

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: August 4, 2022

[†] 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 the 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 a weak unconfoundedness assumption. In a nonseparable structural model, applying our method amounts to testing monotonicity of the average structural function in the continuous treatment 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 the Monte-Carlo simulations. As an empirical illustration, we apply our test to the Job Corps study 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 also considers the assessment of continuously distributed treatments, e.g. hours spent in a training program whose effect on labor market performance is of interest. 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 a weak unconfoundedness assumption, implying that 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.

To construct our test, we first transform the null hypothesis of a monotonic relationship to countably many moment inequalities based on the generalized instrumental func-

tion approach of Hsu et al. (2019) and Hsu and Shen (2020). 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 parametric 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 marginal effects, which would converge at slower nonparametric rates. To compute the critical value 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 estimators do not affect the first-order large sample distribution of the DML estimators. Specifically, we give the high-level conditions on the mean-squared convergence rates on the first-step estimators, as for the semiparametric models in Chernozhukov et al. (2018). The nuisance 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 strength and weakness depending on the data generating process and applications, it is desired 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.

We investigate the finite sample behavior of the proposed test approach in a simulation study and also extend our method to testing conditional monotonicity given observed covariates. As an empirical illustration, we apply our test to data from an experimental study on 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 3 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 with larger treatment values, we neither reject 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 in 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 propose 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 the Job Corps data. Section 6 adapts the method to testing monotonicity with conditional (rather than unconditional) mean potential outcomes given observed covariates. Section

7 concludes. The technical proofs are relegated to the Appendix. An online supplement contains monotonicity tests under nonparametric and parametric estimations of generalized treatment propensity score.

2 Monotonicity of Continuous Treatment Effect

Let $Y(t)$ denote the potential outcome corresponding to the level of treatment intensity $t \in \mathcal{T}$, where $\mathcal{T} = [a, b]$ with $-\infty < a < b < \infty$. $Y(t)$ is called the unit-level dose-response function in Hirano and Imbens (2004). Let $\mu(t) = E[Y(t)]$ for $t \in \mathcal{T}$ denote the average of the potential outcome function, also known as the average dose-response function or the average structural function. In this paper, we are interested in testing if the average dose-response function is weakly increasing in the treatment intensity within a specific range. We define the null hypothesis of our interest as

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

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 $[t_\ell, t_u] = [0, 1]$.¹

Note that 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. 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.² To be specific, 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

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

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

²The generalized instrumental function approach is a generalization of the instrumental function approach in Andrews and Shi (2013, 2014).

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

for any $q = 2, \dots$, and for any $t_1 \geq t_2$ such that $q \cdot t_1, q \cdot t_2 \in \{0, 1, 2, \dots, q-1\}$. 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, by Hsu et al. (2019), Equations (2.2) and (2.3) contain the same information as the null hypothesis.

In the following, we discuss the identification of $\int_t^{t+q^{-1}} \mu(s) \cdot h(s) ds$.

Assumption 2.1. (Weak Unconfoundedness): $Y(t) \perp T \mid X$ for all $t \in \mathcal{T}$.

Assumption 2.1 is a commonly invoked identifying assumption based on observational data, also known as conditional independence and selection on observables. It assumes that conditional on observables X , T is as good as randomly assigned, or conditionally exogenous. The observed outcome Y satisfies that $Y = Y(T)$. We then have the following lemma concerning the identification of $\int_t^{t+q^{-1}} \mu(s) \cdot h(s) ds$. Let $p(t, x) = f_{T|X}(t|x)$ be the generalized propensity score, which is the conditional density of the treatment given the covariates and $p(t, x) > 0$ for all t and x .

Lemma 2.1. *Suppose Assumption 2.1 holds. Let $h(t) > 0$ for all t be a known weight function such that $\int_0^1 h(t) dt < \infty$. Then for $r > 0$,*

$$\int_t^{t+r} \mu(s) h(s) ds = E \left[\frac{Y}{p(T, X)} \cdot h(T) \cdot 1(T \in [t, t+r]) \right].$$

We now apply Lemma 2.1 of Hsu and Shen (2020) and the identification result in Lemma 2.1 to transform H_0 in (2.1) to countably many moment inequalities based on which we will construct our test. For $\ell = (t_1, t_2, q^{-1}) \in [0, 1]^2 \times (0, 1]$, define

$$\mathcal{L} = \left\{ \ell = (t_1, t_2, q^{-1}) : q \cdot (t_1, t_2) \in \{0, 1, 2, \dots, q-1\}^2, t_1 > t_2, \text{ and } q = 2, 3, \dots \right\}. \quad (2.4)$$

For each ℓ , we define

$$\nu_1(\ell) = E \left[\frac{Y}{p(T, X)} 1(T \in [t_1, t_1 + q^{-1}]) \right], \quad \nu_2(\ell) = E \left[\frac{Y}{p(T, X)} 1(T \in [t_2, t_2 + q^{-1}]) \right].$$

Lemma 2.2. *Suppose Assumption 2.1 holds. Assume that $\mu(t)$ is continuous in t . Then H_0 in (2.1) is equivalent to*

$$H'_0 : \nu(\ell) = \nu_2(\ell) - \nu_1(\ell) \leq 0 \text{ for any } \ell = (t_1, t_2, q^{-1}) \in \mathcal{L}. \quad (2.5)$$

The proof of Lemma 2.2 is a direct implication of (2.3) and Lemma 2.1. To see this, set $h(t) = 1$ and note that $\int_t^{t+r} h(s)ds = r$. By (2.3) and Lemma 2.1, for any $\ell \in \mathcal{L}$,

$$\begin{aligned} & \int_{t_2}^{t_2+q^{-1}} \mu(s) \cdot h(s)ds \cdot \int_{t_1}^{t_1+q^{-1}} h(s)ds - \int_{t_1}^{t_1+q^{-1}} \mu(s) \cdot h(s)ds \cdot \int_{t_2}^{t_2+q^{-1}} h(s)ds \leq 0 \\ & \text{iff } \nu_2(\ell)r - \nu_1(\ell)r \leq 0 \\ & \text{iff } \nu(\ell) = \nu_2(\ell) - \nu_1(\ell) \leq 0. \end{aligned}$$

In Lemma 2.2, we pick $h(t) = 1$ for simplicity, but the result also holds for any other known valid wight function $h(t)$.

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 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. (2022), Chernozhukov et al. (2018), and the nonparametric models for continuous treatments in Colangelo and Lee (2022).

We construct the moment function for our DML estimator by the Gateaux derivative limit. Denote as $\nu(t, r) = \int_t^{t+r} \mu(s)ds$ and $\gamma(t, x) = E[Y|T = t, X = x]$. Let f^0 be

³Our 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. (2022), Chernozhukov et al. (2018), Rothe and Firpo (2019), and references therein.

the true pdf of $Z = (Y, T, X)$ and f_Z^h be a pdf approaching a point mass at Z as $h \rightarrow 0$. 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 $\nu(t, r)$ is a linear functional of $\mu(t)$, the Gateaux derivative limit of $\nu(t, r)$ is

$$\begin{aligned} & \lim_{h \rightarrow 0} \int_t^{t+r} \left\{ \gamma(s, X) - \mu(s) + \frac{Y - \gamma(s, X)}{p(s, X)} f_T^h(s) \right\} ds \\ &= E \left[\frac{Y \mathbf{1}(T \in [t, t+r])}{p(T, X)} \middle| X \right] - \nu(t, r) + \frac{Y - \gamma(T, X)}{p(T, X)} \mathbf{1}(T \in [t, t+r]), \end{aligned} \quad (3.1)$$

and it follows that

$$\nu(t, r) = E \left[E \left[\frac{Y \mathbf{1}(T \in [t, t+r])}{p(T, X)} \middle| X \right] + \frac{Y - \gamma(T, X)}{p(T, X)} \mathbf{1}(T \in [t, t+r]) \right]. \quad (3.2)$$

We propose a DML estimator for $\nu(\ell)$ based on (3.2):

Step 1. (Cross-fitting) For some fixed $K \in \{2, \dots, n\}$, a K -fold cross-fitting partitions the observation indices into K distinct groups I_k , $k = 1, \dots, K$, 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 $k = 1, \dots, K$. For $k \in \{1, \dots, K\}$, the estimators $\hat{\gamma}_k(t, x)$ and $\hat{p}_k(t, x)$ for $\gamma(t, x)$ and $p(t, x)$ use observations not in I_k and satisfy Assumption 3.1 below.

Step 2. (Double robustness) The DML estimator is defined as

$$\begin{aligned} \hat{\nu}_{DML}(\ell) &= \hat{\nu}_{2,DML}(\ell) - \hat{\nu}_{1,DML}(\ell) \text{ where for } j = 1 \text{ and } 2, \\ \hat{\nu}_{j,DML}(\ell) &= \frac{1}{K} \sum_{k=1}^K \frac{1}{n_k} \sum_{i \in I_k} \left\{ \int_{t_j}^{t_j+q^{-1}} \hat{\gamma}_k(s, X_i) ds + \frac{Y_i - \hat{\gamma}_k(T_i, X_i)}{\hat{p}_k(T_i, X_i)} \mathbf{1}(T_i \in [t_j, t_j + q^{-1}]) \right\} \end{aligned}$$

and $\int_{t_j}^{t_j+q^{-1}} \hat{\gamma}_k(s, X_i) ds$ is approximated by a numerical integration $M^{-1} \sum_{m=1}^M \hat{\gamma}_k(s_m, X_i) \mathbf{1}(s_m \in [t_j, t_j + q^{-1}])$, with a set of equally spaced grid points $\{s_0 = t_\ell, s_1, \dots, s_M = t_u\}$ over $[t_\ell, t_u]$.

We use $\|\cdot\|_2$ to denote the L_2 -norm, e.g. $\|\hat{\gamma}_k - \gamma\|_2 = \left(\int_{\mathcal{X}} \int_{\mathcal{T}} (\hat{\gamma}_k(t, x) - \gamma(t, x))^2 f_{TX}(t, x) dt dx \right)^{1/2}$ and $\|\hat{p}_k - p\|_2 = \left(\int_{\mathcal{X}} \int_{\mathcal{T}} (\hat{p}_k(t, x) - p(t, x))^2 f_{TX}(t, x) dt dx \right)^{1/2}$.

Assumption 3.1 (DML). For any $k \in \{1, \dots, K\}$,

(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) The total variation of $\hat{\gamma}_k$ is finite with probability approaching one.

(iv) $p(T, X)$ is bounded away from zero and $\text{var}(Y|T, X)$ is bounded above almost surely.

Assumptions 3.1(i) and (ii) are the typical conditions on the mean-squared convergence rates, as in Chernozhukov et al. (2018). Assumption 3.1(iii) is to control the approximation error of the numerical integration.

Lemma 3.1 (DML). Let Assumptions 2.1 and 3.1 hold. Let $\sqrt{n}/M \rightarrow 0$. Then uniformly over $\ell \in \mathcal{L}$,

$$\begin{aligned} \sqrt{n}(\hat{\nu}_{DML}(\ell) - \nu(\ell)) &= n^{-1/2} \sum_{i=1}^n \phi_{\ell, DML}(Y_i, T_i, X_i) + o_p(1) \text{ where} \\ \phi_{\ell, DML}(Y, T, X) &= E \left[\frac{Y \mathbf{1}(T \in [t_2, t_2 + q^{-1}])}{p(T, X)} \middle| X \right] + \frac{Y - \gamma(T, X)}{p(T, X)} \mathbf{1}(T \in [t_2, t_2 + q^{-1}]) \\ &\quad - E \left[\frac{Y \mathbf{1}(T \in [t_1, t_1 + q^{-1}])}{p(T, X)} \middle| X \right] - \frac{Y - \gamma(T, X)}{p(T, X)} \mathbf{1}(T \in [t_1, t_1 + q^{-1}]) - \nu(\ell). \end{aligned} \tag{3.3}$$

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

Lemma 3.1 establishes the limiting behavior of DML estimators for ν 's. Let $\hat{\sigma}_{\nu, DML}^2(\ell) = K^{-1} \sum_{k=1}^K n_k^{-1} \sum_{i \in I_k} \hat{\phi}_{\ell, DML}^2(Y_i, T_i, X_i)$ where

$$\begin{aligned} \hat{\phi}_{\ell, DML}(Y_i, T_i, X_i) &= \left\{ \int_{t_2}^{t_2+q^{-1}} \hat{\gamma}_k(s, X_i) ds + \frac{Y_i - \hat{\gamma}_k(T_i, X_i)}{\hat{p}_k(T_i, X_i)} \mathbf{1}(T_i \in [t_2, t_2 + q^{-1}]) \right\} \\ &\quad - \left\{ \int_{t_1}^{t_1+q^{-1}} \hat{\gamma}_k(s, X_i) ds + \frac{Y_i - \hat{\gamma}_k(T_i, X_i)}{\hat{p}_k(T_i, X_i)} \mathbf{1}(T_i \in [t_1, t_1 + q^{-1}]) \right\} - \hat{\nu}_{DML}(\ell) \end{aligned} \tag{3.4}$$

and $\hat{\sigma}_{\nu, DML}^2(\ell)$ will be a consistent estimator for the asymptotic variance of $\sqrt{n}(\hat{\nu}_{DML}(\ell) - \nu(\ell))$ under the assumptions Lemma 3.1. Let $\hat{\sigma}_{\nu, \epsilon, DML}(\ell) = \max\{\hat{\sigma}_{\nu, DML}(\ell), \epsilon \cdot \hat{\sigma}_{\nu, DML}(0, 1/2, 1/2)\}$,

by which we manually bound the variance estimator away from zero. To test the null hypothesis H'_0 , we make use of a Cramér-von Mises test statistic defined as

$$\widehat{T}_{DML} = \sum_{\ell \in \mathcal{L}} \max \left\{ \sqrt{n} \frac{\hat{\nu}_{DML}(\ell)}{\hat{\sigma}_{\nu, \epsilon, DML}(\ell)}, 0 \right\}^2 Q(\ell), \quad (3.5)$$

where Q is a weighting function such that $Q(\ell) > 0$ 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 first introduce a multiplier bootstrap method that can simulate a process that converges to the same limit as $\sqrt{n}(\hat{\nu}_{DML}(\ell) - \nu(\ell))$. Let $\{U_i : 1 \leq i \leq n\}$ be a sequence of i.i.d. random variables that satisfy Assumption 3.2. We construct the simulated process as

$$\widehat{\Phi}_{\nu, DML}^u(\ell) = \frac{1}{\sqrt{n}} \sum_{i=1}^n U_i \cdot \hat{\phi}_{\ell, DML}(Y_i, T_i, X_i), \quad (3.6)$$

where $\hat{\phi}_{\ell, np}(Y_i, T_i, X_i)$ is the estimated influence function defined in (3.4). Under specific regularity conditions, we can show that the simulated process weakly converges to a Gaussian process conditional on the sample path with probability approaching one and that this limiting Gaussian process corresponds to the limiting process of $\sqrt{n}(\hat{\nu}_{np}(\ell) - \nu(\ell))$.

We adopt the GMS method to construct the simulated critical value as

$$\begin{aligned} \hat{c}_{DML}^\eta(\alpha) &= \sup \left\{ q \mid P^u \left(\sum_{\ell \in \mathcal{L}} \max \left\{ \frac{\widehat{\Phi}_{\nu, DML}^u(\ell)}{\hat{\sigma}_{\nu, \epsilon, DML}(\ell)} + \hat{\psi}_{\nu, DML}(\ell), 0 \right\}^2 Q(\ell) \leq q \right) \leq 1 - \alpha + \eta \right\} + \eta, \\ \hat{\psi}_{\nu, DML}(\ell) &= -B_n \cdot 1 \left(\sqrt{n} \cdot \frac{\hat{\nu}_{DML}(\ell)}{\hat{\sigma}_{\nu, \epsilon, DML}(\ell)} < -a_n \right), \end{aligned}$$

in which a_n and B_n satisfy Assumption 3.3.⁴

The decision rule is then given by

$$\text{Reject } H'_0 \text{ if } \widehat{T}_{DML} > \hat{c}_{DML}^\eta(\alpha). \quad (3.7)$$

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 $\{(Y_i, X_i, T_i) : 1 \leq i \leq n\}$ such that $E[U_i] = 0$, $E[U_i^2] = 1$, and $E[|U_i|^{2+\delta}] < C$ for some $\delta > 0$ and $C > 0$.

⁴The 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).

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

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

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

(a) Under H_0 , $\lim_{n \rightarrow \infty} P(\widehat{T}_{DML} > \widehat{c}_{DML}^\eta(\alpha)) \leq \alpha$;

(b) Under H_1 , $\lim_{n \rightarrow \infty} P(\widehat{T}_{DML} > \widehat{c}_{DML}^\eta(\alpha)) = 1$.

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. Recently Farrell et al. (2021) provide $\|\widehat{\gamma}_k - \gamma\|_2$ of deep neural networks. Colangelo and Lee (2022) propose 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 using the deep neural networks in Farrell et al. (2021), i.e. $\|\widehat{p}_k - p\|_2$.⁵ So Assumptions 3.1(i) and (ii) are attainable by the deep neural networks in Farrell et al. (2021) and Colangelo and Lee (2022). In the rest of this section, we provide the sufficient low-level conditions for lasso methods.

3.1 Step 1 lasso

We illustrate how to employ lasso methods to estimate the nuisance conditional mean function $\gamma(t, x)$ and the generalized propensity score $f_{T|X}(t|x)$. We provide sufficient conditions to verify the high-level Assumption 3.1. We modify the penalized local least squares estimator of $\gamma(t, X)$ in Su, Ura, and Zhang (2019) (SUZ, hereafter). We 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.

⁵The convergence rates in Lemmas 1 and 2 of Colangelo and Lee (2022) can be shown to hold uniformly over $t \in \mathcal{T}$, so we can obtain $\|\widehat{p}_k - p\|_2$. The additional assumption for the MultiGPS estimator in Lemma 2 Colangelo and Lee (2022) is $\sup_{t \in \mathcal{T}} \left\| \widehat{\mu} \left(h_1^{dt} g_{h_1}(T-t); X \right) - \mathbb{E}[h_1^{dt} g_{h_1}(T-t)|X] \right\|_{F_X} = O_p(R_{1n})$. Then Lemma 3 in Colangelo and Lee (2022) provides the sufficient conditions.

Let $b(T, X)$ be a $p \times 1$ vector of basis functions. We approximate $\gamma(t, x)$ by $b(t, x)' \theta$. The lasso estimator $\hat{\gamma}_k(t, x) = b(t, x)' \hat{\theta}_k$ for $k \in \{1, \dots, K\}$, where

$$\hat{\theta}_k = \arg \min_{\theta} \frac{1}{2(n - n_k)} \sum_{i \notin I_k} (Y_i - b(T_i, X_i)' \theta)^2 + \frac{\lambda}{n - n_k} \|\hat{\Xi}_k \theta\|_1, \quad (3.8)$$

where $n_k = \sum_{i=1}^n \mathbf{1}\{i \in I_k\}$, $\|\cdot\|_1$ denotes the L_1 norm, $\lambda = \ell_n(\log(p \vee n)n)^{1/2}$ for some slowly diverging sequence ℓ_n , and $\hat{\Xi}_k = \text{diag}(\tilde{l}_{k1}, \dots, \tilde{l}_{kp})$ is a generic penalty loading matrix computed by Algorithm 1 below from the iterative Algorithm 3.1 in SUZ. Denote as $\|f(X)\|_{\mathbb{P}_{n_k,2}} = ((n - n_k)^{-1} \sum_{i \notin I_k} f(X_i)^2)^{-1/2}$ for a generic function $f(\cdot)$.

Algorithm 1 (SUZ Algorithm 3.1). For $k \in \{1, \dots, K\}$,

1. Let $\hat{\Xi}_k^0 = \text{diag}(\tilde{l}_{k1}^0, \dots, \tilde{l}_{kp}^0)$, where $\tilde{l}_{kj}^0 = \|Y b_j(T, X)\|_{\mathbb{P}_{n_k,2}}$. Compute $\hat{\theta}_k^0$ by (3.8) with $\hat{\Xi}_k^0$ in place of $\hat{\Xi}_k$. Let $\hat{\gamma}_k^0(T_i, X_i) = b(T_i, X_i)' \hat{\theta}_k^0$.
2. For $s = 1, \dots, S$ for some fixed positive integer S , compute $\hat{\Xi}_k^s = \text{diag}(\tilde{l}_{k1}^s, \dots, \tilde{l}_{kp}^s)$, where $\tilde{l}_{kj}^s = \|(Y - \hat{\gamma}_k^{s-1}(T, X)) b_j(T, X)\|_{\mathbb{P}_{n_k,2}}$. Compute $\hat{\theta}_k^s$ by (3.8) with $\hat{\Xi}_k^s$ in place of $\hat{\Xi}_k$. and $\hat{\gamma}_k^s(T_i, X_i) = b(T_i, X_i)' \hat{\theta}_k^s$.

Let the final penalty loading matrix $\hat{\Xi}_k = \hat{\Xi}_k^S$ from Algorithm 1. Then the lasso estimator $\hat{\gamma}_k(t, x) = b(t, x)' \hat{\theta}_k$ for $k \in \{1, \dots, K\}$ from (3.8).

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 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 Λ is the logistic CDF. For $k \in \{1, \dots, K\}$, $\hat{F}_{T|X_k}(t|x) = \Lambda(b(X)' \hat{\beta}_{tk})$, where

$$\hat{\beta}_{tk} = \arg \min_{\beta} \frac{1}{n - n_k} \sum_{i \notin I_k} M(\mathbf{1}\{T_i \leq t\}, X_i; \beta) + \frac{\tilde{\lambda}}{n - n_k} \|\hat{\Psi}_{tk} \beta\|_1 \quad (3.9)$$

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

Algorithm 2 (SUZ Algorithm 3.2). For $k \in \{1, \dots, K\}$,

1. Let $\hat{\Psi}_{tk}^0 = \text{diag}(l_{tk,1}^0, \dots, l_{tk,p}^0)$, where $l_{tk,j}^0 = \|\mathbf{1}\{T \leq t\} b_j(X)\|_{\mathbb{P}_{nk,2}}$. Compute $\hat{\beta}_{tk}^0$ by (3.9) with $\hat{\Psi}_{tk}^0$ in place of $\hat{\Psi}_{tk}$ and $\hat{F}_{T|X_k}^0(t|x) = \Lambda(b(x)'\hat{\beta}_{tk}^0)$.
2. For $s = 1, \dots, S$, compute $\hat{\Psi}_{tk}^s = \text{diag}(l_{tk,1}^s, \dots, l_{tk,p}^s)$, where $l_{tk,j}^s = \left\| \left(\mathbf{1}\{T \leq t\} - \hat{F}_{T|X_k}^{s-1}(t|X) \right) b_j(X) \right\|_{\mathbb{P}_{nk,2}}$. Compute $\hat{\beta}_{tk}^s$ by (3.9) with $\hat{\Psi}_{tk}^s$ in place of $\hat{\Psi}_{tk}$ and $\hat{F}_{T|X_k}^s(t, x) = \Lambda(b(x)'\hat{\beta}_{tk}^s)$.

Let the final penalty loading matrix $\hat{\Psi}_{tk} = \hat{\Psi}_{tk}^S$ from Algorithm 2. Compute $\hat{F}_{T|X_k}(t|x) = \Lambda(b(X)'\hat{\beta}_{tk})$ from (3.9). 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. Following SUZ's notations, denote as $\|\cdot\|_{\mathcal{Q},q}$ the L^q norm under measure \mathcal{Q} and \mathbb{P} assigns probability $1/n$ to each observation.

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

1. (a) $\|\max_{j \leq p} |b_j(T, X)|\|_{\mathbb{P},\infty} \leq \zeta_n$ and $\underline{C} \leq \mathbb{E} [b_j(T, X)^2] \leq 1/\underline{C}$, for some positive constant \underline{C} , $j = 1, \dots, p$.
- (b) $\sup_{t \in \mathcal{T}} \max(\|\beta_t\|_0, \|\theta\|_0) \leq s$ for some s which possibly depends on n , where $\|\theta\|_0$ denotes the number of nonzero coordinates of θ .
- (c) For the approximation error, $\sup_{t \in \mathcal{T}} \|F_{T|X}(t|X) - \Lambda(b(X)'\beta_t)\|_{\mathbb{P},\infty} = O_p((s^2 \zeta_n^2 \log(p \vee n)/n)^{1/2})$ and $\|\gamma(T, X) - b(T, X)'\theta\|_{\mathbb{P},\infty} = o_p\left(\left((s^2 \zeta_n^2 \log(p \vee n)/n)^{1/2}\right)\right)$.
- (d) $p(t, x)$ is second-order differentiable w.r.t. t with bounded derivatives uniformly over $(t, x) \in \mathcal{T} \times \mathcal{X}$.
- (e) $\zeta_n^2 s^2 \ell_n^2 \log(p \vee n)/(nh_1) \rightarrow 0$, $nh_1^5/(\log(p \vee n)) \rightarrow 0$.
2. (a) There exists some positive constant $\underline{C} < 1$ such that $\underline{C} \leq p(t, x) \leq 1/\underline{C}$ uniformly over $(t, x) \in \mathcal{T} \times \mathcal{X}$.
- (b) $\gamma(t, x)$ is three times differentiable with all three derivatives being bounded uniformly over $(t, x) \in \mathcal{T} \times \mathcal{X}$.

3. There exists a sequence $\ell_n \rightarrow \infty$ such that, with probability approaching one, $0 < \kappa' \leq \inf_{\delta \neq 0, \|\delta\|_0 \leq s\ell_n} \frac{\|b(T, X)' \delta\|_{\mathbb{P}_{n,2}}}{\|\delta\|_2} \leq \sup_{\delta \neq 0, \|\delta\|_0 \leq s\ell_n} \frac{\|b(T, X)' \delta\|_{\mathbb{P}_{n,2}}}{\|\delta\|_2} \leq \kappa'' < \infty$.

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 $A_n = \ell_n (\log(p \vee n) s^2 \zeta_n^2 / n)^{1/2}$ and $\sup_{(t,x) \in \mathcal{T} \times \mathcal{X}} |\hat{p}_k(t, x) - p(t, x)| = O_p(B_n)$, where $B_n = h_1^{-1} (\log(p \vee n) s^2 \zeta_n^2 / n)^{1/2}$. 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 $A_n \rightarrow 0$ and $B_n \rightarrow 0$. And a sufficient condition of Assumption 3.4(ii) is $\sqrt{n} A_n B_n \rightarrow 0$.

4 Simulation

This section provides a simulation study to examine the finite sample performance of the proposed 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.

1. Instrumental functions: We opt for using a set of indicator functions of countable hypercubes. For $\ell = (t_1, t_2, q^{-1}) \in [0, 1]^2 \times (0, 1]$, define

$$\mathcal{L} = \left\{ \ell = (t_1, t_2, q^{-1}) : q \cdot (t_1, t_2) \in \{0, 1, 2, \dots, q-1\}^2, \right. \\ \left. t_1 > t_2, \text{ and } q = 2, \dots, q_1 \right\}, \quad (4.1)$$

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 last element of ℓ equal to q^{-1} . Recall that for each q , there are $(q(q+1)/2)$ instrumental functions with the last element of ℓ equal to q^{-1} .
3. a_n, B_n, ϵ, η : 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.

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

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

where the (i, j) -entry $\Sigma_{ij} = (0.5)^{|i-j|}$ for $i, j = 1, \dots, 100$, $U_t \sim \mathcal{N}(0, 1)$, 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 . Three 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: $Y = X'\beta T + \sin(\pi T) + X'\beta + U_y$.

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, $\mu(t) = \sin(\pi T)$, and H_0 does not hold. In this case, we expect the power will increase with the sample size.

In these DGPs, $1 + d_x = 101$. We consider samples of sizes $n = 200, 400, 800$, and 1600 . For q_1 , we set $q_1 = 4$ for $n = 200$, $q_1 = 8$ for $n = 400$, $q_1 = 16$ for $n = 800$, and $q_1 = 32$ for $n = 1600$. The number of subsamples used for cross-fitting is $K \in \{2, 5, 10\}$. 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). Also, all lasso estimations include an intercept and the covariates. For numerical integration in Step 2, we set $M = \lceil n^{2/3} \rceil$, where $\lceil \cdot \rceil$ 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%.^{6,7}

Table 1: Rejection probabilities of our test for $N = 50$

DGP	n	$\beta_j = 1/j^2$			$\beta_j = 1/j$		
		K=2	K=5	K=10	K=2	K=5	K=10
1	200	0.121	0.107	0.117	0.118	0.116	0.119
1	400	0.093	0.099	0.100	0.098	0.112	0.118
1	800	0.102	0.120	0.111	0.109	0.114	0.094
1	1600	0.109	0.097	0.110	0.095	0.106	0.086
2	200	0.001	0.002	0.000	0.003	0.000	0.000
2	400	0.000	0.000	0.000	0.000	0.002	0.000
2	800	0.000	0.000	0.000	0.000	0.000	0.000
2	1600	0.000	0.000	0.000	0.000	0.000	0.000
3	200	0.182	0.207	0.229	0.066	0.088	0.092
3	400	0.423	0.517	0.495	0.143	0.222	0.213
3	800	0.870	0.906	0.911	0.454	0.539	0.550
3	1600	0.999	1.000	1.000	0.898	0.921	0.937

Table 1 shows the rejection probabilities of our test for DGPs 1-3, and the results are consistent with our theoretical findings. For the mild dependence case, the proposed test controls size well in DGP 1 and DGP 2, and the rejection probabilities increase with

⁶In 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 IPW estimator whose variance is not affected by extremely low scores.

⁷Based on this trimming rule, around 0.5% of the samples are trimmed.

the sample size and are greater than the nominal size 0.1 in DGP 3. For the strong dependence case, our test still control size will in both DGP 1 and DGP 2. The power increases with the sample size in DGP 3, but the rejection probabilities are a bit less than the nominal size 0.1 for $n = 200$. Overall, we do not find significant difference for different choices of K .

Table 2: Rejection probabilities of our test for $K = 5$ and different N

DGP	n	$\beta_j = 1/j^2$				$\beta_j = 1/j$			
		N=33	N=40	N=50	N=66	N=33	N=40	N=50	N=66
1	200	0.133	0.100	0.107	0.113	0.101	0.121	0.116	0.111
1	400	0.118	0.116	0.099	0.099	0.127	0.114	0.112	0.110
1	800	0.136	0.103	0.120	0.091	0.122	0.108	0.114	0.109
1	1600	0.114	0.101	0.097	0.115	0.100	0.127	0.106	0.130
2	200	0.003	0.000	0.002	0.000	0.002	0.000	0.000	0.001
2	400	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.000
2	800	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	1600	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	200	0.210	0.247	0.207	0.170	0.100	0.072	0.088	0.084
3	400	0.521	0.503	0.517	0.465	0.220	0.221	0.222	0.202
3	800	0.911	0.912	0.906	0.895	0.549	0.554	0.539	0.547
3	1600	0.999	0.999	1.000	1.000	0.933	0.932	0.921	0.933

We next investigate the robustness of the performance of our test to the choice of q_1 . Let N denote the expected sample size of the smallest cube. We consider three alternative choices of q_1 , each resulting in $N = 33, 40$, and 66 , respectively. Table 2 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 .

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. Relatedly, Lee (2018) and Colangelo and Lee (2022) assess the effect of hours in 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 either study point to a concave association with the treatment dose.⁸

⁸See 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.

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, 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).⁹ 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), 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 in 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, such that there are only 11,313 individuals with completed follow-up inter-

⁹A 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.

views 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, for which we account by the inclusion of 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 3 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 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.¹⁰

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

Table 4 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 $T \in [40, 3000]$.

¹⁰In our empirical study, we do not get unstable IPW $\nu(\ell)$ estimates, so we decide 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 3: Descriptives 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	2.142	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
mum's years of education	11.504	2.599	0.000	20.000	3397
dad'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 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

Table 4: Test statistic and p-value, $40 \leq T \leq 3000$

	N=40, $t_1 > t_2$				N=50, $t_1 > t_2$			
	$\mu(t_1) \geq \mu(t_2)$		$\mu(t_1) \leq \mu(t_2)$		$\mu(t_1) \geq \mu(t_2)$		$\mu(t_1) \leq \mu(t_2)$	
$H_0 :$	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: Test statistic and p-value, $40 \leq T \leq 1000$

	N=40, $t_1 > t_2$				N=50, $t_1 > t_2$			
	$\mu(t_1) \geq \mu(t_2)$		$\mu(t_1) \leq \mu(t_2)$		$\mu(t_1) \geq \mu(t_2)$		$\mu(t_1) \leq \mu(t_2)$	
$H_0 :$	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.

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, we in a second step 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: Test statistic and p-value, $1000 \leq T \leq 2000$

	N=40, $t_1 > t_2$				N=50, $t_1 > t_2$			
	$\mu(t_1) \geq \mu(t_2)$		$\mu(t_1) \leq \mu(t_2)$		$\mu(t_1) \geq \mu(t_2)$		$\mu(t_1) \leq \mu(t_2)$	
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.

Table 5 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 monotonicity. For the intermediate treatment range of $[1000, 2000]$ considered in Table 6, 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 7: Test statistic and p-value, $2000 \leq T \leq 3000$

	N=40, $t_1 > t_2$				N=50, $t_1 > t_2$			
	$\mu(t_1) \geq \mu(t_2)$		$\mu(t_1) \leq \mu(t_2)$		$\mu(t_1) \geq \mu(t_2)$		$\mu(t_1) \leq \mu(t_2)$	
$H_0 :$	stat	p-value	stat	p-value	stat	p-value	stat	p-value
earn4	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 ‘earn4’, ‘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 (rather than unconditional) mean potential outcomes given observed covariates X . In this case, the null hypothesis considered corresponds to

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

where $\mu(t, x) = E[Y(t)|X = x]$ is the conditional average of the potential outcome function or the average dose-response function. For simplicity and without loss of generality, we henceforth assume that X is a scalar with $\mathcal{X} = [0, 1]$. By Lemma 2.1 of Hsu and Shen (2020), H_0 in (6.1) is equivalent to

$$\int_x^{x+q^{-1}} \int_{t_2}^{t_2+q^{-1}} \mu(s, \tilde{x}) \cdot h(s, \tilde{x}) ds d\tilde{x} \cdot \int_x^{x+q^{-1}} \int_{t_1}^{t_1+q^{-1}} h(s, \tilde{x}) ds d\tilde{x} - \int_x^{x+q^{-1}} \int_{t_1}^{t_1+q^{-1}} \mu(s, \tilde{x}) \cdot h(s, \tilde{x}) ds d\tilde{x} \cdot \int_x^{x+q^{-1}} \int_{t_2}^{t_2+q^{-1}} h(s, \tilde{x}) ds d\tilde{x} \leq 0 \quad (6.2)$$

for any $q = 2, \dots$, and for any $t_1 \geq t_2$ such that $q \cdot t_1, q \cdot t_2, q \cdot x \in \{0, 1, 2, \dots, q-1\}$. Define $h(t, x) = f(x)$ to be the density function of X . Following Lemma 2.1 and (3.2), we have for $r > 0$,

$$\begin{aligned} \int_x^{x+r} \int_t^{t+r} \mu(s, \tilde{x}) h(s, \tilde{x}) ds d\tilde{x} &= E \left[\frac{Y}{p(T, X)} \cdot \mathbf{1}(T \in [t, t+r]) \cdot \mathbf{1}(X \in [x, x+r]) \right] \\ &= E \left[\left\{ E \left[\frac{Y \mathbf{1}(T \in [t, t+r])}{p(T, X)} \middle| X \right] + \frac{Y - \gamma(T, X)}{p(T, X)} \mathbf{1}(T \in [t, t+r]) \right\} \mathbf{1}(X \in [x, x+r]) \right], \\ \int_x^{x+r} \int_t^{t+r} h(s, \tilde{x}) ds d\tilde{x} &= E [\mathbf{1}(X \in [x, x+r])]. \end{aligned}$$

For $\ell_x = (t_1, t_2, x, q^{-1}) \in [0, 1]^3 \times (0, 1]$, we let

$$\begin{aligned} \mathcal{L}_x = \left\{ \ell_x = (t_1, t_2, x, q^{-1}) : q \cdot (t_1, t_2, x) \in \{0, 1, 2, \dots, q-1\}^3, t_1 > t_2 \right. \\ \left. , \text{ and } q = 2, 3, \dots \right\}. \end{aligned} \quad (6.3)$$

Similar to (3.2), for each ℓ_x , we define

$$\begin{aligned} \nu_1(\ell_x) &= E \left[\left\{ E \left[\frac{Y \mathbf{1}(T \in [t_1, t_1 + q^{-1}])}{p(T, X)} \middle| X \right] + \frac{Y - \gamma(T, X)}{p(T, X)} \mathbf{1}(T \in [t_1, t_1 + q^{-1}]) \right\} \mathbf{1}(X \in [x, x + q^{-1}]) \right], \\ \nu_2(\ell_x) &= E \left[\left\{ E \left[\frac{Y \mathbf{1}(T \in [t_2, t_1 + q^{-1}])}{p(T, X)} \middle| X \right] + \frac{Y - \gamma(T, X)}{p(T, X)} \mathbf{1}(T \in [t_1, t_1 + q^{-1}]) \right\} \mathbf{1}(X \in [x, x + q^{-1}]) \right]. \end{aligned}$$

This permits establishing the following lemma.

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

$$H'_0 : \nu(\ell_x) = \nu_2(\ell_x) - \nu_1(\ell_x) \leq 0 \text{ for any } \ell_x = (t_1, t_2, x, q^{-1}) \in \mathcal{L}_x. \quad (6.4)$$

Similar to Section 3, we estimate $\nu(\ell_x)$ with $\ell_x = (t_1, t_2, x, q^{-1})$ as the following:

Step 1. (Cross-fitting) For some fixed $K \in \{2, \dots, n\}$, a K -fold cross-fitting partitions the observation indices into K distinct groups I_k , $k = 1, \dots, K$, such that the sample size of each group is the largest integer smaller than n/K . For $k \in \{1, \dots, K\}$, the estimators $\hat{\gamma}_k(t, x)$ and $\hat{p}_k(t, x)$ use observations not in I_k and satisfy Assumption 3.1 below.

Step 2. (Double robustness) The DML estimator is defined as

$$\hat{\nu}_{DML}(\ell_x) = \hat{\nu}_{2, DML}(\ell_x) - \hat{\nu}_{1, DML}(\ell_x), \text{ where for } j = 1 \text{ and } 2,$$

$$\begin{aligned}
& \hat{\nu}_{j,DML}(\ell_x) \\
&= \frac{1}{K} \sum_{k=1}^K \frac{1}{n_k} \sum_{i \in I_k} \left\{ \int_{t_j}^{t_j+q^{-1}} \hat{\gamma}_k(s, X_i) ds \right. \\
&\quad \left. + \frac{Y_i - \hat{\gamma}_k(T_i, X_i)}{\hat{p}_k(T_i, X_i)} \mathbf{1}(T_i \in [t_j, t_j + q^{-1}]) \right\} \mathbf{1}(X_i \in [x, x + q^{-1}]),
\end{aligned}$$

and $\int_{t_j}^{t_j+q^{-1}} \hat{\gamma}_k(s, X_i) ds$ is approximated as in Section 3.

Similar to Lemma 3.1, we can show that uniformly over $\ell_x \in \mathcal{L}_x$,

$$\sqrt{n}(\hat{\nu}_{DML}(\ell_x) - \nu(\ell_x)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_{\ell_x, DML}(Y_i, T_i, X_i) + o_p(1), \quad (6.5)$$

where

$$\begin{aligned}
& \phi_{\ell_x, DML}(Y_i, T_i, X_i) = \phi_{2, \ell_x, DML}(Y_i, T_i, X_i) - \phi_{1, \ell_x, DML}(Y_i, T_i, X_i), \text{ and for } j = 1 \text{ and } 2, \\
& \phi_{j, \ell_x, DML}(Y, T, X) \\
&= \mathbf{1}(X \in [x, x + q^{-1}]) \left(E \left[\frac{Y \mathbf{1}(T \in [t_j, t_j + q^{-1}])}{p(T, X)} \middle| X \right] + \frac{Y - \gamma(T, X)}{p(T, X)} \mathbf{1}(T \in [t_j, t_j + q^{-1}]) \right) - \nu_j(\ell_x).
\end{aligned}$$

Let $\hat{\phi}_{\ell_x, DML}(Y, T, X)$ be the estimated influence function similar to (3.4) and let $\hat{\sigma}_{\nu, DML}^2(\ell_x) = K^{-1} \sum_{k=1}^K n_k^{-1} \sum_{i \in I_k} \hat{\phi}_{\ell_x, DML}^2(Y_i, T_i, X_i)$ which will be a consistent estimator for the asymptotic variance of $\sqrt{n}(\hat{\nu}_{DML}(\ell_x) - \nu(\ell_x))$ under proper regularity conditions. Furthermore, let $\hat{\sigma}_{\nu, \epsilon, DML}(\ell_x) = \max\{\hat{\sigma}_{\nu, DML}(\ell_x), \epsilon \cdot \hat{\sigma}_{\nu, DML}(0, 1/2, 0, 1/2)\}$. The Cramér-von Mises test statistic is defined as

$$\hat{T}_{x, DML} = \sum_{\ell_x \in \mathcal{L}_x} \max \left\{ \sqrt{n} \frac{\hat{\nu}_{DML}(\ell_x)}{\hat{\sigma}_{\nu, \epsilon, DML}(\tau, \ell_x)}, 0 \right\}^2 Q(\ell_x), \quad (6.6)$$

where Q is a weighting function such that $Q(\ell_x) > 0$ for all $\ell_x \in \mathcal{L}_x$ and $\sum_{\ell_x \in \mathcal{L}_x} Q(\ell_x) < \infty$. The simulated process is constructed as

$$\hat{\Phi}_{\nu, x, DML}^u(\ell_x) = \frac{1}{\sqrt{n}} \sum_{i=1}^n U_i \cdot \hat{\phi}_{\ell_x, DML}(Y_i, T_i, X_i). \quad (6.7)$$

The GMS simulated critical value is given by

$$\begin{aligned}
& \hat{c}_{x, DML}^\eta(\alpha) = \sup \left\{ q \middle| P^u \left(\sum_{\ell_x \in \mathcal{L}_x} \max \left\{ \frac{\hat{\Phi}_{\nu, x, DML}^u(\ell_x)}{\hat{\sigma}_{\nu, \epsilon, DML}(\ell_x)} + \hat{\psi}_{\nu, DML}(\ell_x), 0 \right\} Q(\ell_x) \leq q \right) \leq 1 - \alpha + \eta \right\} + \eta, \\
& \hat{\psi}_{\nu, DML}(\ell_x) = -B_n \cdot \mathbf{1} \left(\sqrt{n} \cdot \frac{\hat{\nu}_{DML}(\ell_x)}{\hat{\sigma}_{\nu, \epsilon, DML}(\ell_x)} < -a_n \right).
\end{aligned}$$

Finally, the decision rule is given by

$$\text{Reject } H_0^l \text{ if } \widehat{T}_{x,DML} > \widehat{c}_{x,DML}^\eta(\alpha).$$

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

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 it to perform 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 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 Proof of Lemma 2.1

Under Assumption 2.1, Hirano and Imbens (2004) show that

$$\mu(t) = E[E[Y|T = t, p(t, X)]] = \int_{\mathcal{X}} E[Y|T = t, p(T, X) = p(t, X)] f(x) dx.$$

Then

$$\begin{aligned} \int_t^{t+r} \mu(t) h(s) ds &= \int_t^{t+r} \int_{\mathcal{X}} E[Y|T = s, p(T, X) = p(s, X)] f(x) h(s) dx ds \\ &= E \left[E[Y|T, p(T, X)] \frac{f(X) h(T) \mathbf{1}(T \in [t, t+r])}{f_{TX}(T, X)} \right] \\ &= E \left[E[Y|T, p(T, X)] \frac{h(T) \mathbf{1}(T \in [t, t+r])}{p(T, X)} \right] \\ &= E \left[\frac{Y}{p(T, X)} h(T) \mathbf{1}(T \in [t, t+r]) \right]. \end{aligned}$$

□

B Appendix for Section 3

Proof of Lemma 3.1:

We give an outline of deriving the asymptotically linear representation, following Chernozhukov et al.

(2022). Let $\nu(t, r) = \int_t^{t+r} \mu(s) ds$ and

$$\hat{\nu}_{DML}(t, r) = \frac{1}{n} \sum_{i=1}^n \left\{ \int_t^{t+r} \hat{\gamma}_k(s, X_i) ds + \frac{Y_i - \hat{\gamma}_k(T_i, X_i)}{\hat{p}_k(T_i, X_i)} \mathbf{1}(T_i \in [t, t+r]) \right\},$$

To show Lemma 3.1, it is sufficient to show that uniformly over $(t, r) \in [0, 1]^2$,

$$\begin{aligned} &\sqrt{n}(\hat{\nu}_{DML}(t, r) - \nu(t, r)) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n E \left[\frac{Y \mathbf{1}(T \in [t, t+r])}{p(T, X)} \middle| X \right] + \frac{Y - \gamma(T, X)}{p(T, X)} \mathbf{1}(T \in [t, t+r]) - \nu(t, r) + o_p(1). \end{aligned} \tag{B.1}$$

For notational ease, let $Z_i = (Y_i, X_i, T_i)$, $\gamma_i \equiv \gamma(T_i, X_i)$ and $\lambda_i \equiv \lambda(T_i, X_i) = 1/f_{T|X}(T_i|X_i)$. Let the doubly robust moment function in equation (3.1) be

$$\phi_{(t,r)}(Z_i, \gamma, \lambda)$$

$$\equiv E \left[Y \mathbf{1}(T \in [t, t+r]) \lambda(T, X) \middle| X = X_i \right] - \nu(t, r) + (Y_i - \gamma(T_i, X_i)) \lambda(T_i, X_i) \mathbf{1}(T_i \in [t, t+r]).$$

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

$$\begin{aligned} & \sqrt{n} \frac{1}{n} \sum_{i=1}^n \left\{ \hat{\phi}_{(t,r)}(Z_i, \hat{\gamma}, \hat{\lambda}) - \phi_{(t,r)}(Z_i, \gamma, \lambda) \right\} \\ &= \frac{1}{\sqrt{n}} \sum_{k=1}^K \sum_{i \in I_k} \left\{ \int_t^{t+r} (\hat{\gamma}_k(s, X_i) - \gamma(s, X_i)) ds - E \left[\int_t^{t+r} (\hat{\gamma}_k(s, X_i) - \gamma(s, X_i)) ds \middle| Z_k^c \right] \right\} \end{aligned} \quad (\text{R1-1})$$

$$+ \mathbf{1}(T_i \in [t, t+r]) \lambda_i(\gamma_i - \hat{\gamma}_{ik}) - E \left[\mathbf{1}(T_i \in [t, t+r]) \lambda_i(\gamma_i - \hat{\gamma}_{ik}) \middle| Z_k^c \right] \quad (\text{R1-2})$$

$$\left. + \mathbf{1}(T_i \in [t, t+r]) (\hat{\lambda}_{ik} - \lambda_i)(Y_i - \gamma_i) - E \left[\mathbf{1}(T_i \in [t, t+r]) (\hat{\lambda}_{ik} - \lambda_i)(Y_i - \gamma_i) \middle| Z_k^c \right] \right\} \quad (\text{R1-3})$$

$$\begin{aligned} & + \sqrt{n} \left\{ E \left[\int_t^{t+r} (\hat{\gamma}_k(s, X_i) - \gamma(s, X_i)) ds \middle| Z_k^c \right] - E \left[\mathbf{1}(T_i \in [t, t+r]) \lambda_i(\hat{\gamma}_{ik} - \gamma_i) \middle| Z_k^c \right] \right. \\ & \left. + E \left[(\hat{\lambda}_{ik} - \lambda_i) \mathbf{1}(T_i \in [t, t+r]) (Y_i - \gamma_i) \middle| Z_k^c \right] \right\} \end{aligned} \quad (\text{R1-DR})$$

$$- \frac{1}{\sqrt{n}} \sum_{k=1}^K \sum_{i \in I_k} \mathbf{1}(T_i \in [t, t+r]) (\hat{\lambda}_{ik} - \lambda_i) (\hat{\gamma}_{ik} - \gamma_i). \quad (\text{R2})$$

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).

Note that we can express

$$\int_t^{t+r} \gamma_k(s, X_i) ds = E \left[\frac{\gamma(T, X) \mathbf{1}(T \in [t, t+r])}{p(T, X)} \middle| X = X_i \right]. \quad (\text{B.2})$$

By the law of iterated expectations, $E \left[\int_t^{t+r} (\hat{\gamma}_k(s, X) - \gamma(s, X)) ds \middle| Z_k^c \right] = E \left[\lambda(T, X) (\hat{\gamma}_k(T, X) - \gamma(T, X)) \mathbf{1}(T \in [t, t+r]) \middle| Z_k^c \right]$. So (R1-DR) is zero.

The approximation error of the Riemann sum is

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

by Assumption 3.1(iii). By the condition $\sqrt{n}/M \rightarrow 0$, the approximation error is asymptotically ignorable.

To show (R1-1), (R1-2) and (R1-3) are $o_p(1)$ uniformly over ℓ , we show these terms weakly converge to Gaussian processes indexed by ℓ 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 ω in the probability space Ω 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 $F_{ni} = |\lambda_i(\hat{\gamma}_{ik} - \gamma_i)|/\sqrt{n}$. $F_{n_k} = (F_{n_1}, \dots, F_{n_{n_k}})'$ is a R^{n_k} -valued function on the underlying probability space.

(ii) Let $X_n(t) = X_n(\omega, t) = \sum_{i \in I_k} (f_{ni}(t) - E[f_{ni}(t)|Z_k^c])$. 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 (iv). Let $s \leq t \in [0, 1]$, without loss of generality. $H(s, t) = \lim_{n \rightarrow \infty} E[X_n(s)X_n(t)|Z_k^c] = \lim_{n \rightarrow \infty} E[\mathbf{1}(T \in (s, t])\lambda_i^2(\hat{\gamma}_{ik} - \gamma_i)^2|Z_k^c] = 0$.

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

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

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

(v) For any $s < t$,

$$\begin{aligned} \rho_n(s, t) &= \left(\sum_{i \in I_k} E[|f_{ni}(s) - f_{ni}(t)|^2 | Z_k^c] \right)^{1/2} = \left(E[\mathbf{1}(T_i \in (s, t])\lambda_i(\hat{\gamma}_{ik} - \gamma_i)^2 | Z_k^c] \right)^{1/2} \\ &= O_p(\|\hat{\gamma}_k - \gamma\|_2) = o_p(1) \end{aligned}$$

and the last equality holds by Assumption 3.1(i). Hence, $\rho(s, t) = \lim_{n \rightarrow \infty} \rho_n(s, t) = 0$. The condition (v) holds: for all deterministic sequences $\{s_n\}$ and $\{t_n\}$, if $\rho(s_n, t_n) \rightarrow 0$ then $\rho_n(s_n, t_n) \rightarrow 0$.

Then Theorem 10.6 in Pollard (1990) implies that the finite dimensional distributions of \mathbf{X}_n have Gaussian limits, with zero means and covariances given by H . Therefore, $\mathbf{X}_n = o_p(1)$ 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), $E[f_{ni}(t)^2 | Z_k^c] = O_p(\|\hat{\lambda}_k - \lambda\|_2^2) = o_p(1)$ by the smoothness condition and Assumption 3.1(i).

For (R1-1), define $f_i(t) = \int_0^t (\hat{\gamma}_k(s, X_i) - \gamma(s, X_i)) ds$. By (B.2),

$$\begin{aligned} E[f_i(t)^2 | Z_k^c] &\leq \int \left(E \left[\frac{(\hat{\gamma}_k(T, X) - \gamma(T, X)) \mathbf{1}(T \leq t)}{p(T, X)} \middle| X = X_i \right] \right)^2 f_X(X_i) dX_i \\ &\leq \int E \left[\left(\frac{\hat{\gamma}_k(T, X) - \gamma(T, X)}{p(T, X)} \mathbf{1}(T \leq t) \right)^2 \middle| X = X_i \right] f_X(X_i) dX_i \\ &= \int \int \left(\frac{\hat{\gamma}_k(T_i, X_i) - \gamma(T_i, X_i)}{p(T_i, X_i)} \right)^2 \mathbf{1}(T_i \leq t) f_{T|X}(T_i | X_i) dT_i dX_i \\ &= O_p \left(\int \int (\hat{\gamma}_k(T_i, X_i) - \gamma(T_i, X_i))^2 f_{T|X}(T_i | X_i) dT_i dX_i \right) \\ &= o_p(1) \end{aligned}$$

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

For (R2),

$$\begin{aligned} &E \left[\sup_{\ell} \left| n^{-1/2} \sum_{i \in I_k} \mathbf{1}(T_i \in [t, t+r]) (\hat{\lambda}_{ik} - \lambda_i) (\gamma_i - \hat{\gamma}_{ik}) \right| \middle| Z_k^c \right] \\ &\leq \sqrt{n} \int_{\mathcal{X}} \int_{\mathcal{T}} \sup_{\ell} \mathbf{1}(T_i \in [t, t+r]) \left| (\hat{\lambda}_{ik} - \lambda_i) (\gamma_i - \hat{\gamma}_{ik}) \right| f_{TX}(T_i, X_i) dT_i dX_i \\ &\leq \sqrt{n} \left(\int_{\mathcal{X}} \int_{\mathcal{T}} (\hat{\lambda}_{ik} - \lambda_i)^2 f_{TX}(T_i, X_i) dT_i dX_i \right)^{1/2} \left(\int_{\mathcal{X}} \int_{\mathcal{T}} (\hat{\gamma}_{ik} - \gamma_i)^2 f_{TX}(T_i, X_i) dT_i dX_i \right)^{1/2} \\ &\xrightarrow{p} 0 \tag{B.3} \end{aligned}$$

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

By the triangle inequality, we obtain the asymptotically linear representation

$$n^{-1/2} \sum_{i=1}^n (\hat{\phi}_{t,r}(Z_i, \hat{\gamma}, \hat{\lambda}) - \phi_{t,r}(Z_i, \gamma, \lambda)) = o_p(1),$$

and (B.1) follows.

Then by the fact that $\nu(\ell) = \nu(t_1, q^{-1}) - \nu(t_2, q^{-1})$, then it follows that uniformly over $\ell \in \mathcal{L}$,

$$\sqrt{n}(\hat{\nu}_{DML}(\ell) - \nu(\ell)) = n^{-1/2} \sum_{i=1}^n \phi_{\ell, DML}(Y_i, T_i, X_i) + o_p(1),$$

and this shows the first half of Lemma 3.1.

For the second part, similar to Hsu et al. (2019), it is straightforward to see that $\{\phi_{\ell, DML}(Y, T, X) : \ell \in \mathcal{L}\}$ is a VC class of functions and by functional central limit theorem of Pollard (1990), it follows that $\sqrt{n}(\hat{\nu}_{DML}(\cdot) - \nu(\cdot)) \Rightarrow \Phi_{h_{DML}}(\cdot)$ where $\Phi_{h_{DML}}(\cdot)$ is a Gaussian process with variance-covariance kernel $h_{DML}(\ell_1, \ell_2) = E[\phi_{\ell_1, DML}(Y, T, X)\phi_{\ell_2, DML}(Y, T, X)]$. This completes the proof of Lemma 3.1. \square

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

Proof of Lemma B.1:

The fact that $\{\phi_{\ell, DML} : \ell \in \mathcal{L}\}$ is a VC type class of functions implies that $\{\phi_{\ell, np}^2 : \ell \in \mathcal{L}\}$ is also a VC type. In addition, given that $E[\bar{\phi}_{np}^{2+\delta}] < \infty$, we have by the uniform weak law of large numbers that $\sup_{\ell \in \mathcal{L}} |\tilde{\sigma}_{\nu, DML}^2(\ell) - \sigma_{\nu, DML}^2(\ell)| \xrightarrow{P} 0$, where $\tilde{\sigma}_{\nu, DML}^2(\ell) = n^{-1} \sum_{i=1}^n \phi_{\ell, DML}^2(Y_i, T_i, X_i)$. By Assumption 3.1, we have that $\sup_{\ell \in \mathcal{L}} |\tilde{\sigma}_{\nu, DML}^2(\ell) - \hat{\sigma}_{\nu, DML}^2(\ell)| \xrightarrow{P} 0$. 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. \square

Proof of Theorem 3.1:

The proof of Theorem 3.1 follows from the same arguments as Theorem 5.1 of Hsu (2017) once Lemmas 3.1 and B.1 are established and is omitted for the sake of brevity. \square

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, 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.
- 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.
- 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 (2022). Locally robust semiparametric estimation. *Econometrica* 90(4), 1501–1535.
- Chetverikov, D. (2019). Testing regression monotonicity in econometric models. *Econometric Theory* 35(4), 1146–1200.

- 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 regression qualitative hypotheses. *Statistics and Probability Letters* 63(4), 425–433.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.