respect to nuisance parameters.[3] Our DML estimator for $\nu(\ell)$ uses the moment function $\phi_{j,q}(Z)$ in (2.6). Cross-fitting removes bias induced by overfitting and achieves stochastic equicontinuity without strong entropy conditions. Our work builds on the results for semiparametric models in Ichimura and Newey (2022), Chernozhukov et al. (2018), Chernozhukov et al. (2018), and the nonparametric models for continuous treatments in Colangelo and Lee (2022).

Next we introduce our testing procedure and give a step-by-step algorithm for practical implementation. The first step is to estimate the nuisance functions $\gamma(t,x) \equiv E[Y|T=t, X=x]$ and $p(t,x)$ by a $K$-fold cross-fitting. The second step is to plug in the nuisance function estimates to the DML estimator $\hat{\nu}(\ell)$ of $\nu(\ell) = \nu_2(\ell) - \nu_1(\ell)$ defined in (2.6).

To test the null hypothesis $H_0'$, we make use of a Cramér-von Mises test statistic defined as

$$\widehat{T} = \sum_{\ell \in \mathcal{L}} \max\left\{\sqrt{n}\frac{\hat{\nu}(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)}, 0\right\}^2 Q(\ell), \tag{3.1}$$

where $\hat{\sigma}^2_{\nu,\epsilon}(\ell)$ is a consistent estimator for the asymptotic variance of $\sqrt{n}(\hat{\nu}(\ell) - \nu(\ell))$. A weighting function $Q$ satisfies $Q(\ell) > 0$ for all $\ell \in \mathcal{L}$ and $\sum_{\ell \in \mathcal{L}} Q(\ell) < \infty$.

We next define the simulated critical value for our test. We introduce a multiplier bootstrap method that can simulate a process that converges to the same limit as $\sqrt{n}(\hat{\nu}(\ell) - \nu(\ell))$. Under specific regularity conditions, we can show that the simulated process weakly converges to a Gaussian process conditional on the sample path $w.p.a.1$ and that this limiting Gaussian process

[3]Our estimator is doubly robust in the sense that it consistently estimates $\nu(\ell)$ if either one of the nuisance functions $E[Y|T,X]$ or $f_{T|X}$ is misspecified. The rapidly growing ML literature has utilized this doubly robust property to reduce regularization and modeling biases in estimating the nuisance parameters by ML or nonparametric methods; for example, Belloni et al. (2014), Farrell (2015), Belloni et al. (2017), Farrell et al. (2021), Chernozhukov et al. (2018), Chernozhukov et al. (2018), Rothe and Firpo (2019), and references therein.

corresponds to the limiting process of $\sqrt{n}(\hat{\nu}(\ell) - \nu(\ell))$. We adopt the GMS method to construct the simulated critical value.[4]

The algorithm below summarizes the implementation of our test.

Step 1. **(Nuisance functions)** For some fixed $K \in \{2, ..., n\}$, a $K$-fold cross-fitting partitions the observation indices into $K$ distinct groups $I_k$, $k = 1, ..., K$, such that the sample size of each group is the largest integer smaller than $n/K$. Let $n_k$ denote the number of observations in group $I_k$ for $k = 1, ..., K$. For $k \in \{1, ..., K\}$, the estimator $\hat{\gamma}_k(t, x)$ for $\gamma(t, x) \equiv E[Y|T = t, X = x]$ and the estimator $\hat{p}_k(t, x)$ for $p(t, x)$ use observations not in $I_k$ and satisfy Assumption 3.1.

Step 2. **(DML estimator)** $\hat{\nu}(\ell) = n^{-1} \sum_{i=1}^{n} \left(\hat{\phi}_{2,q}(Z_i) - \hat{\phi}_{1,q}(Z_i)\right)$, where for $j = 1, 2$,

$$\hat{\phi}_{j,q}(Z_i) = \int_{t_j}^{t_j + q^{-1}} \hat{\gamma}_{-i}(s, X_i) ds + \frac{Y_i - \hat{\gamma}_{-i}(T_i, X_i)}{\hat{p}_{-i}(T_i, X_i)} \mathbf{1}(T_i \in [t_j, t_j + q^{-1}]),$$

$\hat{\gamma}_{-i}(T_i, X_i) = \hat{\gamma}_k(T_i, X_i)$, and $\hat{p}_{-i}(T_i, X_i) = \hat{p}_k(T_i, X_i)$ for $i \in I_k$.[5]

Step 3. **(Test statistic)** $\hat{\sigma}^2_\nu(\ell) = n^{-1} \sum_{i=1}^{n} \hat{\phi}^2_\ell(Z_i)$, where

$$\hat{\phi}_\ell(Z_i) = \hat{\phi}_{2,q}(Z_i) - \hat{\phi}_{1,q}(Z_i) - \hat{\nu}(\ell). \tag{3.2}$$

$\hat{\sigma}_{\nu,\epsilon}(\ell) = \max\{\hat{\sigma}_\nu(\ell), \epsilon \cdot \hat{\sigma}_\nu(0, 1/2, 1/2)\}$, by which we manually bound the variance estimator away from zero. Compute the Cramér-von Mises test statistic $\widehat{T}$ in (3.1).

Step 4. **(Critical values)** Let $\{U_i : 1 \leq i \leq n\}$ be a sequence of i.i.d. random variables that satisfy Assumption 3.2. The simulated process is $\widehat{\Phi}^u_\nu(\ell) = n^{-1/2} \sum_{i=1}^{n} U_i \cdot \hat{\phi}_\ell(Z_i)$, where $\hat{\phi}_\ell(Z_i)$ is the estimated influence function given in (3.2).

$$\hat{\psi}_\nu(\ell) = -B_n \cdot 1\left(\sqrt{n} \cdot \frac{\hat{\nu}(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)} < -a_n\right),$$

where $a_n$ and $B_n$ satisfy Assumption 3.3. The critical value is

$$\hat{c}^\eta(\alpha) = \sup\left\{q \middle| P^u\left(\sum_{\ell \in \mathcal{L}} \max\left\{\frac{\widehat{\Phi}^u_\nu(\ell)}{\hat{\sigma}_{\nu,\epsilon}(\ell)} + \hat{\psi}_\nu(\ell), 0\right\}^2 Q(\ell) \leq q\right) \leq 1 - \alpha + \eta\right\} + \eta,$$

[4]The GMS approach is similar to the recentering method of Hansen (2005) and Donald and Hsu (2016), and the contact approach of Linton et al. (2010).

[5]An equivalent expression for the DML estimator is $\hat{\nu}(\ell) = K^{-1} \sum_{k=1}^{K} n_k^{-1} \sum_{i \in I_k} \hat{\phi}_{k,2,q}(Z_i) - \hat{\phi}_{k,1,q}(Z_i)$, where for $j = 1, 2$, $\hat{\phi}_{k,j,q}(Z_i) = \int_{t_j}^{t_j + q^{-1}} \hat{\gamma}_k(s, X_i) ds + (Y_i - \hat{\gamma}_k(T_i, X_i)/\hat{p}_k(T_i, X_i) \mathbf{1}(T_i \in [t_j, t_j + q^{-1}])$. We do not use this alternative expression to avoid the subscript $k$, in order to simplify the notation.

where $P^u$ denotes the multiplier probability measure given the observed samples.

Step 5. **(Decision rule)** Reject $H_0'$ if $\widehat{T} > \hat{c}^{\eta}(\alpha)$.

**Remark 3.1** We discuss the tuning parameters. The number of folds in cross-fitting $K$ is fixed and does not affect asymptotic theory. The choice of $K$ may affect the small-sample performance, as larger values of $K$ provide more observations in the training sample used to estimate the nuisance functions. We choose $K = 5$ in our empirical application following the recommendations in the DML literature, such as Chernozhukov et al. (2018). For the numerical integration in Step 2, we may choose $M = [n^{2/3}]$, where $[\cdot]$ is the nearest integer, so the condition $\sqrt{n}/M \to 0$ in Theorem 3.1 is satisfied. Let $N$ denote the expected sample size of the smallest cube corresponding to $q_1$. We choose $q_1$ such that $N = 50$ for all the sample sizes in the simulations. We set $a_n = 0.15 \cdot \ln(n)$, $B_n = 0.85 \cdot \ln(n)/\ln\ln(n)$, and $\eta = \epsilon = 10^{-6}$, as recommended in Hsu et al. (2019). We also try $a_n = \sqrt{0.3 \cdot \ln(n)}$ and $B_n = \sqrt{0.4 \cdot \ln(n)/\ln\ln(n)}$ as suggested by Andrews and Shi (2013, 2014) and Hsu and Shen (2020) in the simulations.

## 3.1 Asymptotic Size and Power

Let $\|\cdot\|_2$ denote the $L_2$-norm for the root-mean-square convergence rate, e.g., $\|\hat{\gamma}_k - \gamma\|_2 = \left(\int_{\mathcal{X}}\int_{\mathcal{T}} \left(\hat{\gamma}_k(t,x) - \gamma(t,x)\right)^2 f_{TX}(t,x)dtdx\right)^{1/2}$ and $\|\hat{p}_k - p\|_2 = \left(\int_{\mathcal{X}}\int_{\mathcal{T}} \left(\hat{p}_k(t,x) - p(t,x)\right)^2 f_{TX}(t,x)dtdx\right)^{1/2}$.

**Assumption 3.1** *For any $k \in \{1, ..., K\}$,*

*(i)* $\|\hat{\gamma}_k - \gamma\|_2 = o_P(1)$ *and* $\|\hat{p}_k - p\|_2 = o_P(1)$.

*(ii)* $\sqrt{n}\|\hat{\gamma}_k - \gamma\|_2\|\hat{p}_k - p\|_2 = o_P(1)$.

*(iii) There exists a positive constant $c$ such that* $\inf_{t\in\mathcal{T},x\in\mathcal{X}} p(t,x) \geq c$. *There exists a positive constant $c$ such that* $\sup_{t\in\mathcal{T},x\in\mathcal{X}} var(Y|T=t, X=x) \leq c$.

Assumptions 3.1(i) and (ii) are the typical conditions on the mean-squared convergence rates, as in Chernozhukov et al. (2018). Theorem 3.1 establishes the limiting behavior of the DML estimator for $\nu$.

**Theorem 3.1 (Uniform asymptotics)** *Let Assumptions 2.1 and 3.1 hold. Then uniformly over $\ell \in \mathcal{L}$, $\sqrt{n}(\hat{\nu}(\ell) - \nu(\ell)) = n^{-1/2}\sum_{i=1}^n \phi_\ell(Z_i) + o_P(1)$, where $\phi_\ell(Z_i) = \phi_{2,q}(Z_i) - \phi_{1,q}(Z_i) - \nu(\ell)$ defined in (2.6). Also, $\sqrt{n}(\hat{\nu}(\cdot) - \nu(\cdot)) \Rightarrow \Phi_{h_{DML}}(\cdot)$ where $\Phi_{h_{DML}}(\cdot)$ is a Gaussian process with variance-covariance kernel $h_{DML}(\ell_1, \ell_2) = E[\phi_{\ell_1}(Z)\phi_{\ell_2}(Z)]$.*

*Consider approximating* $\int_{t_j}^{t_j+q^{-1}} \hat{\gamma}_k(s, X_i)ds$ *by a numerical integration with a set of equally spaced grid points* $\{s_0 = t_\ell, s_1, ..., s_M = t_u\}$ *over* $[t_\ell, t_u]$, $M^{-1}\sum_{m=1}^{M} \hat{\gamma}_k(s_m, X_i)1(s_m \in [t_j, t_j + q^{-1}])$. *Let* $\sqrt{n}/M \to 0$, *and let the total variation of* $\hat{\gamma}_k$ *be smaller than a positive constant* $c$ *with probability approaching one (w.p.a.*1*), so that the approximation error of the numerical integration is asymptotically first-order ignorable.*

**Assumption 3.2** $\{U_i : 1 \leq i \leq n\}$ *is a sequence of i.i.d. random variables that is independent of the sample path of* $\{(Z_i) : 1 \leq i \leq n\}$ *such that* $E[U_i] = 0$, $E[U_i^2] = 1$, *and* $E[|U_i|^{2+\delta}] < C$ *for some* $\delta > 0$ *and* $C > 0$.

**Assumption 3.3** *(i)* $a_n$ *is a sequence of non-negative numbers satisfying* $\lim_{n\to\infty} a_n = \infty$ *and* $\lim_{n\to\infty} a_n/\sqrt{n} = 0$.

*(ii)* $B_n$ *is a sequence of non-negative numbers satisfying that* $B_n$ *is non-decreasing,* $\lim_{n\to\infty} B_n = \infty$ *and* $\lim_{n\to\infty} B_n/a_n = 0$.

**Theorem 3.2** *Suppose that Assumptions 2.1, 3.1, 3.2 and 3.3 hold. Then the following statements are true:*

1. *Under* $H_0$, $\lim_{n\to\infty} P(\widehat{T} > \hat{c}^{\eta}(\alpha)) \leq \alpha$;
2. *Under* $H_1$, $\lim_{n\to\infty} P(\widehat{T} > \hat{c}^{\eta}(\alpha)) = 1$.

Theorem 3.2 shows that our test can control the asymptotic size under the significance level under the null hypothesis and it is consistent against any fixed alternative hypothesis. It is straightforward for us to show that our test would have non-trivial power against some classes of local alternatives similar to those in Hsu et al. (2019), but we omit the details.

The high-level conditions in Assumption 3.1 are attainable by various estimators, in particular, kernel, series, deep neural networks, and Lasso. The theory of the conventional nonparametric kernel and series methods is well established. The rate conditions are based on the standard root-mean-squared norm (or the $L_2$ norm), rather than the partial $L_2$ norm with a fixed value of $t$ as for estimating $\mu(t)$ in Colangolo and Lee (2022). The main reason is that $\nu(\ell)$ in our test is an integration of $\mu(t)$ over a range of $t$ and can be estimated at a regular root-$n$ rate. This advantageous feature enables a broader class of machine learning or nonparametric methods whose $L_2$-norm convergence rates are available in the literature; for example, Lasso (Bickel et al., 2009), neural networks (Chen and White, 1999; Schmidt-Hieber, 2020; Farrell et al., 2021), random forests (Syrgkanis and Zampetakis, 2020), and empirical $L_2$ rate for boosting in Luo and Spindler (2016), as discussed in Chernozhukov et al. (2018) and Chernozhukov et al. (2022).

Recently Farrell et al. (2021) provide $\|\hat{\gamma}_k - \gamma\|_2$ of deep neural networks. Colangelo and Lee (2022) propose conditional density (GPS) estimators that utilize generic estimators of the conditional mean function. Specifically, Lemmas 1 and 2 in Colangelo and Lee (2022) provide the convergence rates for their GPS estimators $\|\hat{p}_k - p\|_2$ using the deep neural networks in Farrell et al. (2021). So Assumptions 3.1(i) and (ii) are attainable by the deep neural networks in Farrell et al. (2021) and Colangelo and Lee (2022), summarized in Section 3.2. Alternative estimators for estimating the GPS can be the kernel density estimator, the artificial neural networks in Chen and White (1999), and Belloni et al. (2019), or the series cross-validated method in Zhang (2022). In Section 3.3, we illustrate how to employ Lasso methods to estimate the nuisance conditional mean function $\gamma(t,x)$ and the generalized propensity score $p(t,x)$ in a high-dimensional setting. We provide sufficient conditions to verify the high-level Assumption 3.1.

## 3.2 Conditional density estimation

Colangelo and Lee (2022) propose estimating the reciprocal of the conditional density function $1/p(t,x)$ using a generic mean regression estimator $\hat{\Upsilon}(W;x)$ of the conditional mean $E[W|X=x]$ for a random variable $W$. Given the mean-squared convergence rate of $\hat{\Upsilon}$, we can obtain the corresponding mean-squared convergence rate for $\hat{p}$ to verify Assumption 3.1. We summarize the estimator ReGPS below and refer readers to Colangelo and Lee (2022) for details. Moreover the ReGPS estimator avoids plugging in a small estimate in the denominator, and the estimate is positive by construction.

It is known that for any CDF $F$, $\frac{d}{du}F^{-1}(u) = \frac{1}{F'(F^{-1}(u))}$ for $u \in (0,1)$. So $\frac{1}{f_{T|X}(t|x)} = \frac{\partial}{\partial u}F^{-1}_{T|X}(u|x)\big|_{u=F_{T|X}(t|x)}$. Colangelo and Lee (2022) propose the ReGPS estimator of $\frac{1}{f_{T|X}(t|x)}$ using a numerical differentiation

$$\widehat{\frac{1}{p(t|x)}} = \frac{\hat{F}^{-1}_{T|X}\big(\hat{F}_{T|X}(t|x)+\epsilon\big|x\big) - \hat{F}^{-1}_{T|X}\big(\hat{F}_{T|X}(t|x)-\epsilon\big|x\big)}{2\epsilon},$$

where $\epsilon = \epsilon_n$ is a positive sequence vanishing as $n$ grows and $\hat{F}_{T|X}(t|x) \pm \epsilon \in (0,1)$. The conditional CDF is estimated by $\hat{F}_{T|X}(t|x) = \hat{\Upsilon}\left(\Phi\left(\frac{t-T}{h_1}\right);x\right)$, where $\Phi$ is the CDF of a standard normal random variable and $h_1 = h_{1n}$ is a bandwidth sequence vanishing as $n$ grows. The conditional $u$-quantile function $F^{-1}_{T|X}(u|x)$ is estimated by the generalized inverse function $\hat{F}^{-1}_{T|X}(u|x) = \inf_{t\in\mathcal{T}}\{t : \hat{F}_{T|X}(t|x) \geq u\}$.

Suppose that $\sup_{t\in\mathcal{T}}\left\|\hat{\Upsilon}\left(\Phi\left(\frac{t-T}{h_1}\right);X\right) - \mathbb{E}\left[\Phi\left(\frac{t-T}{h_1}\right)\big|X\right]\right\|_2 = O_p(R_{1n})$ for a sequence of constants $R_{1n}$. Assume $\hat{\Upsilon}\left(\Phi\left(\frac{t-T}{h_1}\right);x\right)$ to be continuous in $t \in \mathcal{T}$. We can obtain $\|\hat{p} - p\|_2 = O_p(R_{1n}\epsilon^{-1} + h_1^2\epsilon^{-1} + \epsilon^2)$, by showing the convergence rate in Lemma 1 in Colangelo and Lee (2022)

to hold uniformly over $t$.

The ReGPS estimator can employ the deep neural networks in Farrell et al. (2021) that use the fully connected feedforward neural networks (multilayer perceptron) and the nonsmooth rectified linear units (ReLU) activation function when the dimension of the control variables $d_x$ is fixed. Lemma 3 in Colangelo and Lee (2022), which is based on Theorem 1 in Farrell et al. (2021), formally provides that $R_{1n}^2 = n^{-\frac{r}{r+d_X}} \log^8 n + \log\log n/n$, where the smoothness parameter $r \in \mathcal{N}_+$ such that $\max_{\alpha,|\alpha|\leq r} \operatorname{ess\,sup}_{y\in\mathcal{Y},t\in\mathcal{T},x\in\mathcal{X},} |D_x^\alpha f_{T|X}(t|x)| \leq c$ for some finite positive constant $c$. We do not repeat other low-level regularity conditions in Colangelo and Lee (2022) here to conserve space.

## 3.3 Step 1 Lasso

We illustrate our test by applying Lasso methods in Step 1 to estimate the nuisance functions, when $X$ is potentially high-dimensional. We follow Su, Ura, and Zhang (2019) (SUZ, hereafter) to approximate the outcome and treatment models by a linear regression and a logistic regression, respectively. In particular, the approximation errors satisfy Assumption 3.4 that imposes sparsity structures on $\gamma(t,x)$ and the conditional CDF $F_{T|X}$ so that the number of effective covariates that can affect them is small. See Farrell (2015), Chernozhukov et al. (2022), for example, for in-depth discussions on the specification of high-dimensional sparse models. We modify the penalized local least squares estimator of $\gamma(t,X)$ in SUZ and use the conditional density estimator in SUZ. For completeness, we present the estimators and asymptotic theory in SUZ and refer readers to SUZ for details.

To estimate the conditional density $p(t,x) = f_{T|X}(t|x)$, first estimate the conditional CDF $F_{T|X}$ by the logistic distributional Lasso regression in Belloni et al. (2017) and then take the numerical derivative. Let $b(X)$ be a $p\times 1$ vector of basis functions. We approximate $F_{T|X}(t|x)$ by $\Lambda(b(x)'\beta_t)$, where $\Lambda$ is the logistic CDF. For $k\in\{1,...,K\}$, $\hat{F}_{T|X_k}(t|x) = \Lambda(b(X)'\hat{\beta}_{tk})$, where

$$\hat{\beta}_{tk} = \arg\min_{\beta} \frac{1}{n-n_k} \sum_{i\notin I_k} M(\mathbf{1}\{T_i\leq t\}, X_i; \beta) + \frac{\tilde{\lambda}}{n-n_k} \|\hat{\Psi}_{tk}\beta\|_1 \tag{3.3}$$

where $M(y,x;g) = -\big(y\log(\Lambda(b(x)'g)) + (1-y)\log(1-\Lambda(b(x)'g))\big)$ is the logistic likelihood, the penalty $\tilde{\lambda} = 1.1\Phi^{-1}(1-r/\{p\vee nh_1\})n^{1/2}$, for some $r\to 0$ and $h_1\to 0$, with the standard normal CDF $\Phi$. A generic penalty loading matrix $\hat{\Psi}_{tk}$ is computed by Algorithm 1 below from the iterative Algorithm 3.2 in SUZ.

**Algorithm 1 (SUZ Algorithm 3.2)** *For $k\in\{1,...,K\}$,*

1. *Let $\hat{\Psi}_{tk}^0 = diag(l_{tk,1}^0,...,l_{tk,p}^0)$, where $l_{tk,j}^0 = \|\mathbf{1}\{T\leq t\}b_j(X)\|_{P_{nk},2}$. Compute $\hat{\beta}_{tk}^0$ by (3.3) with*

$\hat{\Psi}^0_{tk}$ *in place of* $\hat{\Psi}_{tk}$. *Let* $\hat{F}^0_{T|X_k}(t|x) = \Lambda(b(x)'\hat{\beta}^0_{tk})$.

2. *Compute* $\hat{\Psi}^s_{tk} = diag(l^s_{tk,1}, ..., l^s_{tk,p})$, *where* $l^s_{tk,j} = \left\| \left( \mathbf{1}\{T \leq t\} - \hat{F}^{s-1}_{T|X_k}(t|X) \right) b_j(X) \right\|_{P_{nk},2}$, *for* $s = 1, ..., S$. *Compute* $\hat{\beta}^s_{tk}$ *by (3.3) with* $\hat{\Psi}^s_{tk}$ *in place of* $\hat{\Psi}_{tk}$. *Let* $\hat{F}^s_{T|X_k}(t, x) = \Lambda(b(x)'\hat{\beta}^s_{tk})$.

Let the final penalty loading matrix $\hat{\Psi}_{tk}$ be $\hat{\Psi}^S_{tk}$ from Algorithm 1. Compute $\hat{F}_{T|X_k}(t|x) = \Lambda(b(X)'\hat{\beta}_{tk})$ from (3.3). Then the conditional density estimator

$$\hat{p}_k(t, x) = \frac{\hat{F}_{T|X_k}(t + h_1|x) - \hat{F}_{T|X_k}(t - h_1|x)}{2h_1}.$$

Assumption 3.4 collects the conditions in Theorems 3.1 and 3.2 in SUZ.

**Assumption 3.4 (Lasso)** *Let* $\mathcal{T}_0$ *be a compact subset of the support of* $T$ *and* $\mathcal{X}$ *be the support of* $X$.

*(i)* *(a)* $\| \max_{j \leq p} |b_j(T, X)| \|_{P,\infty} \leq \zeta_n$ *and* $\underline{C} \leq E\left[b_j(T, X)^2\right] \leq 1/\underline{C}$, *for some positive constant* $\underline{C}$, $j = 1, ..., p$.

*(b)* $\sup_{t \in \mathcal{T}_0} \max(\|\beta_t\|_0, \|\theta\|_0) \leq s$ *for some* $s$ *which possibly depends on* $n$, *where* $\|\theta\|_0$ *denotes the number of nonzero coordinates of* $\theta$.

*(c) For the approximation error,* $\sup_{t \in \mathcal{T}_0} \|F_{T|X}(t|X) - \Lambda(b(X)'\beta_t)\|_{P,\infty} = O((s^2\zeta_n^2 \log(p \vee n)/n)^{1/2})$ *and* $\|\gamma(T, X) - b(T, X)'\theta\|_{P,\infty} = O\left((s^2\zeta_n^2 \log(p \vee n)/n)^{1/2}\right)$.

*(d)* $p(t, x)$ *is second-order differentiable w.r.t.* $t$ *with bounded derivatives uniformly over* $(t, x) \in \mathcal{T}_0 \times \mathcal{X}$.

*(e)* $\zeta_n^2 s^2 \iota_n^2 \log(p \vee n)/(nh_1) \to 0$, $nh_1^5/(\log(p \vee n)) \to 0$.

*(ii)* *(a) There exists some positive constant* $\underline{C} < 1$ *such that* $\underline{C} \leq p(t, x) \leq 1/\underline{C}$ *uniformly over* $(t, x) \in \mathcal{T}_0 \times \mathcal{X}$.

*(b)* $\gamma(t, x)$ *is three times differentiable with all three derivatives being bounded uniformly over* $(t, x) \in \mathcal{T}_0 \times \mathcal{X}$.

*(iii) There exists a sequence* $\iota_n \to \infty$ *such that w.p.a.1* $0 < \kappa' \leq \inf_{\delta \neq 0, \|\delta\|_0 \leq s\iota_n} \frac{\|b(T,X)'\delta\|_{P_n,2}}{\|\delta\|_2} \leq \sup_{\delta \neq 0, \|\delta\|_0 \leq s\iota_n} \frac{\|b(T,X)'\delta\|_{P_n,2}}{\|\delta\|_2} \leq \kappa'' < \infty$.

Let Assumption 3.4 hold. Then Theorems 3.1 and 3.2 in SUZ imply that $\sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |\hat{\gamma}_k(t,x) - \gamma(t,x)| = O_P(A_n)$, where $A_n = \iota_n(\log(p \vee n)s^2\zeta_n^2/n)^{1/2}$ and $\sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |\hat{p}_k(t,x) - p(t,x)| = O_P(R_n)$, where $R_n = h_1^{-1}(\log(p \vee n)s^2\zeta_n^2/n)^{1/2}$. Then we can obtain the same rates for the root-mean-squared rates $\|\hat{\gamma}_k - \gamma\|_2$ and $\|\hat{p}_k - p\|_2$ to verify Assumption 3.1. Therefore a sufficient condition of Assumption 3.4(i) is $A_n \to 0$ and $R_n \to 0$. And a sufficient condition of Assumption 3.4(ii) is $\sqrt{n}A_nR_n \to 0$.

It may be interesting to note that the conditional density estimators discussed above are based on the mean regression estimators for the conditional CDF. The numerical differentiation slows down the convergence rate by the step size, i.e., $h_1^{-1}$ in the Lasso method and $\epsilon^{-1}$ in the ReGPS in Colangelo and Lee (2022).[6]

As we use a Neyman orthogonal and doubly robust moment function for $\nu(\ell)$, the final estimator is less sensitive to the estimation errors for the Step 1 nuisance functions and hence the corresponding tuning parameters. The Lasso penalty could be chosen by the rule-of-thumb method in Su et al. (2019). Specifically following Su et al. (2019), we can let $\iota_n = \sqrt{\log\log(n)}$ in $\lambda$, $r = 1/\log(n)$ in $\tilde{\lambda}$, and the rule-of-thumb bandwidth $h_1 = 1.06 \times sd(T) \times n^{-1/5}$. See also Farrell (2015) for the choice of penalty. An alternative widespread practice is cross-validation. Chetverikov et al. (2021) provide theoretical justification of the cross-validated Lasso estimator by showing that it has nearly optimal rates of convergence.

# 4 Simulation

This section provides a simulation study to examine the finite sample performance of the proposed test. We will examine the size and power properties of our test, and compare our test with SUZ's method. We also examine the local power of our test and SUZ's method. Finally, we examine the robustness of our test against some tuning parameters in the test.

To implement our test in practice, one has to choose several tuning parameters in advance. We make the following propositions concerning the choice of these parameters and present related Monte Carlo simulation results further below.[7]

1. Instrumental functions: We opt for using a set of indicator functions of countable hypercubes.

[6] By Theorem 6.2 and Comment 6.1 in Belloni et al. (2017), the rate of the conditional CDF regression estimator is $\sup_{t\in\mathcal{T}} \|\hat{F}_{T|X}(t,X) - F_{T|X}(t,X)\|_{P_n,2} = O_p(R_{1n})$ with $R_{1n} = (\log(p\vee n)s/n)^{1/2}$. Theorem 3.2 in Su et al. (2019) shows that the rate of the conditional density estimator $\sup_{t\in\mathcal{T}} \|\hat{p}(t,X) - p(t,X)\|_{P_n,2} = O_p(R_{1n}h_1^{-1})$.

[7] The MATLAB code is available upon request.

Define

$$\mathcal{L} = \Big\{ \ell = (t_1, t_2, q^{-1}) : (t_1, t_2) \in [0,1]^2, t_1 > t_2, q = 2, \cdots, q_1, \\ q \cdot (t_1, t_2) \in \{0, 1, 2, \cdots, q-1\}^2 \Big\},$$

where $q_1$ is a natural number and is chosen such that the expected sample size of the smallest cube is around 50. Our simulations show that the results are robust to various expected sample sizes.

2. $Q(\ell)$**:** The distribution $Q(\ell)$ assigns weight $\propto q^{-2}$ to each $q$, and for each $q$, $Q(\ell)$ assigns an equal weight to each instrumental function with the last element of $\ell$ equal to $q^{-1}$. Recall that for each $q$, there are $(q(q+1)/2)$ instrumental functions with the last element of $\ell$ equal to $q^{-1}$.

3. $a_n$**,** $B_n$**,** $\epsilon$**,** $\eta$**:** We set $a_n = 0.15 \cdot \ln(n)$, $B_n = 0.85 \cdot \ln(n)/\ln\ln(n)$, $\epsilon = 10^{-6}$, and $\eta = 10^{-6}$ as suggested by Hsu et al. (2019). These choices are used in all the simulations that we report below and seem to perform well.

Note that if $\mu(t)$ is differentiable for all $t$, then $H_0$ in (2.1) is equivalent to

$$H_0'' : \ d\mu(t)/dt \geq 0, \text{ for } t \in [t_\ell, t_u]. \tag{4.1}$$

Therefore, one can employ SUZ's method to test weather the average partial effect $d\mu(t)/dt$ is greater or equal to zero for all $t \in [t_\ell, t_u]$; see Appendix for details of the procedure.

For all data generating processes (DGPs), the continuous treatment variable $T$, the control variables $X$, and the error term $U_y$ are generated as follows

$$\begin{aligned} T &= (3.6 + X'\beta)/7.2 + 0.5U_t, \\ X &= (X_1, \ldots, X_{100})' \sim \mathcal{N}(0, \Sigma), \\ U_y &\sim \mathcal{N}(0,1), \end{aligned}$$

where the $(i,j)$-entry $\Sigma_{ij} = (0.5)^{|i-j|}$ for $i, j = 1, \ldots, 100$, $U_t \sim \mathcal{N}(0,1)$, and $U_y$, $U_t$, and $X$ are mutually independent. We set $\beta_j = 1/j^2$ for mild dependence between $X_j$ and $\beta_j = 1/j$ for strong dependence between $X_j$. Four cases of the potential outcomes are studied:

DGP-1: $Y = U_y$,

DGP-2: $Y = X'\beta T + T^2 + X'\beta + U_y$,

DGP-3: $Y = X'\beta T - T + X'\beta + U_y$,

DGP-4: $Y = X'\beta T + \sin(\pi T) + X'\beta + U_y$.

In DGP-1, $\mu(t) = 0$, and $H_0$ holds with moment equalities. In this case, we expect that the size of the proposed test will achieve the nominal level, since every moment would hold with equality. In DGP-2, $\mu(t) = t^2$, and $H_0$ holds with strict moment inequalities. In this case, we expect the size will converge to zero since every moment would hold with strict inequality. This is because the test statistics will converge to zero and the critical value is bounded away from zero. In DGP-3, $\mu(t) = -t$, and in DGP-4, $\mu(t) = \sin(\pi t)$. In both cases, $H_0$ does not hold, and we expect the power will increase with the sample size.

In these DGPs, we have $1 + d_X = 101$ regressors. We consider samples of sizes $n = 200$, 400, 800, and 1200. For $q_1$, we set $q_1 = 4$ for $n = 200$, $q_1 = 8$ for $n = 400$, $q_1 = 16$ for $n = 800$, and $q_1 = 24$ for $n = 1200$. Recall that $N$ denotes the expected sample size of the smallest cube corresponding to $q_1$. We choose $q_1$ such that $N = 50$ for all the sample sizes. For the $K$-fold cross-fitting, we consider the number of split subsamples $K \in \{2, 5, 10\}$. All our Monte Carlo results are based on 1000 simulations. In each simulation, the critical value is approximated by 1000 bootstrap replications. The nominal size of the test is set at 10%.

To estimate the conditional mean function $\gamma(t, x) = E[Y|T = t, X = x]$, we employ the Lasso regression, where the penalization parameter is chosen via grid search utilizing 10-fold cross validation. To estimate the conditional density estimation $p(t, x)$, we first estimate $F_{T|X}(t|x)$ by the logistic distributional Lasso regression, and then take the numerical derivative. The penalization parameter of the distributional Lasso regression is estimated by Algorithm 3.2 of Su et al. (2019), described in Section 3.3. Also, all Lasso estimations include an intercept and the covariates. For numerical integration in Step 2, we set $M = [n^{2/3}]$, where $[\cdot]$ is the nearest integer. Our test is based on the trimmed generalized propensity score estimator, defined as $\tilde{p}(T_i, X_i) = \max\{\hat{p}(T_i, X_i), 0.025\}$, implying that conditional treatment densities below 2.5% are set to 2.5%.[8,9]

Table 1 shows the rejection probabilities of our test for DGPs 1-4, and the results are consistent with our theoretical findings. For the mild dependence case ($\beta_j = 1/j^2$), the proposed test controls size well in DGP-1 and DGP-2, and the rejection probabilities increase with the sample size and are greater than the nominal size 0.1 in DGP-3 and DGP-4. For the strong dependence case ($\beta_j = 1/j$), our test still controls size well in both DGP-1 and DGP-2. The power increases

[8] In general, one can follow Donald et al. (2014) and Hsu et al. (2020) and trim the estimated generalized propensity scores to prevent them from being too close zero, in order to obtain a more stable estimator whose variance is not affected by extremely low scores.

[9] Based on this trimming rule, around 0.5% of the samples are trimmed.

Table 1: Rejection probabilities for $N = 50$ and different $K$

| | | $\beta_j = 1/j^2$ | | | | $\beta_j = 1/j$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| DGP | n | K=2 | K=5 | K=10 | SUZ | K=2 | K=5 | K=10 | SUZ |
| 1 | 200 | 0.105 | 0.108 | 0.120 | 0.124 | 0.133 | 0.098 | 0.129 | 0.088 |
| 1 | 400 | 0.112 | 0.118 | 0.126 | 0.100 | 0.107 | 0.110 | 0.109 | 0.112 |
| 1 | 800 | 0.097 | 0.103 | 0.118 | 0.114 | 0.092 | 0.128 | 0.125 | 0.106 |
| 1 | 1200 | 0.113 | 0.121 | 0.104 | 0.102 | 0.122 | 0.121 | 0.120 | 0.099 |
| 2 | 200 | 0.001 | 0.000 | 0.001 | 0.000 | 0.002 | 0.000 | 0.003 | 0.000 |
| 2 | 400 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2 | 800 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2 | 1200 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 3 | 200 | 0.931 | 0.976 | 0.978 | 0.666 | 0.605 | 0.768 | 0.833 | 0.131 |
| 3 | 400 | 1.000 | 1.000 | 1.000 | 0.904 | 0.976 | 0.995 | 0.996 | 0.310 |
| 3 | 800 | 1.000 | 1.000 | 1.000 | 0.973 | 1.000 | 1.000 | 1.000 | 0.617 |
| 3 | 1200 | 1.000 | 1.000 | 1.000 | 0.979 | 1.000 | 1.000 | 1.000 | 0.760 |
| 4 | 200 | 0.197 | 0.210 | 0.205 | 0.925 | 0.061 | 0.083 | 0.080 | 0.453 |
| 4 | 400 | 0.429 | 0.501 | 0.530 | 0.989 | 0.155 | 0.217 | 0.227 | 0.803 |
| 4 | 800 | 0.859 | 0.905 | 0.908 | 0.998 | 0.427 | 0.565 | 0.562 | 0.950 |
| 4 | 1200 | 0.985 | 0.987 | 0.984 | 1.000 | 0.702 | 0.821 | 0.803 | 0.982 |

with the sample size in DGP-3 and DGP-4, but the rejection probabilities are a bit less than the nominal size 0.1 for $n = 200$ in DGP-4. Overall, we do not find significant difference for different choices of $K$.

We next compare the performance between our test and the SUZ method. As we can see from Table 1, the performance of the SUZ method is quite similar to that of our test in DGP-1 and DGP-2. However, our test has better power properties than the SUZ method in DGP-3, but lower power than the SUZ method in DGP-4. The main reason is the following. In DGP-3, $d\mu(t)/dt = -1$ and in this case, the violation of the null hypothesis is the same for each $t$. In contrast, in DGP-4, $d\mu(t)/dt = \pi\cos(\pi t)$ with violation of the null hypothesis is higher when $t$ is closer to 1. For SUZ's method, which is based on a uniform confidence band of direct estimation of $d\mu(t)/dt$, it is equivalent to a supremum-type test, so the power of it can be driven by a large deviation of the null at a specific point such as DGP-4. On the other hand, our test is an integral-type test, so in general, it will have better power if the deviation of the null is more evenly spaced as in DGP-3. As a result, our test has better power in DGP-3 and SUZ's method has better power in DGP-4.

We now consider two extended cases to investigate the local power properties. The data generating processes are based on DGP-3 and DGP-4, and the potential outcomes are modified to

DGP-$3^n$: $Y = X'\beta T - T/\sqrt{n} + X'\beta + U_y$,

DGP-$4^n$: $Y = X'\beta T + \sin(\pi T)/\sqrt{n} + X'\beta + U_y$.

Therefore $\mu(t) = -t/\sqrt{n}$ in DGP-$3^n$ and $\mu(t) = \sin(\pi t)/\sqrt{n}$ in DGP-$4^n$.

Table 2: Rejection probabilities for $N = 50$ and different $K$ in DGP-$3^n$ and DGP-$4^n$

| | | $\beta_j = 1/j^2$ | | | | $\beta_j = 1/j$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| DGP | n | K=2 | K=5 | K=10 | SUZ | K=2 | K=5 | K=10 | SUZ |
| $3^n$ | 200 | 0.091 | 0.104 | 0.086 | 0.038 | 0.050 | 0.060 | 0.066 | 0.008 |
| $3^n$ | 400 | 0.108 | 0.092 | 0.131 | 0.059 | 0.052 | 0.063 | 0.076 | 0.009 |
| $3^n$ | 800 | 0.111 | 0.103 | 0.133 | 0.069 | 0.062 | 0.088 | 0.105 | 0.008 |
| $3^n$ | 1200 | 0.127 | 0.108 | 0.104 | 0.054 | 0.074 | 0.081 | 0.105 | 0.011 |
| $4^n$ | 200 | 0.060 | 0.057 | 0.092 | 0.028 | 0.040 | 0.033 | 0.047 | 0.006 |
| $4^n$ | 400 | 0.073 | 0.082 | 0.073 | 0.054 | 0.037 | 0.030 | 0.035 | 0.009 |
| $4^n$ | 800 | 0.041 | 0.081 | 0.082 | 0.064 | 0.041 | 0.051 | 0.069 | 0.012 |
| $4^n$ | 1200 | 0.063 | 0.076 | 0.077 | 0.065 | 0.042 | 0.065 | 0.052 | 0.015 |

Tables 2 shows the rejection probabilities of our test and the SUZ method in DGP-$3^n$ and DGP-$4^n$. It is clear that our test has a nontrivial power in both DGPs and has better local power properties than the SUZ method. The main reason is the following. The SUZ method is based on the uniform confidence band of nonparametric estimation of $d\mu(t)/dt$ that converges at a nonparametric rate, so it will not have local power against the $n^{-1/2}$ local alternatives in the form of DGP-$3^n$ and DGP-$4^n$ in that the local power is equal to or smaller than the pre-specified significance level; on the other hand, our test can have non-trivial local power against some $n^{-1/2}$ local alternatives as in Hsu et al. (2019).

Now we investigate the robustness of the performance of our test to the choice of $q_1$. We consider three alternative choices of $q_1$, each resulting in the expected sample size of the smallest cube $N = 33$, 40, and 66, respectively. Table 3 shows the rejection probabilities of our test for different choices of $N$. The results suggest that the choice of $q_1$ does not affect the test performance much. Therefore, the finite sample behavior of our test appears to be reasonably robust to different values of $q_1$.

Next we investigate the robustness of the performance of our test to the choice of $a_n$ and $B_n$. Instead of setting $a_n = 0.15 \cdot \ln(n)$ and $B_n = 0.85 \cdot \ln(n)/\ln\ln(n)$, we consider $a_n = \sqrt{0.3 \cdot \ln(n)}$ and $B_n = \sqrt{0.4 \cdot \ln(n)/\ln\ln(n)}$ as suggested by Andrews and Shi (2014) and Hsu and Shen (2020). Table 4 shows that our test still controls size well in both DGP-1 and DGP-2, the rejection probabilities are greater than the nominal size 0.1 in DGP-3 and DGP-4, and our test has nontrivial power in DGP-$3^n$ and DGP-$4^n$.

Table 3: Rejection probabilities for $K = 5$ and different $N$

| | | $\beta_j = 1/j^2$ | | | | $\beta_j = 1/j$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| DGP | n | N=33 | N=40 | N=50 | N=66 | N=33 | N=40 | N=50 | N=66 |
| 1 | 200 | 0.133 | 0.100 | 0.108 | 0.113 | 0.101 | 0.121 | 0.098 | 0.111 |
| 1 | 400 | 0.118 | 0.116 | 0.118 | 0.099 | 0.127 | 0.114 | 0.110 | 0.110 |
| 1 | 800 | 0.136 | 0.103 | 0.103 | 0.091 | 0.122 | 0.108 | 0.128 | 0.109 |
| 1 | 1200 | 0.102 | 0.104 | 0.121 | 0.108 | 0.111 | 0.097 | 0.121 | 0.104 |
| 2 | 200 | 0.003 | 0.000 | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 | 0.001 |
| 2 | 400 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 |
| 2 | 800 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2 | 1200 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 3 | 200 | 0.982 | 0.982 | 0.976 | 0.984 | 0.812 | 0.799 | 0.768 | 0.789 |
| 3 | 400 | 1.000 | 1.000 | 1.000 | 1.000 | 0.994 | 0.994 | 0.995 | 0.993 |
| 3 | 800 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 3 | 1200 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 4 | 200 | 0.210 | 0.247 | 0.210 | 0.170 | 0.100 | 0.072 | 0.083 | 0.084 |
| 4 | 400 | 0.521 | 0.503 | 0.501 | 0.465 | 0.220 | 0.221 | 0.217 | 0.202 |
| 4 | 800 | 0.911 | 0.912 | 0.905 | 0.895 | 0.549 | 0.554 | 0.565 | 0.547 |
| 4 | 1200 | 0.992 | 0.989 | 0.987 | 0.993 | 0.816 | 0.806 | 0.821 | 0.809 |
| $3^n$ | 200 | 0.102 | 0.120 | 0.104 | 0.093 | 0.047 | 0.066 | 0.060 | 0.051 |
| $3^n$ | 400 | 0.114 | 0.108 | 0.092 | 0.109 | 0.071 | 0.075 | 0.063 | 0.069 |
| $3^n$ | 800 | 0.115 | 0.118 | 0.103 | 0.127 | 0.092 | 0.073 | 0.088 | 0.083 |
| $3^n$ | 1200 | 0.114 | 0.123 | 0.108 | 0.118 | 0.092 | 0.097 | 0.081 | 0.082 |
| $4^n$ | 200 | 0.071 | 0.082 | 0.057 | 0.070 | 0.048 | 0.038 | 0.033 | 0.042 |
| $4^n$ | 400 | 0.086 | 0.065 | 0.082 | 0.068 | 0.056 | 0.039 | 0.030 | 0.032 |
| $4^n$ | 800 | 0.077 | 0.072 | 0.081 | 0.067 | 0.060 | 0.060 | 0.051 | 0.065 |
| $4^n$ | 1200 | 0.072 | 0.066 | 0.076 | 0.089 | 0.059 | 0.056 | 0.065 | 0.060 |

Table 4: Rejection probabilities for $a_n = \sqrt{0.3 \cdot \ln(n)}$ and $B_n = \sqrt{0.4 \cdot \ln(n)/\ln\ln(n)}$

| | | $\beta_j = 1/j^2$ | | | $\beta_j = 1/j$ | | |
|---|---|---|---|---|---|---|---|
| DGP | n | K=2 | K=5 | K=10 | K=2 | K=5 | K=10 |
| 1 | 200 | 0.108 | 0.122 | 0.111 | 0.125 | 0.108 | 0.116 |
| 1 | 400 | 0.103 | 0.105 | 0.115 | 0.110 | 0.112 | 0.104 |
| 1 | 800 | 0.132 | 0.104 | 0.098 | 0.114 | 0.110 | 0.120 |
| 1 | 1200 | 0.100 | 0.095 | 0.090 | 0.101 | 0.108 | 0.091 |
| 2 | 200 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2 | 400 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2 | 800 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2 | 1200 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 3 | 200 | 0.928 | 0.978 | 0.984 | 0.594 | 0.794 | 0.826 |
| 3 | 400 | 1.000 | 1.000 | 1.000 | 0.979 | 0.994 | 0.996 |
| 3 | 800 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 3 | 1200 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 4 | 200 | 0.118 | 0.135 | 0.140 | 0.049 | 0.057 | 0.063 |
| 4 | 400 | 0.324 | 0.400 | 0.419 | 0.070 | 0.157 | 0.165 |
| 4 | 800 | 0.811 | 0.858 | 0.839 | 0.358 | 0.416 | 0.440 |
| 4 | 1200 | 0.969 | 0.984 | 0.985 | 0.611 | 0.695 | 0.716 |
| $3^n$ | 200 | 0.075 | 0.101 | 0.085 | 0.037 | 0.055 | 0.057 |
| $3^n$ | 400 | 0.090 | 0.096 | 0.102 | 0.043 | 0.052 | 0.055 |
| $3^n$ | 800 | 0.091 | 0.112 | 0.136 | 0.056 | 0.076 | 0.074 |
| $3^n$ | 1200 | 0.098 | 0.116 | 0.134 | 0.066 | 0.095 | 0.094 |
| $4^n$ | 200 | 0.071 | 0.065 | 0.063 | 0.023 | 0.031 | 0.028 |
| $4^n$ | 400 | 0.066 | 0.074 | 0.074 | 0.024 | 0.026 | 0.037 |
| $4^n$ | 800 | 0.036 | 0.050 | 0.075 | 0.036 | 0.042 | 0.055 |
| $4^n$ | 1200 | 0.070 | 0.090 | 0.084 | 0.033 | 0.046 | 0.054 |

# 5 Empirical application

As an empirical illustration, we apply our test to data from the Job Corps study. The latter was conducted between November 1994 and February 1996 to evaluate the publicly funded U.S. Job Corps program and used an experimental design that randomly assigned access to the program. Job Corps targets youths from low-income households who are between 16 and 24 years old and legally reside in the U.S. Program participants obtained on average roughly 1200 hours of vocational and/or academic classroom training as well as housing and board over an average duration of 8 months. We refer to Schochet et al. (2001) and Schochet et al. (2008) for a detailed discussion of the study design and the average effects of program assignment on a range of different outcomes. Their results suggest that Job Corps raises educational attainment, reduces criminal activity, and increases labor market performance measured by employment and earnings, at least for some years after the program.

Particularly relevant for our context is the study by Flores et al. (2012), who consider the length of exposure to academic and/or vocational training as continuously distributed treatment to assess its effect on earnings based on regression and weighting estimators (using the inverse of the conditional treatment density as weight). As the length of treatment exposure is (in contrast to Job Corps assignment) not random, they impose a selection-on-observables assumption and control for baseline characteristics at Job Corps assignment. While the authors find overall positive average effects of increasing hours in academic and vocational instruction, the marginal effects appear to decrease with length of exposure, pointing to a potential concavity in the association of earnings and time of instruction. Similarly, Lee (2018) and Colangelo and Lee (2022) assess the effect of hours of training on the proportion of weeks employed in the second year after program assignment based on kernel regression and double machine learning, respectively. Also for this outcome, the plotted regression lines in both studies point to a concave association with the treatment dose.[10]

However, in the light of estimation uncertainty, mere eye-balling of the outcome-treatment associations in empirical applications does not tell us whether specific shape restrictions can be refuted. For this reason, we use our DML method with Lasso regression for nuisance parameter estimation to formally test whether weak positive and negative monotonicity can be rejected in the Job Corps data when considering several labor market outcomes. To this end, we define the treatment variable $T$ as the total hours spent in academic and vocational training in the 12 months following the program assignment. Our outcomes $Y$ include weekly earnings in the fourth year,

[10]See also Huber et al. (2020), who use a causal mediation approach to assess the direct effect of the treatment dose on the number of arrests in the fourth year after program assignment when controlling for employment behavior in the second year based on inverse probability weighting, and find a non-linear association.

earnings and hours worked per week in quarter 16, and a binary employment indicator four years after assignment (i.e., in week 208).

For invoking weak unconfoundedness (Assumption 2.1), we consider the same set of pre-treatment covariates $X$ as Lee (2018), Colangelo and Lee (2022), and Huber et al. (2020), which overlaps with the control variables of Flores et al. (2012).[11] We condition on individual characteristics like age, gender, ethnicity, language competency, education, marital status, household size and income, previous receipt of social aid, family background (e.g., parents' education), and criminal activity, as well as health and health-related behavior (e.g., smoking, alcohol, or drug consumption). Conditioning on such a rich set of socio-economic variables appears important, as the satisfaction of weak unconfoundedness relies on successfully controlling for all factors jointly affecting treatment duration and labor market behavior. Furthermore, we include variables that might be associated with the duration of training, namely expectations about Job Corps and interaction with the recruiters, which might serve as proxies for unobserved personality traits (like motivation) that could also affect the outcomes. Finally, we control for pre-treatment outcomes, namely previous labor market participation and earnings, to tackle any confounders that affect the outcomes of interest through their respective pre-treatment values.

The original Job Corps data set consists of $15,386$ individuals prior to program assignment, but a substantial share never enrolled in the program and dropped out of the study, leaving only $11,313$ individuals with completed follow-up interviews four years after randomization. Among those, $6,828$ had been randomized into Job Corps and had thus access to academic or vocational training. To define our final evaluation sample, we follow Flores et al. (2012), Lee (2018), Colangelo and Lee (2022), and Huber et al. (2020) and consider observations with at least 40 hours (or one working week) of training for our analysis, all in all $4,166$ individuals. Among these, there are cases of item non-response in various elements of $X$ measured at the baseline survey, which we account by including missing dummies as additional regressors, while observations with missing values in the outcome of interest need to be dropped when running the respective test. Table 9 in the Appendix provides descriptive statistics for selected covariates $X$ (see Huber et al. (2020) for a full list of control variables) as well as for the treatment $T$ and all outcomes $Y$, including the respective number of nonmissing observations (nonmissing).

The choices of nuisance parameters are the same as in the simulations (see the previous section). The number of subsamples used for cross-fitting is 5, and the expected sample size of the smallest cube is either 40 or 50. The Lasso estimations include an intercept, the covariates and

[11] A control variable in Flores et al. (2012) we do not have access to is the local unemployment rate, which was constructed by matching county-level unemployment rates to individual postal codes of residence, which are only available in a restricted-use data set.

the squared terms of any non-binary covariates. The $p$-values of the tests for the various outcomes are calculated based on 1000 bootstrap replications.[12]

In a first step, we apply the test to a treatment interval of $T \in [40, 3000]$, where choosing 3000 hours of training as the upper bound of the analysis is motivated by the quickly decreasing number of observations beyond that point.

Table 5: Test statistic and p-value, $40 \leq T \leq 3000$

| | N=40, $t_1 > t_2$ | | | | N=50, $t_1 > t_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| $H_0$ : | $\mu(t_1) \geq \mu(t_2)$ | | $\mu(t_1) \leq \mu(t_2)$ | | $\mu(t_1) \geq \mu(t_2)$ | | $\mu(t_1) \leq \mu(t_2)$ | |
| $Y$ | stat | p-value | stat | p-value | stat | p-value | stat | p-value |
| earny4 | 0.001 | 1.000 | 7.205 | 0.000 | 0.001 | 1.000 | 6.078 | 0.000 |
| earnq16 | 0.001 | 1.000 | 8.740 | 0.000 | 0.001 | 1.000 | 8.435 | 0.000 |
| hrswq16 | 0.001 | 1.000 | 9.985 | 0.000 | 0.001 | 1.000 | 9.613 | 0.000 |
| work208 | 0.001 | 0.997 | 10.397 | 0.000 | 0.001 | 0.998 | 9.478 | 0.000 |

Note: Outcomes 'earny4', 'earnq16', 'hrswq16', and 'work208' are weekly earnings in the fourth year, earnings and hours worked per week in quarter 16, and a binary employment indicator four years after assignment (i.e., in week 208). 'Stat' denotes the test statistic.

Table 5 reports the test statistics and p-values for all outcomes under both null hypotheses of weakly increasing mean potential outcomes in the treatment ($\mu(t_1) \geq \mu(t_2)$ for $t_1 > t_2$) and weakly decreasing mean potential outcomes ($\mu(t_1) \leq \mu(t_2)$), respectively. Our tests clearly reject the latter hypothesis of weakly negative monotonicity for any labor market outcome at the 1% level of statistical significance. In contrast, weak positive monotonicity is never rejected, as any test yields p-values close to or equal to 1 (or 100%). Our findings therefore suggest that an increase in the treatment does either increase or at least not reduce the outcome over the treatment range $T \in [40, 3000]$.

It is worth mentioning that the concavities in the outcome-treatment associations spotted in the previously mentioned empirical applications suggest decreasing marginal effects when increasing the treatment. In our testing context, this implies that weakly negative monotonicity should be more clearly rejected for lower rather than higher ranges of treatment values by our method. To verify this suspicion, in a second step we partition the treatment support into three sets of $[40, 1000]$, $[1000, 2000]$, and $[2000, 3000]$ and run the tests separately within each set.

Table 6 presents the results for $T \in [40, 1000]$. None of the tests rejects weakly positive monotonicity at any conventional level of significance, while all tests strongly reject weakly negative

[12] In our empirical study, we do not get unstable $\nu(\ell)$ estimates, so we decided not to apply the trimming method. Also, we note that all estimated generalized propensity scores are greater than 0.0001 in our empirical study.

Table 6: Test statistic and p-value, $40 \leq T \leq 1000$

| | N=40, $t_1 > t_2$ | | | | N=50, $t_1 > t_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| $H_0$ : | $\mu(t_1) \geq \mu(t_2)$ | | $\mu(t_1) \leq \mu(t_2)$ | | $\mu(t_1) \geq \mu(t_2)$ | | $\mu(t_1) \leq \mu(t_2)$ | |
| $Y$ | stat | p-value | stat | p-value | stat | p-value | stat | p-value |
| earny4 | 0.004 | 0.750 | 11.402 | 0.000 | 0.004 | 0.750 | 11.562 | 0.000 |
| earnq16 | 0.017 | 0.535 | 5.556 | 0.000 | 0.016 | 0.540 | 5.427 | 0.000 |
| hrswq16 | 0.007 | 0.631 | 7.157 | 0.000 | 0.007 | 0.666 | 6.998 | 0.000 |
| work208 | 0.001 | 0.991 | 11.675 | 0.000 | 0.001 | 0.985 | 11.081 | 0.000 |

Note: Outcomes 'earny4', 'earnq16', 'hrswq16', and 'work208' are weekly earnings in the fourth year, earnings and hours worked per week in quarter 16, and a binary employment indicator four years after assignment (i.e., in week 208). 'Stat' denotes the test statistic.

Table 7: Test statistic and p-value, $1000 \leq T \leq 2000$

| | N=40, $t_1 > t_2$ | | | | N=50, $t_1 > t_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| $H_0$ : | $\mu(t_1) \geq \mu(t_2)$ | | $\mu(t_1) \leq \mu(t_2)$ | | $\mu(t_1) \geq \mu(t_2)$ | | $\mu(t_1) \leq \mu(t_2)$ | |
| $Y$ | stat | p-value | stat | p-value | stat | p-value | stat | p-value |
| earny4 | 0.075 | 0.672 | 0.485 | 0.206 | 0.076 | 0.631 | 0.468 | 0.238 |
| earnq16 | 0.554 | 0.200 | 0.029 | 0.860 | 0.524 | 0.207 | 0.038 | 0.814 |
| hrswq16 | 0.563 | 0.183 | 0.088 | 0.580 | 0.552 | 0.194 | 0.088 | 0.552 |
| work208 | 0.419 | 0.226 | 0.232 | 0.393 | 0.415 | 0.225 | 0.264 | 0.346 |

Note: Outcomes 'earny4', 'earnq16', 'hrswq16', and 'work208' are weekly earnings in the fourth year, earnings and hours worked per week in quarter 16, and a binary employment indicator four years after assignment (i.e., in week 208). 'Stat' denotes the test statistic.

monotonicity. For the intermediate treatment range of $[1000, 2000]$ considered in Table 7, however, neither positive nor negative monotonicity is ever rejected at the 10% level of statistical significance. This implies that marginal treatment effects are generally less positive than for lower values of $T$. The same findings apply to the highest treatment bracket $[2000, 3000]$, where all tests yield p-values which are beyond conventional levels of significance. Summing up, our empirical findings are consistent with a concave mean potential outcome-treatment dependence, implying that initially strongly positive marginal treatment effects decrease as the treatment value considered (hours in training) increases. A potential explanation for the concavity could be that individuals attending more training in the first year might be induced to attain more education also in the following years rather than to participate in the labor market.

Table 8: Test statistic and p-value, $2000 \leq T \leq 3000$

| | N=40, $t_1 > t_2$ | | | | N=50, $t_1 > t_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| $H_0$ : | $\mu(t_1) \geq \mu(t_2)$ | | $\mu(t_1) \leq \mu(t_2)$ | | $\mu(t_1) \geq \mu(t_2)$ | | $\mu(t_1) \leq \mu(t_2)$ | |
| $Y$ | stat | p-value | stat | p-value | stat | p-value | stat | p-value |
| earny4 | 0.029 | 0.600 | 0.487 | 0.211 | 0.023 | 0.641 | 0.472 | 0.199 |
| earnq16 | 0.008 | 0.889 | 0.591 | 0.178 | 0.007 | 0.876 | 0.543 | 0.210 |
| hrswq16 | 0.132 | 0.353 | 0.205 | 0.353 | 0.149 | 0.346 | 0.176 | 0.385 |
| work208 | 0.465 | 0.229 | 0.020 | 0.723 | 0.457 | 0.231 | 0.014 | 0.758 |

Note: Outcomes 'earny4', 'earnq16', 'hrswq16', and 'work208' are weekly earnings in the fourth year, earnings and hours worked per week in quarter 16, and a binary employment indicator four years after assignment (i.e., in week 208). 'Stat' denotes the test statistic.

# 6 Testing Monotonicity Conditional on Covariates

In this section, we adapt our method to testing monotonicity with conditional mean potential outcomes given continuous covariates $X_1 \subseteq X = (X_1', X_2')'$. In this case, the null hypothesis considered corresponds to

$$H_0: \ \mu(t_1, x_1) \geq \mu(t_2, x_1), \quad \text{for all } t_1 \geq t_2, \text{ for } t_1, t_2 \in [0,1] \text{ and } \ x_1 \in \mathcal{X}_1, \tag{6.1}$$

where $\mu(t, x_1) = E[Y(t)|X_1 = x_1]$ is the conditional average of the potential outcome function or the average dose response function. The conditional ATE (CATE) of a continuous treatment can be defined as $\mu(t_2, x_1) - \mu(t_1, x_1)$ when the continuous treatment is changed from $t_1$ to $t_2$. The identification of $\mu(t, x_1)$ follows the same arguments in the unconditional ADF $\mu(t)$. For example,

$X_1$ could be "age," which is sometimes treated as continuous variable in empirical studies, and we can study the heterogeneous effect over different subpopulations defined by age.[13]

We allow $X_2$ to be potentially high-dimensional and let $X_1$ be of fixed dimension. For simplicity and without loss of generality, we henceforth assume that $X_1$ is a scalar with $\mathcal{X}_1 = [0,1]$. By Lemma 2.1 of Hsu and Shen (2020), $H_0$ in (6.1) is equivalent to

$$\int_{x_1}^{x_1+q^{-1}} \int_{t_2}^{t_2+q^{-1}} \mu(s,\tilde{x}_1)h(s,\tilde{x}_1)dsd\tilde{x}_1 \cdot \int_{x_1}^{x_1+q^{-1}} \int_{t_1}^{t_1+q^{-1}} h(s,\tilde{x}_1)dsd\tilde{x}_1 - \int_{x_1}^{x_1+q^{-1}} \int_{t_1}^{t_1+q^{-1}} \mu(s,\tilde{x}_1)h(s,\tilde{x}_1)dsd\tilde{x}_1 \cdot \int_{x_1}^{x_1+q^{-1}} \int_{t_2}^{t_2+q^{-1}} h(s,\tilde{x}_1)dsd\tilde{x}_1 \leq 0 \tag{6.2}$$

for any $q = 2, \cdots$, and for any $t_1 \geq t_2$ such that $t_1, t_2, x_1 \in \{0, 1/q, 2/q, \cdots, 1-1/q\}$. Similar to the unconditional case, define

$$\mathcal{L}_x = \Big\{ \ell_x = (t_1, t_2, x_1, q^{-1}) : (t_1,t_2) \in [0,1]^2, t_1 > t_2, q = 2,3,\cdots, \\ q \cdot (t_1, t_2, x_1) \in \{0,1,2,\cdots,q-1\}^3 \Big\} \tag{6.3}$$

and $\nu_j(\ell_x) \equiv \int_{x_1}^{x_1+q^{-1}} \int_{t_j}^{t_j+q^{-1}} \mu(s,\tilde{x}_1)h(s,\tilde{x}_1)dsd\tilde{x}_1$ for $\ell_x = (t_1, t_2, x_1, q^{-1})$.

We derive an identifying moment function for $\nu_j(\ell_x)$ that is doubly robust and Neyman orthogonal, based on the Gateaux derivative of $\mu(t, x_1)$ derived in (A.3) in Appendix. We choose the weighting function to be the density function of $X_1$, i.e., $h(t,x) = f_{X_1}(x_1)$. So we avoid estimating the additional nuisance function $f_{X_1}(x_1)$ in the moment function. This permits establishing the following lemma.

**Lemma 6.1** *Suppose Assumption 2.1 holds. Assume that $\mu(t, x_1)$ is continuous in $t$ for all $x_1 \in [0,1]$. Then $H_0$ in (6.1) is equivalent to*

$$H_0' : \nu(\ell_x) = \nu_2(\ell_x) - \nu_1(\ell_x) \leq 0 \text{ for any } \ell_x = (t_1, t_2, x_1, q^{-1}) \in \mathcal{L}_x, \tag{6.4}$$

*where $\nu_j(\ell_x) = E\left[\phi_{j,q}(Z) 1(X_1 \in [x_1, x_1 + q^{-1}])\right]$ for $j = 1, 2$, with $\phi_{j,q}(Z)$ given in (2.6).*

Similar to Theorem 3.1, we can show that uniformly over $\ell_x \in \mathcal{L}_x$, $\sqrt{n}(\hat{\nu}(\ell_x) - \nu(\ell_x)) = n^{-1/2} \sum_{i=1}^n \phi_{\ell_x}(Z_i) + o_P(1)$. As the unconditional case in Section 3, we implement our test by the following algorithm.

Step 1. **(Nuisance functions)** For some fixed $K \in \{2, ..., n\}$, a $K$-fold cross-fitting partitions

[13] We focus on the case when $X_1$ is a continuous variable. If $X_1$ is a discrete variable taking on a finite number of values, then we can just split the sample and conduct a joint test over different values of $X_1$.

the observation indices into $K$ distinct groups $I_k$, $k = 1, ..., K$, such that the sample size of each group is the largest integer smaller than $n/K$. For $k \in \{1, ..., K\}$, the estimators $\hat{\gamma}_k(t, x)$ and $\hat{p}_k(t, x)$ use observations not in $I_k$ and satisfy Assumption 3.1.

Step 2. **(DML estimator)** $\hat{\nu}(\ell_x) = n^{-1} \sum_{i=1}^{n} \left( \hat{\phi}_{\ell_x 2}(Z_i) - \hat{\phi}_{\ell_x 1}(Z_i) \right)$, where $\hat{\phi}_{\ell_x j}(Z) = \hat{\phi}_{j,q}(Z) 1(X_1 \in [x_1, x_1 + q^{-1}])$ and $\phi_{j,q}(Z)$ given in (2.6), for $j = 1, 2$.

Step 3. **(Test statistic)** $\hat{\sigma}^2_{\nu}(\ell_x) = n^{-1} \sum_{i=1}^{n} \hat{\phi}^2_{\ell_x}(Z_i)$, where $\hat{\phi}_{\ell_x}(Z_i) = \hat{\phi}_{\ell_x 2}(Z_i) - \hat{\phi}_{\ell_x 1}(Z_i) - \hat{\nu}(\ell_x)$. $\hat{\sigma}_{\nu,\epsilon}(\ell_x) = \max\{\hat{\sigma}_{\nu}(\ell_x), \epsilon \cdot \hat{\sigma}_{\nu}(0, 1/2, 0, 1/2)\}$.

Compute the Cramér-von Mises test statistic $\widehat{T}_{x_1} = \sum_{\ell_x \in \mathcal{L}_x} \max\left\{ \sqrt{n} \frac{\hat{\nu}(\ell_x)}{\hat{\sigma}_{\nu,\epsilon}(\ell_x)}, 0 \right\}^2 Q(\ell_x)$, where $Q$ is a weighting function such that $Q(\ell_x) > 0$ for all $\ell_x \in \mathcal{L}_x$ and $\sum_{\ell_x \in \mathcal{L}_x} Q(\ell_x) < \infty$.

Step 4. **(Critical values)** Let $\{U_i : 1 \leq i \leq n\}$ be a sequence of i.i.d. random variables that satisfy Assumption 3.2.

The simulated process is constructed as $\widehat{\Phi}^u_{\nu,x}(\ell_x) = n^{-1/2} \sum_{i=1}^{n} U_i \cdot \hat{\phi}_{\ell_x}(Z_i)$, where $\hat{\phi}_{\ell_x}(Z_i)$ is the estimated influence function in Step 3.

$$\hat{\psi}_{\nu}(\ell_x) = -B_n \cdot 1\left( \sqrt{n} \cdot \frac{\hat{\nu}(\ell_x)}{\hat{\sigma}_{\nu,\epsilon}(\ell_x)} < -a_n \right).$$

where $a_n$ and $B_n$ satisfy Assumption 3.3. The critical value is

$$\hat{c}^{\eta}_{x_1}(\alpha) = \sup\left\{ q \middle| P^u\left( \sum_{\ell_x \in \mathcal{L}_x} \max\left\{ \frac{\widehat{\Phi}^u_{\nu,x}(\ell_x)}{\hat{\sigma}_{\nu,\epsilon}(\ell_x)} + \hat{\psi}_{\nu}(\ell_x), 0 \right\} Q(\ell_x) \leq q \right) \leq 1 - \alpha + \eta \right\} + \eta,$$

where $P^u$ denotes the multiplier probability measure given the observed samples.

Step 5. **(Decision rule)** Reject $H_0'$ if $\widehat{T}_{x_1} > \hat{c}^{\eta}_{x_1}(\alpha)$.

The size and power properties are similar to the unconditional potential outcome cases, and the details are omitted for brevity.

**Remark 6.2** Our test has the advantage of being easily extended to the conditional case, compared with the uniform inference method based on Su et al. (2019) considered in our simulation. This is because such a supremum-type test would require nonparametric estimation of the $d\mu(t, x_1)/dx$ that is not a trivial extension of Su et al. (2019). More specifically, based on the Gateaux derivative limit of $\mu(t, x_1) = E[Y(t)|X_1 = x_1]$ derived in (A.3), we can extend the DML

estimator of $E[Y(t)]$ in Colangelo and Lee (2022) to estimate $\mu(t, x_1)$ by

$$\hat{\mu}(t, x_1) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \left\{ \hat{\gamma}_k(t, X_i) + \frac{Y_i - \hat{\gamma}_k(t, X_i)}{\hat{p}_k(t|X_i)} K_h(T_i - t) \right\} \frac{K_{h_x}(X_{1i} - x_1)}{\hat{f}_{X_1}(x_1)},$$

where $K_h(T-t) \equiv k((T-t)/h)/h$ with a suitable second-order symmetric kernel function $k(\cdot)$ and a bandwidth $h$, and $\hat{f}_{X_1}(x_1) = n^{-1} \sum_{i=1}^{n} K_{h_x}(X_{1i} - x_1)$. A uniform inference theory for $\hat{\mu}(t, x_1)$ or $\partial \hat{\mu}(t, x_1) \partial t$ could be an alternative approach to our method. This interesting extension is beyond the scope of the paper and is worthy of a separate research project.

# 7 Conclusion

In this paper, we propose Cramér-von Mises-type tests for testing whether a mean potential outcome is weakly monotonic in a continuously distributed treatment under a weak unconfoundedness assumption. To flexibly employ nonparametric or machine learning estimators in the presence of possibly high-dimensional nuisance parameters, we propose a double debiased machine learning estimator for the moments entering the test. Furthermore, we extend our method to testing monotonicity conditional on observed covariates. We also investigate the test's finite sample behavior in a simulation study and find that it performs decently under our suggested choices of tuning parameters.

As an empirical illustration, we apply our test to the Job Corps study, investigating the associations of several labor market outcomes (earnings, employment, and hours worked) with hours in training as the treatment. We find that an increase in the treatment does either increase or at least not reduce the outcome. When splitting the treatment range into subsets, our testing results are consistent with a concave mean potential outcome-treatment dependence, implying that initially stronger marginal treatment effects decrease as the treatment value (i.e., hours already spent in training) increases.

# APPENDIX

## A Gateaux derivative limit

Let $f^0$ be the true pdf of $Z = (Y, T, X)$ and $f_Z^h$ be a pdf approaching a point mass at $Z$ as $h \to 0$. Consider $f^{\tau h} = (1-\tau)f^0 + \tau f_Z^h$ for $\tau \in [0,1]$. Colangelo and Lee (2022) derive the Gateaux derivative of $\mu(t)$ with respect to a deviation from the true distribution $f_Z^h - f^0$ to be $\gamma(t,X) - \mu(t) + \frac{Y-\gamma(t,X)}{p(t,X)} f_T^h(t)$. Since $\nu(t,r) \equiv \int_t^{t+r} \mu(s)ds$ is a linear functional of $\mu$, the Gateaux derivative limit of $\nu(t,r)$ is

$$\begin{aligned}
&\lim_{h\to 0} \int_t^{t+r} \left\{ \gamma(s,X) - \mu(s) + \frac{Y-\gamma(s,X)}{p(s,X)} f_T^h(s) \right\} ds \\
&= \int_t^{t+r} \gamma(s,X)ds - \nu(t,r) + \frac{Y-\gamma(T,X)}{p(T,X)} \mathbf{1}(T \in [t, t+r]),
\end{aligned} \tag{A.1}$$

and it follows that

$$\nu(t,r) = E\left[ \int_t^{t+r} \gamma(s,X)ds + \frac{Y-\gamma(T,X)}{p(T,X)} \mathbf{1}(T \in [t, t+r]) \right]. \tag{A.2}$$

**Conditional ADF:** By the same arguments as for the identification of the unconditional ADF $\mu(t)$, we identify the conditional ADF $\mu(t, x_1) = \int_{\mathcal{X}_2} \gamma(t, x_1, x_2) f_{X_2|X_1}(x_2|x_1)dx_2$. Let the Dirac delta function $\delta_t(s) = \infty$ for $s = t$, $\delta_t(s) = 0$ for $s \neq t$, and $\int g(s)\delta_t(s)ds = g(t)$, for any continuous compactly supported function $g$. Write

$$\begin{aligned}
\mu(t, x_1) &= \int_{\mathcal{X}} \int_{\mathcal{T}} \gamma(s, \tilde{x}_1, x_2) \delta_t(s) ds f_{X_2|X_1}(x_2|\tilde{x}_1) \delta_{x_1}(\tilde{x}_1) d\tilde{x}_1 dx_2 \\
&= \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} y \delta_t(s) \delta_{x_1}(\tilde{x}_1) f_{Y|TX}(y|s, \tilde{x}_1, x_2) f_{X_2|X_1}(x_2|\tilde{x}_1) dy ds d\tilde{x}_1 dx_2.
\end{aligned}$$

Following Colangelo and Lee (2022), we derive the Gateaux derivative of $\mu(t, x_1)$ with respect to a deviation from the true distribution $f_Z^h - f^0$ by

$$\begin{aligned}
\frac{d}{d\tau}\mu(t, x_1) &= \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} y \delta_t(s) \delta_{x_1}(\tilde{x}_1) \frac{d}{d\tau} \left( \frac{f_Z(y, s, \tilde{x}_1, x_2) f_X(\tilde{x}_1, x_2)}{f_{TX}(s, \tilde{x}_1, x_2) f_{X_1}(\tilde{x}_1)} \right) dy ds d\tilde{x}_1 dx_2 \\
&= \int_{\mathcal{X}} \int_{\mathcal{T}} \int_{\mathcal{Y}} y \delta_t(s) \delta_{x_1}(\tilde{x}_1) \Bigg\{ \frac{f_X(\tilde{x}_1, x_2)}{f_{TX}(s, \tilde{x}_1, x_2) f_{X_1}(\tilde{x}_1)} \left( -f_Z^0(y, s, \tilde{x}_1, x_2) + f_Z^h(y, s, \tilde{x}_1, x_2) \right) \\
&\quad - \frac{f_Z(y, s, \tilde{x}_1, x_2) f_{X_2|X_1}(x_2|\tilde{x}_1)}{f_{TX}(s, \tilde{x}_1, x_2)^2} \left( -f_{TX}^0(s, \tilde{x}_1, x_2) + f_{TX}^h(s, \tilde{x}_1, x_2) \right)
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{f_Z(y,s,\tilde{x}_1,x_2)}{f_{TX}(s,\tilde{x}_1,x_2)f_{X_1}(\tilde{x}_1)}\Big( - f_X^0(\tilde{x}_1,x_2) + f_X^h(\tilde{x}_1,x_2) \\
&- \frac{f_X(\tilde{x}_1,x_2)}{f_{X_1}(\tilde{x}_1)}\left(-f_{X_1}^0(\tilde{x}_1) + f_{X_1}^h(\tilde{x}_1)\right)\Big)\Big\} dyds d\tilde{x}_1 dx_2 \\
&= -\mu(t,x_1) + \frac{Y f_{TX_1}^h(t,x_1)}{f_{T|X}(t,x_1,X_2)f_{X_1}(x_1)} + \mu(t,x_1) - \frac{\gamma(t,x_1,X_2)f_{TX_1}^h(t,x_1)}{f_{T|X}(t|x_1,X_2)f_{X_1}(x_1)} \\
&- \mu(t,x_1) + \gamma(t,x_1,X_2)\frac{f_{X_1}^h(x_1)}{f_{X_1}(x_1)} + \mu(t,x_1) - \mu(t,x_1)\frac{f_{X_1}^h(x_1)}{f_{X_1}(x_1)} \\
&= \frac{Y-\gamma(t,x_1,X_2)}{f_{T|X}(t|x_1,X_2)}\frac{f_{TX_1}^h(t,x_1)}{f_{X_1}(x_1)} + (\gamma(t,x_1,X_2) - \mu(t,x_1))\frac{f_{X_1}^h(x_1)}{f_{X_1}(x_1)}. \qquad \text{(A.3)}
\end{aligned}
$$

Since $\nu(t,x_1,r) \equiv \int_{x_1}^{x_1+r}\int_t^{t+r}\mu(s,\tilde{x}_1)f_{X_1}(\tilde{x}_1)ds d\tilde{x}_1 = E\left[\int_t^{t+r}\mu(s,X_1)ds\mathbf{1}(X_1\in[x_1,x_1+r])\right]$ is a linear functional of $\mu$, the Gateaux derivative limit of $\nu(t,x_1,r)$ is

$$
\begin{aligned}
&\lim_{h\to 0}\int_{x_1}^{x_1+r}\int_t^{t+r}\left\{(\gamma(s,\tilde{x}_1,X_2)-\mu(s,\tilde{x}_1))\frac{f_{X_1}^h(\tilde{x}_1)}{f_{X_1}(\tilde{x}_1)} + \frac{Y-\gamma(s,\tilde{x}_1,X_2)}{p(s,\tilde{x}_1,X_2)}\frac{f_{TX_1}^h(s,\tilde{x}_1)}{f_{X_1}(\tilde{x}_1)}\right\}f_{X_1}(\tilde{x}_1)ds d\tilde{x}_1 \\
&= \left\{\int_t^{t+r}\gamma(s,X)ds - \int_t^{t+r}\mu(s,X_1)ds + \frac{Y-\gamma(T,X)}{p(T,X)}\mathbf{1}(T\in[t,t+r])\right\}\mathbf{1}(X_1\in[x_1,x_1+r]).
\end{aligned}
$$

It follows that

$$
\nu(t,x_1,r) = E\left[\left\{\int_t^{t+r}\gamma(s,X)ds + \frac{Y-\gamma(T,X)}{p(T,X)}\mathbf{1}(T\in[t,t+r])\right\}\mathbf{1}(X_1\in[x_1,x_1+r])\right]. \qquad \text{(A.4)}
$$

So we obtain $\nu_j(\ell_x) = \nu(t_j,x_1,1/q) = E\left[\phi_{j,q}(Z)\mathbf{1}(X_1\in[x_1,x_1+r])\right]$.

# B Appendix for Section 3

**Proof of Theorem 3.1**:
We give an outline of deriving the asymptotically linear representation, following Chernozhukov et al. (2018). Let $\nu(t,r) = \int_t^{t+r}\mu(s)ds$, $\gamma_i \equiv \gamma(T_i,X_i)$, and $\lambda_i \equiv \lambda(T_i,X_i) = 1/f_{T|X}(T_i|X_i)$. Let

$$
\phi_{(t,r)}(Z_i,\gamma,\lambda) \equiv \int_t^{t+r}\gamma(s,X_i)ds + (Y_i - \gamma(T_i,X_i))\lambda(T_i,X_i)\mathbf{1}(T_i\in[t,t+r]).
$$

So $\hat{\nu}(t,r) = n^{-1}\sum_{i=1}^{n}\phi_{(t,r)}(Z_i,\hat{\gamma},\hat{\lambda})$. To show Theorem 3.1, it is sufficient to show that uniformly over $(t,r)\in[0,1]^2$,

$$\sqrt{n}(\hat{\nu}(t,r)-\nu(t,r)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_{t,r}(Z_i,\gamma,\lambda)-\nu(t,r)+o_P(1). \tag{B.1}$$

Let $Z_k^c$ denote the observations $Z_i$ for $i\neq I_k$ and $\hat{\gamma}_{ik}=\hat{r}_k(T_i,X_i)$ using $Z_k^c$ for $i\in I_k$. We decompose the remainder term

$$\begin{aligned}
&\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\Big\{\hat{\phi}_{(t,r)}(Z_i,\hat{\gamma},\hat{\lambda})-\phi_{(t,r)}(Z_i,\gamma,\lambda)\Big\}\\
&=\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in I_k}\Bigg\{\int_t^{t+r}(\hat{\gamma}_k(s,X_i)-\gamma(s,X_i))\,ds - E\left[\int_t^{t+r}(\hat{\gamma}_k(s,X_i)-\gamma(s,X_i))\,ds\middle|Z_k^c\right] && \text{(R1-1)}\\
&\quad+\mathbf{1}(T_i\in[t,t+r])\lambda_i(\gamma_i-\hat{\gamma}_{ik})-E\big[\mathbf{1}(T_i\in[t,t+r])\lambda_i(\gamma_i-\hat{\gamma}_{ik})\big|Z_k^c\big] && \text{(R1-2)}\\
&\quad+\mathbf{1}(T_i\in[t,t+r])(\hat{\lambda}_{ik}-\lambda_i)(Y_i-\gamma_i)-E\big[\mathbf{1}(T_i\in[t,t+r])(\hat{\lambda}_{ik}-\lambda_i)(Y_i-\gamma_i)\big|Z_k^c\big]\Bigg\} && \text{(R1-3)}\\
&\quad+\sqrt{n}\Bigg\{E\left[\int_t^{t+r}(\hat{\gamma}_k(s,X_i)-\gamma(s,X_i))ds\middle|Z_k^c\right]-E\left[\mathbf{1}(T_i\in[t,t+r])\lambda_i(\hat{\gamma}_{ik}-\gamma_i)|Z_k^c\right]\\
&\quad+E[(\hat{\lambda}_{ik}-\lambda_i)\mathbf{1}(T_i\in[t,t+r])(Y_i-\gamma_i)|Z_k^c]\Bigg\} && \text{(R1-DR)}\\
&\quad-\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in I_k}\mathbf{1}(T_i\in[t,t+r])\big(\hat{\lambda}_{ik}-\lambda_i\big)\big(\hat{\gamma}_{ik}-\gamma_i\big). && \text{(R2)}
\end{aligned}$$

The remainder terms (R1-1), (R1-2) and (R1-3) are stochastic equicontinuous terms that are controlled to be $o_P(1)$ by the mean-squared consistency conditions in Assumption 3.1(i) and cross-fitting. The second-order remainder term (R2) is controlled by Assumption 3.1(ii). By the law of iterated expectations, $E\Big[\int_t^{t+r}\big(\hat{\gamma}_k(s,X)-\gamma(s,X)\big)ds\Big|Z_k^c\Big] = E\Big[\lambda(T,X)\big(\hat{\gamma}_k(T,X)-\gamma(T,X)\big)\mathbf{1}(T\in[t,t+r])\Big|Z_k^c\Big]$. So (R1-DR) is zero.

To show (R1-1), (R1-2) and (R1-3) are $o_P(1)$ uniformly over $\ell$, we show these terms weakly converge to Gaussian processes indexed by $\ell$ with zero covariance kernel. It suffices to show the results with $\mathbf{1}(T_i\leq t)$ replacing $\mathbf{1}(T_i\in[t,t+r])$. We apply the functional central limit theorem in Theorem 10.6 in Pollard (1990). Following the notation in Pollard (1990), for any $\omega$ in the probability space $\Omega$ and for $i\in I_k$, define $f_i(t)=f_i(\omega,t)=\mathbf{1}(T_i\leq t)\lambda_i(\hat{\gamma}_{ik}-\gamma_i)$ for (R1-2) and $f_{ni}(t)=f_i(t)/\sqrt{n}$. Due to cross-fitting, the processes from the triangular array $\{f_{ni}(t)\}$ given $Z_k^c$ are independent within rows. Let $n_k=\sum_{i=1}^{n}\mathbf{1}(i\in I_k)$. Since $K$ is fixed, $n/n_k=O(1)$. We verify

the conditions in Theorem 10.6 in Pollard (1990).

(i) $\{\mathbf{1}(T_i \leq t) : t \in [0,1], i \in I_k\}$ is manageable since it is monotone increasing in $t$ (p.221 in Kosorok (2008)). The triangular array processes $\{f_{ni}(t)\}$ are manageable with respect to the envelopes $F_{ni} = \left|\lambda_i(\hat{\gamma}_{ik} - \gamma_i)\right|/\sqrt{n}$. $F_{n_k} = (F_{n1}, ..., F_{nn_k})'$ is an $R^{n_k}$-valued function on the underlying probability space.

(ii) Let $\mathsf{X}_n(t) = \mathsf{X}_n(\omega, t) = \sum_{i \in I_k} \left(f_{ni}(t) - E\left[f_{ni}(t)\middle|Z_k^c\right]\right)$. By construction and independence of $Z_k^c$ and $z_i, i \in I_k$, $E[f_{ni}(t)|Z_k^c] = 0$ and $E[f_{ni}(t)f_{nj}(t)|Z_k^c] = 0$ for $i, j \in I_k$. For $i \in I_k$, $E[f_i(t)^2|Z_k^c] = O_P(\|\hat{\gamma}_{ik} - \gamma_i)\|_2^2) = o_P(1)$ by Assumption 3.1(i) and (iii). Let $s \leq t \in [0,1]$, without loss of generality. $H(s,t) = \lim_{n\to\infty} E\left[\mathsf{X}_n(s)\mathsf{X}_n(t)\middle|Z_k^c\right] = \lim_{n\to\infty} E\left[\mathbf{1}(T \in (s,t])\lambda_i^2(\hat{\gamma}_{ik} - \gamma_i)^2\middle|Z_k^c\right] = 0$.

(iii) By the argument in (ii), $H(t,t) = 0$.

(iv) For each $\epsilon > 0$,

$$\sum_{i \in I_k} E[F_{ni}^2 1(F_{ni} \geq \epsilon)|Z_k^c] \leq \sum_{i \in I_k} E[F_{ni}^2|Z_k^c] = O_P\left(\|\hat{\gamma}_k - \gamma\|_2^2\right) = o_P(1).$$

(v) For any $s < t$,

$$\begin{aligned}
\rho_n(s,t)^2 &= \sum_{i \in I_k} E\left[\left|f_{ni}(s) - f_{ni}(t)\right|^2 \middle| Z_k^c\right] \\
&= \sum_{i \in I_k} E[1(T_i \in (s,t])\lambda_i(\hat{\gamma}_{ik} - \gamma_i)^2|Z_k^c] \\
&= \sum_{i \in I_k} \int_{\mathcal{X}} \int_s^t \lambda(T_i, X_i)(\hat{\gamma}_k(T_i, X_i) - \gamma(T_i, X_i))^2 f_{TX}(T_i, X_i) dT_i dX_i \\
.I &= O_P\left(\int_{\mathcal{X}} \int_{\mathcal{T}} (\hat{\gamma}_k(T_i, X_i) - \gamma(T_i, X_i))^2 f_{TX}(T_i, X_i) dT_i dX_i\right) \\
&= O_P\left(\|\hat{\gamma}_k - \gamma\|_2^2\right)
\end{aligned}$$

by the definition of $\|\cdot\|_2$ and Assumption 3.1(iii). By Assumption 3.1(i), $\rho(s,t) = \lim_{n\to\infty} \rho_n(s,t) = 0$. The condition (v) holds: for all deterministic sequences $\{s_n\}$ and $\{t_n\}$, if $\rho(s_n, t_n) \to 0$ then $\rho_n(s_n, t_n) \to 0$.

Then Theorem 10.6 in Pollard (1990) implies that the finite dimensional distributions of $\mathsf{X}_n$ have Gaussian limits, with zero means and covariances given by $H$. Therefore, $\mathsf{X}_n = o_P(1)$ uniformly over $t \in [0,1]$.

The analogous results also hold for $f_i(t) = \mathbf{1}(T_i \leq t)(\hat{\lambda}_{ik} - \lambda_i)(Y_i - \gamma_i)$ in (R1-3). In particular, for (R1-3), $E[f_{ni}(t)^2|Z_k^c] = O_P\left(\|\hat{\lambda}_k - \lambda\|_2^2\right) = o_P(1)$ by the smoothness condition and Assumption 3.1(i).

For (R1-1), define $f_i(t) = \int_0^t (\hat{\gamma}_k(s, X_i) - \gamma(s, X_i))\, ds$.

Note that we can express $\int_t^{t+r} \gamma_k(s, X_i)ds = E\left[\gamma(T, X)\mathbf{1}(T \in [t, t+r])/p(T, X)|X = X_i\right]$. So

$$\begin{aligned}
E[f_i(t)^2|Z_k^c] &\leq \int \left(E\left[\frac{(\hat{\gamma}_k(T, X) - \gamma(T, X))}{p(T, X)}\mathbf{1}(T \leq t)\Big|X = x\right]\right)^2 f_X(x)dx \\
&\leq \int E\left[\left(\frac{\hat{\gamma}_k(T, X) - \gamma(T, X)}{p(T, X)}\mathbf{1}(T \leq t)\right)^2 \Big|X = x\right] f_X(x)dx \\
&= \int\int \left(\frac{\hat{\gamma}_k(s, x) - \gamma(s, x)}{p(s, x)}\right)^2 \mathbf{1}(s \leq t)p(s, x)dsdx \\
&= O_P\left(\int\int (\hat{\gamma}_k(s, x) - \gamma(s, x))^2 p(s, x)dsdx\right) \\
&= o_P(1)
\end{aligned}$$

and the last equality holds by Assumption 3.1(i).

For (R2),

$$\begin{aligned}
&E\left[\sup_\ell \left|n^{-1/2}\sum_{i\in I_k} \mathbf{1}(T_i \in [t, t+r])(\hat{\lambda}_{ik} - \lambda_i)(\gamma_i - \hat{\gamma}_{ik})\right|\Big|Z_k^c\right] \\
&\leq \sqrt{n}\int_{\mathcal{X}}\int_{\mathcal{T}} \sup_\ell \mathbf{1}(T_i \in [t, t+r])\Big|(\hat{\lambda}_{ik} - \lambda_i)(\gamma_i - \hat{\gamma}_{ik})\Big| f_{TX}(T_i, X_i)dT_i dX_i \\
&\leq \sqrt{n}\Big(\int_{\mathcal{X}}\int_{\mathcal{T}} (\hat{\lambda}_{ik} - \lambda_i)^2 f_{TX}(T_i, X_i)dT_i dX_i\Big)^{1/2}\Big(\int_{\mathcal{X}}\int_{\mathcal{T}} (\hat{\gamma}_{ik} - \gamma_i)^2 f_{TX}(T_i, X_i)dT_i dX_i\Big)^{1/2} \\
&\xrightarrow{p} 0
\end{aligned} \tag{B.2}$$

by Cauchy-Schwartz inequality and Assumption 3.1(ii). By the conditional Markov and triangle inequalities, (R2)$\xrightarrow{p} 0$ uniformly over $\ell$.

The triangle inequality yields the asymptotically linear representation $n^{-1/2}\sum_{i=1}^n \left(\hat{\phi}_{t,r}(Z_i, \hat{\gamma}, \hat{\lambda}) - \phi_{t,r}(Z_i, \gamma, \lambda)\right) = o_P(1)$, and (B.1) follows.

Then by the fact that $\nu(\ell) = \nu(t_1, q^{-1}) - \nu(t_2, q^{-1})$, it follows that uniformly over $\ell \in \mathcal{L}$, $\sqrt{n}(\hat{\nu}(\ell) - \nu(\ell)) = n^{-1/2}\sum_{i=1}^n \phi_\ell(Z_i) + o_P(1)$, and this shows the first half of Theorem 3.1.

For the second part, similar to Hsu et al. (2019), it is straightforward to see that $\{\phi_\ell(Z) : \ell \in \mathcal{L}\}$ is a VC class of functions and by functional central limit theorem of Pollard (1990), it follows that

$\sqrt{n}(\hat{\nu}(\cdot) - \nu(\cdot)) \Rightarrow \Phi_{h_{DML}}(\cdot)$ where $\Phi_{h_{DML}}(\cdot)$ is a Gaussian process with variance-covariance kernel $h_{DML}(\ell_1, \ell_2) = E[\phi_{\ell_1}(Z)\phi_{\ell_2}(Z)]$.

Finally, the approximation error of the Riemann sum is

$$\left|M^{-1}\sum_{m=1}^{M}\hat{\gamma}_k(t_m, X_i) - \int_t^{t+r}\hat{\gamma}_k(s, X_i)ds\right| \leq M^{-1}\sum_{m=1}^{M}\left|\hat{\gamma}_k(t_m, X_i) - \hat{\gamma}_k(t_{m-1}, X_i)\right| = O_P(M^{-1}),$$

assuming finite total variation of $\hat{\gamma}_k$ with probability approach one. By the condition $\sqrt{n}/M \to 0$, the approximation error is asymptotically ignorable. □

**Lemma B.1** *Suppose the Assumptions 2.1, 3.1 and 3.2 hold. Then,* $\sup_{\ell\in\mathcal{L}}|\hat{\sigma}_\nu(\ell) - \sigma_\nu(\ell)| \xrightarrow{p} 0$ *where* $\sigma_\nu^2(\ell) = E[\phi_\ell^2]$*, and* $\widehat{\Phi}_\nu^u \Rightarrow \Phi_{h_{DML}}$ *conditional on sample path w.p.a.*1.

**Proof of Lemma B.1:**
The fact that $\{\phi_\ell : \ell \in \mathcal{L}\}$ is a VC type class of functions implies that $\{\phi_\ell^2 : \ell \in \mathcal{L}\}$ is also a VC type. In addition, given that $E[\bar{\phi}^{2+\delta}] < \infty$, we have by the uniform weak law of large numbers that $\sup_{\ell\in\mathcal{L}}|\tilde{\sigma}_\nu^2(\ell) - \sigma_\nu^2(\ell)| \xrightarrow{p} 0$, where $\tilde{\sigma}_\nu^2(\ell) = n^{-1}\sum_{i=1}^n \phi_\ell^2(Z_i)$. By Assumption 3.1, we have that $\sup_{\ell\in\mathcal{L}}|\tilde{\sigma}_\nu^2(\ell) - \hat{\sigma}_\nu^2(\ell)| \xrightarrow{p} 0$. Then the first part follows. The proof of the second part follows from the standard arguments for the multiplier bootstrap such as Lemma 4.1 of Hsu (2017), and is omitted for the sake of brevity. □

**Proof of Theorem 3.2:**
The proof of Theorem 3.2 follows from the same arguments as Theorem 5.1 of Hsu (2017) once Theorem 3.1 and Lemma B.1 are established, and is omitted for the sake of brevity. □

# C The SUZ method

Equation (3.2) in SUZ or Equation (1) in Colangelo and Lee (2022) shows that

$$\mu(t) \equiv E[Y(t)] = \lim_{h\to 0} E\left[\frac{(Y - \gamma(t,x))K_h(T_j - t)}{p(t,x)} + \gamma(t,x)\right],$$

where $p(t,x) \equiv f_{T|X}(t|x)$, $\gamma(t,x) \equiv E[Y|X = x, T = t]$, and $K_h(T-t) \equiv k((T-t)/h)/h$ uses a suitable second-order symmetric kernel function $k(\cdot)$ with a bandwidth $h$.

SUZ propose the following three-stage procedure to estimate $\mu(t)$ and the average partial effect $\theta(t) \equiv d\mu(t)/dt$. We briefly describe the estimation procedure proposed in SUZ and refer readers to SUZ for details.

1. Estimate $\gamma(t,x)$ and $p(t,x)$ by $\hat{\gamma}(t,x)$ and $\hat{p}(t,x)$ respectively with the first-stage bandwidth $h_1$.

2. Estimate $\mu(t)$ by

$$\hat{\mu}(t) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y-\hat{\gamma}(t,X_i)}{\hat{p}(t,X_i)}K_{h_2}(T_i-t)+\hat{\gamma}(t,X_i)\right\}$$

   with the second-stage bandwidth $h_2$.

3. Estimate $\theta(t)$ by $\hat{\theta}(t)$ which is the estimator of the slope coefficient in the local linear regression of $\hat{\mu}(T_i)$ on $T_i$.

Let $\sigma(t)$ be the standard deviation of $\sqrt{nh_2^3}\hat{\theta}(t)$. We estimate $\sigma(t)$ by $\hat{\sigma}(t)$ based on the asymptotic property of $\hat{\theta}(t)$; see the equation after Theorem 3.5 of SUZ. We next follow Fan et al. (2022) and propose the following algorithm to test the average partial effect $\theta(t)$.

1. Compute $\hat{\theta}(t)$ and $\hat{\sigma}(t)$ for a suitably fine grid over $\mathcal{T}$.

2. Compute $\hat{\theta}^b(t)$, the multiplier bootstrap version of $\hat{\theta}(t)$, over the same grid for $b=1,\ldots,B$, while generating a new set of i.i.d. random variables $\{\eta_i\}_{i=1}^n$ from the distribution of $\eta$ such that it has sub-exponential tails and unit mean and variance in each step $b$.

3. For $b=1,\ldots,B$, compute

$$M_b^{\text{1-sided}} = \sup_{t\in\mathcal{T}}\frac{\sqrt{nh_2^3}(\hat{\theta}^b(t)-\hat{\theta}(t))}{\hat{\sigma}(t)},$$

   where the supremum is approximated by the maximum over the chosen grid points.

4. Given a confidence level $1-\alpha$, find the empirical $(1-\alpha)$ quantile of the sets of numbers $\{M_b^{\text{1-sided}} : b=1,\ldots,B\}$ and denote the quantile as $\hat{C}_\alpha^{\text{1-sided}}$.

5. The decision rule is given by

$$\text{Reject } H_0'' \text{ if } \min_{t\in\mathcal{T}}\left(\hat{\theta}(t)+\hat{C}_\alpha^{\text{1-sided}}\frac{\hat{\sigma}(t)}{\sqrt{nh_2^3}}\right)<0.$$

Table 9: Descriptive statistics for selected covariates, treatment, and outcomes

| variable | mean | median | minimum | maximum | nonmissing |
|---|---|---|---|---|---|
| female | 0.432 | 0.495 | 0.000 | 1.000 | 4166 |
| age | 18.325 | 21.42 | 16.000 | 24.000 | 4166 |
| White | 0.249 | 0.433 | 0.000 | 1.000 | 4166 |
| Black | 0.502 | 0.500 | 0.000 | 1.000 | 4166 |
| Hispanic | 0.172 | 0.378 | 0.000 | 1.000 | 4166 |
| years of education | 10.045 | 1.535 | 0.000 | 20.000 | 4102 |
| married | 0.016 | 0.126 | 0.000 | 1.000 | 4166 |
| has children | 0.178 | 0.382 | 0.000 | 1.000 | 4166 |
| ever worked | 0.145 | 0.352 | 0.000 | 1.000 | 4166 |
| mean gross weekly earnings | 19.429 | 97.749 | 0.000 | 2000.000 | 4166 |
| household size | 3.536 | 2.006 | 0.000 | 15.000 | 4101 |
| mother's years of education | 11.504 | 2.599 | 0.000 | 20.000 | 3397 |
| father's years of education | 11.459 | 2.900 | 0.000 | 20.000 | 2604 |
| welfare receipt during childhood | 2.064 | 1.189 | 1.000 | 4.000 | 3871 |
| poor or fair general health | 0.124 | 0.330 | 0.000 | 1.000 | 4166 |
| physical or emotional problems | 0.043 | 0.203 | 0.000 | 1.000 | 4166 |
| extent of marijuana use | 2.540 | 1.549 | 0.000 | 4.000 | 1534 |
| extent of smoking | 1.526 | 0.971 | 0.000 | 4.000 | 2171 |
| extent of alcohol consumption | 3.140 | 1.210 | 0.000 | 4.000 | 2383 |
| ever arrested | 0.241 | 0.428 | 0.000 | 1.000 | 4166 |
| recruiter support | 1.592 | 1.059 | 1.000 | 5.000 | 4068 |
| idea about desired training | 0.839 | 0.368 | 0.000 | 1.000 | 4166 |
| expected months in Job Corps | 6.622 | 9.794 | 0.000 | 36.000 | 4166 |
| hours in training ($T$) | 1192.130 | 966.945 | 0.857 | 6188.571 | 4166 |
| weekly earnings in fourth year ($Y$) | 215.521 | 202.619 | 0.000 | 1879.172 | 4024 |
| weekly earnings in quarter 16 ($Y$) | 220.933 | 223.078 | 0.000 | 1970.445 | 4015 |
| weekly hours worked in quarter 16 ($Y$) | 28.187 | 22.746 | 0.000 | 84.000 | 4102 |
| employed in week 208 ($Y$) | 0.627 | 0.484 | 0.000 | 1.000 | 4007 |

# References


Andrews, D. W. K. and X. Shi (2013). Inference based on conditional moment inequalities. *Econometrica 81*(2), 609–666.

Andrews, D. W. K. and X. Shi (2014). Nonparametric inference based on conditional moment inequalities. *Journal of Econometrics 179*(1), 31–45.

Athey, S. and G. Imbens (2019). Machine learning methods economists should know about. arxiv:1903.10075v1.

Baraud, Y., S. Huet, and B. Laurent (2005). Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *The Annals of Statistics 23*(1), 214–257.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics 186*(2), 345–366. High Dimensional Problems in Econometrics.

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

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

Belloni, A., V. Chernozhukov, and K. Kato (2019). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association 114*(526), 749–758.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics 37*(4), 1705–1732.

Blundell, R. and J. L. Powell (2003). *Endogeneity in Nonparametric and Semiparametric Regression Models*, Volume II. Cambridge University Press, Cambridge, U.K.

Bowman, A. W., M. C. Jones, and I. Gijbels (1998). Testing monotonicity of regression. *Journal of Computational and Graphical Statistics 7*(4), 489–500.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2022). On binscatter. arxiv:1902.09608.

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2020). Large sample properties of partitioning-based series estimators. *The Annals of Statistics 48*(3), 1718 – 1741.

Chen, X. and H. White (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory 45*.

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

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2018). Locally robust semiparametric estimation. arxiv:1608.00033.

Chernozhukov, V., W. Newey, and R. Singh (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica 90*(3), 967–1027.

Chetverikov, D. (2019). Testing regression monotonicity in econometric models. *Econometric Theory 35*(4), 1146–1200.

Chetverikov, D., Z. Liao, and V. Chernozhukov (2021). On cross-validated Lasso in high dimensions. *The Annals of Statistics 49*(3), 1300–1317.

Colangelo, K. and Y.-Y. Lee (2022). Double debiased machine learning nonparametric inference with continuous treatments. arxiv:2004.03036.

Donald, S. G. and Y.-C. Hsu (2016). Improving the power of tests of stochastic dominance. *Econometric Reviews 35*(4), 553–585.

Donald, S. G., Y.-C. Hsu, and R. P. Lieli (2014). Testing the unconfoundedness assumption via inverse probability weighted estimators of (L)ATT. *Journal of Business & Economic Statistics 32*(3), 395–415.

Dümbgen, L. and V. G. Spokoiny (2001). Multiscale testing of qualitative hypotheses. *The Annals of Statistics 29*(1), 124–152.

Durot, C. (2003). Multiscale testing a Kolmogorov-type test for monotonicity of regressionf qualitative hypotheses. *Statistics and Probability Letters 63*(4), 425–433.

Fan, Q., Y.-C. Hsu, R. P. Lieli, and Y. Zhang (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics 40*(1), 313–327.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 189*(1), 1–23.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep neural networks for estimation and inference. *Econometrica 89*(1), 181–213.

Flores, C. A. (2007). Estimation of dose-response functions and optimal doses with a continuous treatment. *Working Paper*.

Flores, C. A., A. Flores-Lagunes, A. Gonzalez, and T. C. Neumann (2012). Estimating the effects of length of exposure to instruction in a training program: The case of Job Corps. *The Review of Economics and Statistics 94*(1), 153–171.

Fong, C., C. Hazlett, and K. Imai (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics 12*(1), 156–177.

Galvao, A. F. and L. Wang (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association 110*, 1528–1542.

Ghosal, S., A. Senand, and A. W. van der Vaart (2000). Testing monotonicity of regression. *The Annals of Statistics 28*(4), 1054–1082.

Gijbels, I., P. Hall, M. C. Jones, and I. Koch (2000). Tests for monotonicity of a regression mean with guaranteed level. *Biometrika 87*(3), 663–673.

Hall, P. and N. E. Heckman (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *The Annals of Statistics 28*(1), 20–39.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics 23*(4), 365–380.

Hirano, K. and G. W. Imbens (2004). The propensity score with continuous treatments. In A. Gelman and X. Meng (Eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Chapter 7, pp. 73–84. New York: Wiley.

Hsu, Y.-C. (2017). Consistent tests for conditional treatment effects. *Econometrics Journal 20*(1), 1–22.

Hsu, Y.-C., T.-C. Lai, and R. P. Lieli (2020). Estimation and inference for distribution and quantile functions in endogenous treatment effect models. *Econometric Reviews*, forthcoming.

Hsu, Y.-C., C.-A. Liu, and X. Shi (2019). Testing generalized regression monotonicity. *Econometric Theory 35*(6), 1146–1200.

Hsu, Y.-C. and S. Shen (2020). Testing monotonicity of conditional treatment effects under regression discontinuity designs. *Journal of Applied Econometrics*, Forthcoming.

Huber, M., Y.-C. Hsu, Y.-Y. Lee, and L. Lettry (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics 35*(7), 814–840.

Ichimura, H. and W. K. Newey (2022). The influence function of semiparametric estimators. *Quantitative Economics 13*(1), 29–61.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–710.

Imbens, G. W. and W. K. Newey (2009, 09). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512.

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

Kluve, J., H. Schneider, A. Uhlendorff, and Z. Zhao (2012). Evaluating continuous training programs using the generalized propensity score. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*(2), 587–617.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer: New York.

Lee, Y.-Y. (2018). Partial mean processes with generated regressors: Continuous treatment effects and nonseparable models. arxiv:1811.00157.

Linton, O., K. Song, and Y.-J. Whang (2010). An improved bootstrap test of stochastic dominance. *Journal of Econometrics 154*(2), 186–202.

Luo, Y. and M. Spindler (2016). High-dimensional l2boosting: Rate of convergence. arxiv:1602.08927.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Conference Board of the Mathematical Science: NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics.

Rothe, C. and S. Firpo (2019). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. *Econometric Theory 35*(5), 1048–1087.

Rothenhäusler, D. and B. Yu (2019). Incremental causal effects. rxiv:1907.13258.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics 48*(4), 1875–1897.

Schochet, P. Z., J. Burghardt, and S. Glazerman (2001). National Job Corps study: The impacts of Job Corps on participants' employment and related outcomes. *Report (Washington, DC: Mathematica Policy Research, Inc.)*.

Schochet, P. Z., J. Burghardt, and S. McConnell (2008). Does Job Corps work? impact findings from the national Job Corps study. *The American Economic Review 98*, 1864–1886.

Su, L., T. Ura, and Y. Zhang (2019). Non-separable models with high-dimensional data. *Journal of Econometrics 212*(2), 646–677.

Syrgkanis, V. and M. Zampetakis (2020). Estimation and inference with trees and forests in high dimensions. arxiv:2007.03210.

Wang, J. C. and M. C. Meyer (2011). Testing the monotonicity or convexity of a function using regression splines. *The Canadian Journal of Statistics 39*(1), 89–107.

Zhang, L. Z. (2022). Cross-validated conditional density estimation and continuous difference-in-differences models. Working paper, Department of Economics, UCLA.