The Annals of Statistics
2024, Vol. 52, No. 4, 1592–1615
https://doi.org/10.1214/24-AOS2405

A NONPARAMETRIC DOUBLY ROBUST TEST FOR A CONTINUOUS TREATMENT EFFECT

BY CHARLES R. DOSS[1,a], GUANGWEI WENG[1,b], LAN WANG[2,c], IRA MOSCOVICE[3,d] AND TONGTAN CHANTARAT[3,e]

[1]*School of Statistics, University of Minnesota,* [a]*cdoss@stat.umn.edu,* [b]*wengx076@umn.edu*
[2]*Miami Herbert Business School, University of Miami,* [c]*lxw611@miami.edu*
[3]*School of Public Health, University of Minnesota,* [d]*mosco001@umn.edu,* [e]*chant083@umn.edu*

The vast majority of literature on evaluating the significance of a treatment effect based on observational data has been confined to discrete treatments. These methods are not applicable to drawing inference for a continuous treatment, which arises in many important applications. To adjust for confounders when evaluating a continuous treatment, existing inference methods often rely on discretizing the treatment or using (possibly misspecified) parametric models for the effect curve. Recently, Kennedy et al. (*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** (2017) 1229–1245) proposed nonparametric doubly robust estimation for a continuous treatment effect in observational studies. However, inference for the continuous treatment effect is a harder problem. To the best of our knowledge, a completely nonparametric doubly robust approach for inference in this setting is not yet available. We develop such a nonparametric doubly robust procedure in this paper for making inference on the continuous treatment effect curve. Using empirical process techniques for local U- and V-processes, we establish the test statistic's asymptotic distribution. Furthermore, we propose a wild bootstrap procedure for implementing the test in practice. In addition, we define a version of the test procedure based on sample splitting. We illustrate the new method(s) via simulations and a study of a constructed dataset relating the effect of nurse staffing hours on hospital performance. We implement our doubly robust dose response test in the R package DRDRtest on CRAN.

1. Introduction.

We are interested in hypothesis testing for a continuous (causal) treatment effect based on observational data. The fundamental challenge of causal inference with observational data is to account for confounding variables, which are variables related to both the outcome and the treatment. In the presence of confounding variables, it is well known that naive regression modeling does not lead to an unbiased estimate for the causal effect curve. While continuous treatments are common in many important applications, much of the existing literature on inference for a treatment effect from observational data has been focused on discrete treatments. Relatively few methods are available for testing hypotheses about a continuous treatment effect curve.

Under the popular "no unmeasured confounders" assumption, there are two broad directions to adjust for the confounding variables. A procedure can start by estimating the outcome regression function, a function that relates the outcome to the treatment and the confounders, and then this can be weighted appropriately to yield an estimate of the causal estimand. Alternatively, a procedure can start by estimating the propensity score function, a function that relates the treatment to confounders, and then allows a variety of methods to be implemented to estimate the causal estimand. For instance, Imbens (2004) and Hill (2011) model only the outcome regression function; while Galvao and Wang (2015), Hirano and Imbens

Received May 2023; revised March 2024.
MSC2020 subject classifications. Primary 62D20; secondary 62G10.
Key words and phrases. Causal inference, double robustness, nonparametric, hypothesis test, U-statistics.

(2004), Imai and van Dyk (2004) model only the propensity score function. Following the terminology of semiparametric statistics, the outcome regression function and the propensity score function are often referred to as nuisance parameters (possibly infinite-dimensional). In the aforementioned approaches, an incorrectly specified model for either nuisance parameter would lead to inconsistent estimates of the treatment effect curve; and, regardless, the conclusions are susceptible to the curse of dimensionality: the rate of convergence of the estimator of the treatment effect curve is the same as that of the estimator of the nuisance parameter, which may be high dimensional.

In the so-called doubly robust approach, widely used for estimating the average treatment effect in the discrete treatment setting, one estimates both nuisance parameters and then combines them. The term "doubly robust" means that only one of the two nuisance parameters needs to be estimated consistently to achieve consistent estimation of the causal treatment effect. Thus even if the model for one of the two nuisance parameters is misspecified, the causal estimand can still be estimated consistently if the other nuisance parameter model is correctly specified; alternatively/similarly, the rate of the leading error term in estimating the causal estimand is determined by the product of the error terms for estimating the two nuisance parameters, allowing for efficient estimation.

If one wishes to apply a doubly robust test in the continuous treatment setting, the simplest and likely the standard approach would be to discretize the treatment and use the methodology for discrete treatments (Robins et al. (2007), Van Der Laan and Dudoit (2003)). Unfortunately, this could result in misleading estimates, and can lead to possibly massive loss of power. Also, in many applications maintaining the treatment as a continuous variable is important for post-analysis interpretation. (As mentioned above, Galvao and Wang (2015) develop inference procedures for the dose response curve but require good, possibly parametric, estimators for the propensity score; see their assumptions N.1, G.IV.) If one wishes to use doubly robust methods without discretization, then Robins (2000) and Neugebauer and van der Laan (2007) allow this, but requires specifying a parametric model for the unknown causal treatment effect curve. If the parametric model is not plausible, then the results can be unreliable. Recently, a nonparametric doubly robust estimation method has been proposed (Kennedy et al. (2017)), allowing for greater flexibility in modeling the nuisance parameters. Although the rates of nonparametrically estimating each nuisance function may be slower than $\sqrt{n}$, the rate of estimating the causal estimand may be much faster than that for estimating either of the individual nuisance parameters (by virtue of the product rate discussed earlier), while alleviating the difficulty of model specification for nuisance parameters.

To the best of our knowledge, a completely nonparametric doubly robust approach for inference for a continuous treatment effect is not yet available. It is worth noting that "double robustness" for estimation does not automatically warrant "double robustness" for inference. See, for instance, related discussions in van der Laan (2014) and Benkeser et al. (2017).

In this paper, we develop a doubly robust procedure for testing the null hypothesis that the treatment effect curve is constant. To do so, we introduce a test statistic based on comparing the integrated squared distance from an estimate under the alternative to the null estimate. We derive the limit distribution of the proposed test statistic. In order to implement the hypothesis test, the unknown parameters in the limit distribution must be estimated. A natural approach is the bootstrap (Efron and Tibshirani (1993)). Unfortunately, the naive bootstrap turns out to be inconsistent. We propose a wild bootstrap procedure, which provides provable guarantees for estimating the limit distribution, and thus allows the test to be implemented. Code that implements our doubly robust dose response test is available in the R package "DRDRtest" on CRAN.

Our main contribution is thus a new doubly robust test procedure which is consistent (in level and against fixed alternatives) as long as at least one of the two nuisance parameters is specified correctly. It requires only nonparametric assumptions on the nuisance parameters unlike Robins (2000) and Neugebauer and van der Laan (2007). The proposed test is doubly robust in the sense that the p-values we generate are reliable (uniformly distributed under the null hypothesis) even if one of the nuisance parameter models is misspecified. Some may argue that one may use machine learning to estimate the nuisance parameters to alleviate model misspecification. Although this is true to some degree, popular machine learning methods such as random forests and neural networks are not immune from model misspecification; without structural assumptions (e.g., sparsity, additive structure) on the underlying model, they may have poor estimation accuracy (very slow rates of convergence). Their practical implementations also often require multiple tuning parameters.

Our statistic is inspired by the test of Härdle and Mammen (1993) (also Dette and Neumeyer (2001)) which were developed in the noncausal setting. Comparing with the noncausal setting, our theory is significantly more complicated due to the two infinite-dimensional nuisance parameters that are present in the causal inference setting. We develop technical results that may be of independent interest and of use in future work involving continuous treatment effect estimation/inference. In Section K of the Supplementary Material Doss et al. (2024), we present (nonasymptotic) moment inequalities for U-processes. We extend the work of Arcones and Giné (1993) to yield moment inequalities (with upper bounds given as a product of entropy-type integrals and envelope function moments) in the flavor of known bounds for empirical processes (like, e.g., Theorem 2.14.1 of van der Vaart and Wellner (1996)) but which are new for U-processes. Then in Lemmas C.1, C.2 and C.3, we apply the moment inequalities to certain local U-processes; these, or similar, local U-processes will arise naturally (see the proof outline after Theorem 3.1) in contexts related to performing estimation or inference for (nonparametrically estimated) continuous treatment effects, especially when one is using estimators related to local smoothing. These lemmas (or their proofs) that bound the size of these remainder terms are likely to be necessary or very useful tools in such contexts.

The rest of the paper is organized as follows. In Sections 1.1 and 1.2, immediately following this one, we provide further discussions on related literature on nonparametric hypothesis testing and on the continuous treatment effect setting, respectively. Section 2 introduces the setup, notation, the new testing procedure, and underlying assumptions. In Section 3, we present the main results. Section 4 presents simulation studies and in Section 5 we present analysis of a dataset relating nurse staffing to hospital effectiveness. Proofs and arguments that do not fit in the main paper are provided in the Supplementary Material (Doss et al. (2024)). References that begin with a capital letter (“A”, “B”, etc.) refer to the Supplementary Material.

1.1. *Literature on nonparametric hypothesis testing.*

The simpler, noncausal, problem of hypothesis testing about a (noncausal) regression function when the alternative is a large nonparametric class has a very large literature already. There are many different approaches to this general problem; to start, one must decide on the definition of the nonparametric alternative class. The full, unrestricted, nonparametric alternative class is generally tested against by using a test based on a primitive of the function of interest: if $m(\cdot)$ is the regression function then $M(a) := \int_{-\infty}^{a} m(x)\,dx$ is the primitive. This approach is possibly more familiar to readers in the density/distribution testing setting where m and M would be replaced by a density and cumulative distribution function, respectively. Such tests based on M are “omnibus” in the sense that in theory they have power approaching one against any fixed alternative. However, for that theory to be relevant with certain fixed alternatives, extremely large sample sizes may be needed; or put another way, there are many alternatives that such tests for practical sample sizes are not well powered.

Another set of procedures is based on taking the alternative class to be some sort of smoothness class (e.g., a Hölder, Sobolev, or Besov class; Giné and Nickl (2016)). Confusion may arise because some tests, for example, those based on primitive functions, may have power against local alternatives converging at rate $n^{-1/2}$ whereas tests based on smoothness assumptions often require local alternatives to converge at a slower rate. However, this is a case where the (local) rates of convergence can be misleading. Rather than local rates, one can use global minimax rates of convergence over a given class or classes to compare procedures. Ingster (1993a, 1993b, 1993c) studies minimax rates in nonparametric hypothesis testing problems (in a white noise model and in density estimation, both with simple null models). We do not recount all the results here, but note that in general tests based on primitives will not attain minimax optimal rates against smoothness-based alternatives (Ingster (1993a), Section 2.5, Pouet (2001)). The minimax results are not just theoretical: Eubank and LaRiccia (1992) provide both theory and simulation results demonstrating that (in the context of density estimation) for any fixed sample size there are alternative sequences such that smoothness-based methods are more powerful than primitive-based methods (the Cramér-von Mises statistic in this case) even though the latter has a local $n^{-1/2}$ rate. The statistic we develop here is built to have uniform power over a large nonparametric smoothness class, as well as to be doubly robust.

One of the difficulties in smoothness-based testing is the issue of bias. Nonparametric smoothness-based estimates generally have nontrivial bias which must be accounted for and the estimation of which entails complications. In our particular testing setting, actually there is no bias under the null (since our null hypothesis is the class of constant functions), so bias is not a major issue in the usual way. The bias will affect the estimator under the alternative and so will affect the power. A related issue is that in nonparametric testing, the asymptotic distributions of many test statistics have the property that their bias (mean) is of a larger order of magnitude than their variance. One consequence of this for us is that it makes the asymptotically negligible error terms in the analysis of our test statistic (in the causal setting of the current paper) much more complicated than they would otherwise be; it turns out that the large bias of the main term gets multiplied by other error terms (after expanding a square) and this requires extra mathematical analysis.

1.2. *Literature on continuous treatment effects.* In the last few years there has been significant and increasing interest in causal inference with continuous treatment effects; this includes interest in the setting of optimal treatment regimes (Chen, Li and Yu (2022), Chen, Zeng and Kosorok (2016), Kallus and Zhou (2018), Schulz and Moodie (2021)), and in specific scientific areas (e.g., Coulombe, Moodie and Platt (2021), Kreif et al. (2015) in the health sciences). These settings are not the same as ours, but do provide motivation for our interest in continuous treatments.

We briefly discuss here several papers that have built theory and/or methods related to causal effect estimation and/or inference in the presence of a continuous treatment (based on observational data). The recent literature on estimation starts with Kennedy et al. (2017), on which other works, including the present paper, build. Kennedy et al. (2017) have developed a method for efficient doubly robust estimation of the treatment effect curve. Denote the outcome regression function by μ or μ_0, and denote the propensity score function by π or π_0. Their method is based on a pseudo-outcome $\xi \equiv \xi(\boldsymbol{Z}; \pi, \mu)$, which depends on the sample point $\boldsymbol{Z}$, and on the nuisance functions π, μ. The pseudo-outcome ξ has the key double robustness property that if *either* $\pi = \pi_0$ or $\mu = \mu_0$, then $\mathbb{E}(\xi(\boldsymbol{Z}; \pi, \mu)|A = a)$ is equal to the treatment effect curve (at the treatment value a). The estimation procedure of Kennedy et al. (2017) is then a natural two-step procedure: (1) estimate the nuisance functions (π_0, μ_0) by some estimators $(\widehat{\pi}, \widehat{\mu})$ which the user can choose as they wish and construct (observable) pseudo-outcomes $\widehat{\xi}_i$ (which approximate ξ_i and depend on $\widehat{\pi}$, $\widehat{\mu}$), and (2) regress the pseudo-outcomes on A using some nonparametric method (e.g., local linear regression). As we described above, the error term from the nuisance parameter estimation is given by the product of the error term for estimating π_0 and for estimating μ_0, so is smaller than either, partially alleviating the curse of dimensionality.

Several works have now made use of the pseudo-outcome approach of Kennedy et al. (2017), or similar approaches. Semenova and Chernozhukov (2021), Westling, Gilbert and Carone (2020) use the pseudo-outcomes of Kennedy et al. (2017) with alternative estimation techniques, and Colangelo and Lee (2020), Su, Ura and Zhang (2019) use similar pseudo-outcomes (and study particular nuisance estimators). Like Kennedy et al. (2017), Westling, Gilbert and Carone (2020) also develop a doubly robust estimator of a continuous treatment effect curve; they develop a different procedure, based on the assumption that the true effect curve satisfies the shape constraint of monotonicity. Colangelo and Lee (2020) provide an alternative motivation for a related pseudo-outcome Kennedy et al. (2017), study a sample-splitting variation of the estimation methodology of Kennedy et al. (2017), and also consider estimating the gradient of the treatment curve.

Many works since Kennedy et al. (2017) have considered doubly robust estimation of structural/causal functions based on nonparametric models. Some include or focus on continuous treatment effects, while others focus more on the problem of conditional average treatment effect (CATE) (or "partially conditional average treatment effect" (PCATE) (Wang et al. (2022)) based on a binary treatment variable, or other related quantities. The (P)CATE setting is of course different than the continuous treatment setting we consider, but may share some features with our setting when the covariates on which the treatment is conditioned are continuous, so we discuss some of the recent literature briefly. Chernozhukov et al. (2018), Chernozhukov, Newey and Singh (2022), Semenova and Chernozhukov (2021) develop general "double/debiased" machine learning approaches to estimating causal estimands that are continuous functions. Chernozhukov, Newey and Singh (2022) develop a Dantzig-type estimator based on estimating equations for the nuisance parameters. They consider four running example estimands, as well as "local" and "perfectly localized" functionals, the latter including the continuous treatment effect at a fixed point. Semenova and Chernozhukov (2021) develop a general theory for debiased machine learning for different causal or missing data estimands, such as conditional average treatment effects, regression functions with partially missing outcomes, and conditional average partial derivatives. They also consider the causal effect curve with continuous treatments. However, in the latter setting, their assumptions are slightly too strong to allow double robustness for (pointwise) consistency.[1]

In summary, the theory and methodology of Chernozhukov, Newey and Singh (2022), Semenova and Chernozhukov (2021) is built for a variety of settings and does not focus exclusively on the setting of continuous treatments, which we do focus on, and derive a powerful procedure for, here.

Lee, Okui and Whang (2017), Luedtke and van der Laan (2016), Wang et al. (2022), Zimmert and Lechner (2019) and Foster and Syrgkanis (2023), consider various doubly robust types of estimation procedures for (P)CATE estimation, generally based on pseudo-outcomes. Here "PCATE" means the effect of a binary/discrete treatment conditional on some covariates which may be a strict subset of the set of all confounding variables. Nie and Wager (2021) and Kennedy (2023) consider a different type of estimator (dubbed the

[1] "The only requirement we impose on the estimation of [the nuisance parameters] is that [they converge] to the true nuisance parameter η_0 at a fast enough rate [$o_p(n^{-1/4-\delta})$] for some $\delta \geq 0$" (Semenova and Chernozhukov (2021), page 271); see their Assumption 4.9.

"R-learner" and "lp-R-Learner", respectively, after Robinson (1988)); they provide double-robust-type conditions under which these two-step estimators can attain the oracle rate of convergence, with Kennedy (2023) able to weaken the conditions on the nuisances given in Nie and Wager (2021) by a new cross-validation technique (inspired by Newey and Robins (2018)) and undersmoothing. Very recently Kennedy, Balakrishnan and Wasserman (2022) find minimax lower and upper bounds for CATE estimation (showing that the estimator/rates of Kennedy (2023) were optimal in some smoothness regimes but not others). We do not know of an analog of the (lp-)R-Learner that has been directly studied for the case of estimation of continuous treatments. An open question is whether the approach we develop here can be applied also in the setting of inference for a (P)CATE.

In terms of inference, none of the above works on (P)CATE estimation consider hypothesis tests or questions of inference except for Lee, Okui and Whang (2017). Inference results for the continuous treatment effect curve are limited. Kennedy et al. (2017) find the pointwise limit distribution for their estimator, but do not operationalize it. Semenova and Chernozhukov (2021), in their general approach for debiased machine learning, develop confidence bands, but when those bands are applied to the case of the continuous treatment effect curve, the band is centered at an approximation of the true function, rather than at the true function itself (i.e., there is an error term that is ignored; see their Theorem 4.7). In a different setup Luedtke, Carone and van der Laan (2019) develop a hypothesis test in a causal inference setting which could allow for continuous treatments. However, their null hypothesis is different than ours and their conditions (Condition 3) rule out our setting. They also consider the causal effect curve with continuous treatments. However, in the latter setting, their assumptions are slightly too strong to allow double robustness for (pointwise) consistency.[2] and their confidence band is centered at an approximation of the true function, rather than at the true function itself (i.e., there is an error term that is ignored; see their Theorem 4.7). Chernozhukov, Newey and Singh (2022) develop Gaussian approximations for the distributions of their estimators at a fixed point, but they do not further develop inference methods, so their focus is distinct from our focus on a (global) testing problem.[3] Thus, none of the above works solve the global inference problem that we address here.

Very recently, Westling (2022) considered a similar problem to the one we consider, but using a different test statistic, based on a "primitive" (or "antiderivative") of the treatment effect curve. The present paper and Westling (2022) were developed entirely separately. A strength of the primitive-based method of Westling (2022) is that it naturally handles mixed discrete–continuous exposures, whereas our method would require further modifications (e.g., locally chosen bandwidths) to do so. Westling (2022)'s method is not quite doubly robust (requiring an $o_p(n^{-1/2})$ rather than $O_p(n^{-1/2})$ product-nuisance-estimation-rate), whereas our method (which allows an $o_p((n\sqrt{h})^{-1/2})$ rate) is. (See Westling (2022)'s Assumption A4 and Theorems 3 and 4, as well as Figure 1.) Also, the two methods have noticeably different power in different scenarios. Westling (2022)'s test has a local $n^{-1/2}$ convergence rate, whereas our test statistic has a local convergence rate of $(n\sqrt{h})^{-1/2}$. The implications are as discussed above: Westling (2022)'s test will have power focused in "one direction" and decaying in directions away from that one direction, and our power will be more uniformly spread over the alternatives. For instance, we will have noticeably higher power when the true (alternative) effect curve has significant peaks or valleys.

[2] "The only requirement we impose on the estimation of [the nuisance parameters] is that [they converge] to the true nuisance parameter η_0 at a fast enough rate [$o_p(n^{-1/4-\delta})$] for some $\delta \geq 0$" (Semenova and Chernozhukov (2021), page 271); see their Assumption 4.9.

[3] In discussing works that focus on continuous treatment effect estimation, they say "These works develop inference on perfectly localized average potential outcomes with continuous treatment effects, using a different approach than what we develop here. Our development is complementary as it covers a much broader collection of functionals."

2. Setup and method.

2.1. *Notation, data and setup.* We will use the following notations throughout this paper. Let $(\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n)$ be the observed sample where each observation is an independent copy of the tuple $\boldsymbol{Z} = (\boldsymbol{L}, A, Y)$ with support $\mathcal{Z} = \mathcal{L} \times \mathcal{A} \times \mathcal{Y}$. Here $\mathcal{L} \subseteq \mathbb{R}^d$ and $\boldsymbol{L}$ is the vector containing the d potential confounder variables (covariates); $\mathcal{A} \subseteq \mathbb{R}$ and A is the continuous treatment dosage received; $\mathcal{Y} \subseteq \mathbb{R}$ and Y is the observable outcome of interest. We let P denote the distribution of $\boldsymbol{Z}$ and $p_0(\boldsymbol{z}) = p_0(y|\boldsymbol{l}, a)p_0(a|\boldsymbol{l})p_0(\boldsymbol{l})$ denote the corresponding density function with respect to some dominating measure ν. Let $\mu_0(\boldsymbol{l}, a) := \mathbb{E}(Y|\boldsymbol{L} = \boldsymbol{l}, A = a)$ denote the outcome regression function. Similarly, let $\pi_0(a|\boldsymbol{l}) := \frac{\partial}{\partial a} P(A \le a|\boldsymbol{L} = \boldsymbol{l})$ denote the conditional density or propensity score function of A given $\boldsymbol{L}$ and $\varpi_0(a) := \frac{\partial}{\partial a} P(A \le a)$ denote the marginal density function of A (both of which densities are assumed to exist). For a function f on $\mathbb{R}$, we let $\mathbb{P}\{f(\boldsymbol{Z})\} := \int_{\mathcal{Z}} f(\boldsymbol{z})\, dP(\boldsymbol{z})$. And for $p \ge 1$, we use $\|f\|_p := \{\int f(\boldsymbol{z})^p\, dP(\boldsymbol{z})\}^{1/p}$ to denote the $L_p(P)$ norm and use $\|f\|_{\mathcal{X}} := \sup_{x \in \mathcal{X}} |f(x)|$ to denote the uniform norm over the range $\mathcal{X}$. We use $\mathbb{P}_n$ to denote the empirical distribution defined on the observed data so that $\mathbb{P}_n\{f(\boldsymbol{Z})\} := \int f(\boldsymbol{z})\, d\mathbb{P}_n(\boldsymbol{z}) = n^{-1} \sum_{i=1}^n f(\boldsymbol{Z}_i)$.

To characterize the problem, let Y^a be the potential outcome (Rubin (1975)) when treatment level a is applied. Then the causal estimand that we are interested in learning about (developing a hypothesis test for) is $\theta_0(a) := \mathbb{E}(Y^a)$, and we wish to test if this function is constant. Specifically, we want to test

$$H_0 \colon \theta_0 \equiv c \in \mathbb{R} \quad \text{versus} \quad H_1 \colon \theta_0 \text{ is nonconstant}, \tag{2.1}$$

where we will assume that $\theta_0(\cdot)$ satisfies some smoothness assumptions if it is nonconstant.

2.2. *Proposed method.* In Kennedy et al. (2017), the authors derived a doubly robust mapping for estimating the continuous treatment effect curve. Like doubly robust estimators in binary treatment cases, the doubly robust mapping depends on both the outcome regression function and the propensity score function, and can be written as

$$\xi(\boldsymbol{Z}; \pi, \mu) = \frac{Y - \mu(\boldsymbol{L}, A)}{\pi(A|\boldsymbol{L})} \int_{\mathcal{L}} \pi(A|\boldsymbol{l})\, dP(\boldsymbol{l}) + \int_{\mathcal{L}} \mu(\boldsymbol{l}, A)\, dP(\boldsymbol{l}), \tag{2.2}$$

where $\pi(a|\boldsymbol{l})$ and $\mu(\boldsymbol{l}, a)$ are some propensity score and outcome regression functions, respectively. The above mapping has the desired property of double robustness in that

$$\mathbb{E}\{\xi(\boldsymbol{Z}; \pi, \mu)|A = a\} = \theta_0(a), \tag{2.3}$$

provided either $\mu = \mu_0$ or $\pi = \pi_0$, under Assumptions I below (Kennedy et al. (2017)). Thus $\theta_0(\cdot)$ could be estimated using standard nonparametric smoothing techniques if either μ_0 or π_0 were known. Since we do not actually know μ_0, π_0, we plug in estimators $\widehat{\mu}, \widehat{\pi}$ for μ_0, π_0. To compute ξ, we also need to know $dP(\boldsymbol{l})$ in two places; since we do not, we plug in $\mathbb{P}_n(\boldsymbol{l})$ for $P(\boldsymbol{l})$, and we denote this by $\widehat{\xi}$. Thus, our estimate of the pseudo-outcome $\xi(\boldsymbol{Z}; \pi_0, \mu_0)$ is

$$\widehat{\xi}(\boldsymbol{Z}; \widehat{\pi}, \widehat{\mu}) = \frac{Y - \widehat{\mu}(\boldsymbol{L}, A)}{\widehat{\pi}(A|\boldsymbol{L})} \int_{\mathcal{L}} \widehat{\pi}(A|\boldsymbol{l})\, d\mathbb{P}_n(\boldsymbol{l}) + \int_{\mathcal{L}} \widehat{\mu}(\boldsymbol{l}, A)\, d\mathbb{P}_n(\boldsymbol{l}). \tag{2.4}$$

We can subsequently apply a nonparametric estimation procedure to the (observed) tuples $\{(\widehat{\xi}(\boldsymbol{Z}_i; \widehat{\pi}, \widehat{\mu}), A_i)\}_{i=1}^n$. Kennedy et al. (2017) show that when at least one of the estimators is consistent then, under some assumptions about complexity and boundedness conditions of $\widehat{\mu}$ and $\widehat{\pi}$ and the product of their convergence rates, the convergence rate of the nonparametric estimator is the same as if we know the true μ_0 or π_0. Kennedy et al. (2017) apply a local linear estimator to the pseudo-outcome to estimate $\theta_0(\cdot)$ and show the above-stated property on the pointwise convergence rate of the nonparametric estimator.

In this paper, we are interested in a different problem: testing if $\theta_0(\cdot)$ is constant or not. As in the estimation problem, in the setting where we (unrealistically) know one of μ_0 or π_0, testing whether $\theta_0(\cdot)$ is constant becomes a standard regression problem, and we can consider many possible nonparametric tests to the tuples $\{(\widehat{\xi}(\boldsymbol{Z}_i;\widehat{\pi},\widehat{\mu}), A_i)\}_{i=1}^n$. As described in Section 1, not all tests will be doubly robust for testing, though.

In Härdle and Mammen (1993), the authors consider the problem of testing parametric null linear models (not in a causal setting) and construct test statistics based on the integrated difference between the nonparametric model estimated using the Nadaraya–Watson estimator and the parametric null model. Alcalá, Cristóbal and González-Manteiga (1999) extended the test to allow using a local polynomial estimator (Fan and Gijbels (1996)) for the nonparametric model. We propose to test our hypothesis (2.1) of a constant treatment effect curve (i.e., no treatment effect) using the following statistic

$$T_n = n\sqrt{h}\int_{\mathcal{A}}\big(\widehat{\theta}_h(a) - \mathbb{P}_n\widehat{\xi}(\boldsymbol{Z})\big)^2 w(a)\,da, \tag{2.5}$$

where $w(\cdot)$ is a user-specified weight function, $\widehat{\theta}_h(a)$ is the local linear estimator applied to $\{(\widehat{\xi}(\boldsymbol{Z}_i;\widehat{\pi},\widehat{\mu}), A_i)\}_{i=1}^n$. To define the local linear estimator, we let $\widehat{\beta}_h(a) = \operatorname{argmin}_{\beta\in\mathbb{R}^2}\mathbb{P}_n[K_{ha}(A)\{\widehat{\xi}(\boldsymbol{Z};\widehat{\pi},\widehat{\mu}) - g_{ha}(A)^T\beta\}^2]$, where $g_{ha}(t) = (1, \frac{t-a}{h})^T$, $K_{ha}(t) = h^{-1}K\{(t-a)/h\}$, and $K(\cdot)$ is a kernel function, and then we let $\widehat{\theta}_h(a) = g_{h,0}(0)^T\widehat{\beta}_h(a)$. Note: we could define the test statistic T_n by a summation over A_i rather than as an integral against (Lebesgue measure) da; as mentioned in Dette and Neumeyer (2001), Horowitz and Spokoiny (2001), similar results as ours would hold although some constants would change. Note that under the null hypothesis of no treatment effect, $\mathbb{P}_n\{\xi(\boldsymbol{Z};\pi_0,\mu_0)\}$ is an $\sqrt{n}$-consistent estimator of the null model and so is $\mathbb{P}_n\widehat{\xi}$ provided $\widehat{\mu}$, $\widehat{\pi}$ are not converging too slowly (see the later discussion). We will see under some mild conditions on the convergence rates of $\widehat{\pi}$ and $\widehat{\mu}$, that T_n converges to a normal distribution similar to the one in Alcalá, Cristóbal and González-Manteiga (1999) under the null model (given in (2.1)). However, similar to Härdle and Mammen (1993) and Alcalá, Cristóbal and González-Manteiga (1999), due to the slow order of convergence of the asymptotically negligible terms that arise in the proof, we do not suggest using the target distribution given in Theorem 3.1 to directly calculate the critical values under the null hypothesis. Instead we advocate using the bootstrap (Efron and Tibshirani (1993)) to estimate the distribution of T_n to improve the finite sample performance and, more specifically, we use the so-called wild bootstrap (Davidson and Flachaire (2008)) as used in Alcalá, Cristóbal and González-Manteiga (1999), Härdle and Mammen (1993) and Dette and Neumeyer (2001). In Härdle and Mammen (1993), the authors show the theoretical properties of three different bootstrap methods: (1) the naive resampling method; (2) the adjusted residual bootstrap; (3) the wild bootstrap and showed only the wild bootstrap gives consistent estimation of the null distribution. These results are again true in our setting: the wild bootstrap is valid whereas the other two are not. Here we provide a brief outline of our proposed test procedure.

1. Estimate (π_0,μ_0) by (black-box estimators) $(\widehat{\pi},\widehat{\mu})$.
2. Calculate the pseudo-outcomes $\widehat{\xi}(\boldsymbol{Z};\widehat{\pi},\widehat{\mu})$ by (2.4) and construct the local linear estimator $\widehat{\theta}_h(a)$ using $\{(\widehat{\xi}(\boldsymbol{Z}_i;\widehat{\pi},\widehat{\mu}), A_i)\}_{i=1}^n$.
3. To generate wild bootstrap samples to estimate the distribution of T_n under the null hypothesis,

   (a) Calculate the estimated residuals as $\hat{\varepsilon}_i = \widehat{\xi}(\boldsymbol{Z}_i;\widehat{\pi},\widehat{\mu}) - \widehat{\theta}_h(A_i)$ (we can also use $\hat{\varepsilon}_i = \widehat{\xi}(\boldsymbol{Z}_i;\widehat{\pi},\widehat{\mu}) - \sum_{i=1}^n\widehat{\xi}(\boldsymbol{Z}_i;\widehat{\pi},\widehat{\mu})/n$),

   (b) Do the following B times, where B is the desired number of bootstrap resamplings: for each $i\in\{1,\dots,n\}$, generate $\varepsilon_i^*\sim\widehat{F}_i$ (defined just below, based on $\{\hat{\varepsilon}_i\}$) and use $(\xi_i^* = \mathbb{P}_n\widehat{\xi} + \varepsilon_i^*, A_i)$ as bootstrap observations,

4. Use the wild bootstrap samples to compute $T^*_{n,j}$, $j = 1, \dots, B$ (according to (2.5) but using the bootstrap samples) and use $\{T^*_{n,j}\}_{j=1}^B$ to estimate the distribution of T_n under the null hypothesis. Let $\widehat{t}^*_{n,1-\alpha}$ denote the $1 - \alpha$ quantile of the estimated distribution, where $0 < \alpha < 1$ is the predetermined significance level. Reject the null hypothesis if $T_n > \widehat{t}^*_{n,1-\alpha}$.

When generating bootstrap samples, we use $\hat{F}_i$ to estimate the conditional distribution of $\xi(\boldsymbol{Z}_i; \bar{\pi}, \bar{\mu})$ based on the single residual $\hat{\varepsilon}_i$. Härdle and Mammen (1993) use a "two point distribution" which matches the first three moments of $\hat{\varepsilon}_i$ and is defined as

$$\varepsilon_i^* = \begin{cases} -\hat{\varepsilon}_i(\sqrt{5} - 1)/2 & \text{with probability } (\sqrt{5} + 1)/(2\sqrt{5}), \\ \hat{\varepsilon}_i(\sqrt{5} + 1)/2 & \text{with probability } (\sqrt{5} - 1)/(2\sqrt{5}). \end{cases} \tag{2.6}$$

We also consider another common choice, a Rademacher type distribution, where ε_i^* equals $\hat{\varepsilon}_i$ or $-\hat{\varepsilon}_i$ with probability $1/2$ each. (Davidson and Flachaire (2008)). Unlike the two point distribution, the Rademacher distribution matches the first two and the fourth (and all even) moments of $\hat{\varepsilon}_i$, but imposes symmetry on $\hat{F}_i$.

REMARK 2.1. In Section H of the Supplementary Material Doss et al. (2024), we also present an extension of the doubly robust pseudo-outcome to allow for possible (discrete or continuous) effect modifiers, and present the natural extension of the test to the case where the effect modifier is discrete.

2.3. *Assumptions*.

Here we introduce the assumptions needed for our theoretical results. Our parameter of interest, $\theta_0(a)$, is defined on the potential outcome Y^a which is not observable. Thus, we need the following identifiability conditions on the observed data.

ASSUMPTION I.

1. Consistency: $A = a$ implies $Y = Y^a$.
2. Positivity: $\pi_0(a|\boldsymbol{l}) \geq \pi_{\min} > 0$ for all $\boldsymbol{l} \in \mathcal{L}$ and all $a \in A$.
3. Ignorability: $\mathbb{E}(Y^a|\boldsymbol{L}, A) = \mathbb{E}(Y^a|\boldsymbol{L})$.

We need some further assumptions to regulate the distribution of the observed data and the treatment effect curve $\theta_0(a)$.

ASSUMPTION D.

1. The support of A (i.e., $\mathcal{A}$), is a compact subset of $\mathbb{R}$.

2. The treatment effect curve $\theta_0(a)$ and the marginal density function $\varpi_0(a)$ are twice continuously differentiable.

3. The conditional density $\pi_0(a|\boldsymbol{l})$ and the outcome regression function $\mu_0(\boldsymbol{l}, a)$ are uniformly bounded.

4. Let $\tau(\boldsymbol{l}, a) := \mathrm{Var}(Y|\boldsymbol{L} = \boldsymbol{l}, A = a)$ be the conditional variance of Y given covariates and treatment level. Assume there exist $\tau_{\max} > 0$ such that $0 < \tau(\boldsymbol{l}, a) \leq \tau_{\max}$ for all $\boldsymbol{l} \in \mathcal{L}$ and $a \in \mathcal{A}$. Moreover, define

$$S_\tau := \{\boldsymbol{l} \in \mathcal{L} : \tau(\boldsymbol{l}, a) \text{ is a continuous function of } a\},$$
$$S_{\pi_0} := \{\boldsymbol{l} \in \mathcal{L} : \pi_0(a|\boldsymbol{l}) \text{ is a continuous function of } a\},$$
$$S_{\mu_0} := \{\boldsymbol{l} \in \mathcal{L} : \mu_0(\boldsymbol{l}, a) \text{ is a continuous function of } a\};$$

assume we have $P(S_\tau \cup S_\pi \cup S_\mu) = 1$.

Statement 4 about the sets S_τ, S_{π_0}, S_μ is just a slight relaxation of the requirement that the given functions all be simultaneously almost surely continuous everywhere. For the assumption on our kernel, we also need to define a *Vapnik–Chervonenkis (VC)* (Dudley (1999)) class. If a class of functions $\mathcal{F}$ is a VC class, we have that

$$\sup_Q N(\tau\|F\|_{2,Q}, \mathcal{F}, L_2(Q)) \le \left(\frac{C}{\tau}\right)^v \tag{2.7}$$

for some positive C, v and all $\tau > 0$ (and again the sup is over all probability measures Q). The assumptions we make on our estimators are as follows.

ASSUMPTION E(A).

1. The bandwidth $h \equiv h_n$ fulfills $c_1^h n^{-1/5} \le \liminf h_n \le \limsup h_n \le c_2^h n^{-1/5}$ for some constants $0 < c_1^h \le c_2^h < \infty$.

2. Let $\bar{\pi}$ and $\bar{\mu}$ denote the limits of the estimators $\widehat{\pi}$ and $\widehat{\mu}$ such that $\|\widehat{\pi} - \bar{\pi}\|_{\mathcal{Z}} = o_p(\sqrt{h})$ and $\|\widehat{\mu} - \bar{\mu}\|_{\mathcal{Z}} = o_p(\sqrt{h})$, where h is the bandwidth used in local linear estimator. And we have either $\bar{\pi} = \pi_0$ or $\bar{\mu} = \mu_0$.

3. The kernel function K for the local linear estimator is a continuous symmetric probability density function with support on $[-1, 1]$. Moreover, we assume the class of functions $\{K((\cdot - a)/h) : a \in \mathbb{R}, h > 0\}$ satisfies condition (2.7).

4. Let r_n^∞ and s_n^∞ be such that

$$\sup_{a\in\mathcal{A}} \|\widehat{\pi}(a|\boldsymbol{L}) - \pi_0(a|\boldsymbol{L})\|_2 = O_p(r_n^\infty),$$

$$\sup_{a\in\mathcal{A}} \|\widehat{\mu}(\boldsymbol{L}, a) - \mu_0(\boldsymbol{L}, a)\|_2 = O_p(s_n^\infty).$$

We assume $s_n^\infty r_n^\infty = o\{(n\sqrt{h})^{-1/2}\}$.

Assumption E(A)1 requires that h is of the order of magnitude for optimal estimation; such h can be achieved a variety of ways (for instance, one can minimize a risk estimate, or perform cross validation (Fan and Gijbels (1996)). Assumption E(A)2 is not a stringent assumption; by definition $\overline{\pi}$, $\overline{\mu}$ are the limits of the estimators $\widehat{\pi}$, $\widehat{\mu}$; here we require the rate of convergence (in $\|\cdot\|_{\mathcal{Z}}$) to these limits (which are not necessarily the truth) to be order $\sqrt{h} = o(n^{-1/10})$ which is quite slow.

Assumption E(A)3 is a standard assumption on the user-chosen kernel. Assumption E(A)4 is a somewhat nonstandard assumption, since it combines L_∞ and L_2 norms. The L_2 aspect arises in the local asymptotics of Kennedy et al. (2017), and the L_∞ aspect arises because we consider a global test. Note that in parametric settings, r_n^∞ or s_n^∞ may attain $\sqrt{n}$ rates. For instance, in a linear regression, if the regression model is $\mu(\boldsymbol{l}, a) = (\boldsymbol{l}^T, a)\boldsymbol{\beta}$, and $\widehat{\boldsymbol{\beta}}$ is an estimator converging to the true parameter $\boldsymbol{\beta}_0$ at $\sqrt{n}$ rate, then $\mathbb{P}((\boldsymbol{l}^T, a)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))^2 \le 2O_p(n^{-1})(a^2 + \mathbb{P}\|\boldsymbol{l}\|_2^2)$, which follows from using the inequality $(a+b)^2 \le 2a^2 + 2b^2$ and the fact that if $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = O_p(n^{-1/2})$ then $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T$ has eigenvalues of order $O_p(n^{-1})$. Taking a square root and a supremum over the bounded set $a \in \mathcal{A}$ shows that in this case r_n^∞ is order $\sqrt{n}$. Similar results hold in other parametric models for μ or for π. Thus indeed, the test is “doubly robust”: if one parametric model is well-specified and attains root-n rates, the other may be misspecified.

Many nonparametric or semiparametric examples will also satisfy Assumption E(A)4; if r_n^∞ and s_n^∞ are both, say, order $n^{-2/5}$ up to polylogarithmic factors, which is the rate one expects from for instance, a generalized additive model under twice differentiability then the assumption is satisfied.

In addition to the above E(A) ("Estimator assumption part A") assumption, we make one more assumption ("Estimator assumption part B") on the estimators of the nuisance parameters and the local linear estimator of the treatment effect curve. We assume the estimators for nuisance parameters fall in classes with finite uniform entropy integrals. For a generic class of functions $\mathcal{F}$, let F denote an envelope function for $\mathcal{F}$, that is, $\sup_{f\in\mathcal{F}}|f| \le \mathcal{F}$. Let $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ denote the covering number, that is, the minimal number of ε-balls (with distance defined on $\|\cdot\|$) needed to cover $\mathcal{F}$. Let

$$J_m(\delta, \mathcal{F}, L_2) := \int_0^\delta \sup_Q \bigl(1 + \log N\bigl(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\bigr)\bigr)^{m/2}\, d\varepsilon, \tag{2.8}$$

where the sup is over all probability measures Q and $L_2(Q) \equiv \|\cdot\|_{2,Q}$ is the L_2 semimetric under the distribution Q, that is, $\|f\|_{2,Q} = (\int f^2\, dQ)^{1/2}$. If $J_1(1, \mathcal{F}, L_2) < \infty$ we say $\mathcal{F}$ has a finite uniform entropy integral, and following standard convention, we sometimes let $J(\cdot,\cdot,\cdot)$ refer to $J_1(\cdot,\cdot,\cdot)$. Differing results require $J_m(1, \mathcal{F}, L_2) < \infty$ for differing values of $m \in \{1, 2, 3, 4\}$. Thus, we subscript this next assumption by m, which corresponds to the requirement that $J_m(1, \mathcal{F}, L_2) < \infty$. We label the below assumption as "E(B)$_m$" and when we want to assume that, for example, $J_3(1, \mathcal{F}, L_2) < \infty$, we refer to the assumption as "E(B)$_3$".

ASSUMPTION E(B)$_m$. The estimators $\widehat{\pi}$, $\widehat{\mu}$ and their limits $\bar{\pi}$, $\bar{\mu}$ are contained in uniformly bounded function classes $\mathcal{F}_\pi$, $\mathcal{F}_\mu$, which satisfy that $J_m(1, \mathcal{F}, L_2) < \infty$ for $\mathcal{F} = \mathcal{F}_\pi$ or $\mathcal{F} = \mathcal{F}_\mu$, with $1/\widehat{\pi}$ also uniformly bounded. Moreover, we assume $P(S_{\bar{\pi}} \cup S_{\bar{\mu}}) = 1$, where we let

$$S_{\bar{\pi}} := \bigl\{\boldsymbol{l} \in \mathcal{L} : \bar{\pi}(a|\boldsymbol{l}) \text{ is a continuous function of } a\bigr\},$$
$$S_{\bar{\mu}} := \bigl\{\boldsymbol{l} \in \mathcal{L} : \bar{\mu}(\boldsymbol{l}, a) \text{ is a continuous function of } a\bigr\}.$$

REMARK 2.2. To control rates of convergence, we make assumptions on the complexity of the classes being considered in E(B). For our first main theorem (limit distribution of the test statistic), we require E(B)$_3$ and for our bootstrap theorems, we require either E(B)$_3$ or E(B)$_4$. For instance, if $\mathcal{F}_\mu$ is a class of Hölder continuous functions with Hölder exponent $\beta > 0$ on $D = d + 1$ dimensional Euclidean space, and we require E(B)$_m$ to hold, then the ε-entropy is of order $\varepsilon^{-D/\beta}$ so we require $mD/2\beta < 1$ or $\beta > mD/2$. When $m = 1$, the condition is the standard one and when $m = 3$ or 4 it is more restrictive.

REMARK 2.3. It may be possible to weaken these assumptions. The assumption E(B)$_m$ with $m > 1$ arises from certain (degenerate) U- or V-process terms in the analysis. Analyzing such terms requires these more stringent entropy conditions. On the other hand, an mth order (degenerate) U-process comes with a faster decay to 0, of order $n^{-m/2}$. When the class $\mathcal{F}$ (i.e., $\mathcal{F}_\mu$, $\mathcal{F}_\pi$) does not depend on n this does not help us. But if we allow a sieve-type approach where the class $\mathcal{F} \equiv \mathcal{F}_n$ depends on n, then we need $J_1(1, \mathcal{F}_n)$ to be $O(1)$ but $J_m(1, \mathcal{F}_n)$ for $m > 1$ can be allowed to grow with n; if we allow such sieve classes $\mathcal{F}_n$, then the only entropy required to stay finite/bounded in n is J_1, and so in this sense we can recover/require the more classical condition. At present, we have phrased the conditions only in terms of independent-of-n classes.

3. Main results.

Now we present the asymptotic distribution of our test statistic T_n under the null hypothesis. To metrize weak convergence, we use the Dudley metric (Shorack (2000), Chapter 14, Section 2) (although any topologically equivalent metric would work), which is defined as

$$d(\mu, \nu) := \sup\biggl\{\int g\, d\mu - \int g\, d\nu : \|g\|_{BL} \le 1\biggr\}, \tag{3.1}$$

where X and Y are random variables with probability distributions/laws μ and ν, respectively, and where $\|g\|_{BL} := \sup_{x\in\mathbb{R}} |g(x)| + \sup_{x\neq y} |g(x)-g(y)|/|x-y|$. For a kernel function K, we use $K^{(s)}$ to denote the s-times convolution product of K, that is $K^{(s)}(x) = \int K^{(s-1)}(y)K(x-y)\,dy$, with $K^{(1)} = K$. And we let $K_h^{(s)}(x) := K^{(s)}(x/h)$. Let

$$\sigma^2(a) = \mathbb{E}\biggl[\frac{\tau(\boldsymbol{L},a) + \{\mu_0(\boldsymbol{L},a) - \bar{\mu}(\boldsymbol{L},a)\}^2}{\{\bar{\pi}(a|\boldsymbol{L})/\bar{\varpi}(a)\}^2/\{\pi_0(a|\boldsymbol{L})/\varpi_0(a)\}}\biggr] - \bigl\{\theta_0(a) - \bar{m}(a)\bigr\}^2, \tag{3.2}$$

where $\bar{\varpi}(a) := \int \bar{\pi}(a|\boldsymbol{l})\,dP(\boldsymbol{l})$ and $\bar{m}(a) := \int \bar{\mu}(\boldsymbol{l},a)\,dP(\boldsymbol{l})$. We can now state our main theorem, which gives the limit distribution of our test statistic under the null hypothesis. Let $\mathcal{L}(X)$ denote the probability law of the random variable X.

THEOREM 3.1. *Let Assumptions* I1–I3, D1–D4, E(A)1–E(A)4, *and* E(B)3 *hold and let $w(\cdot)$ be a continuously differentiable weight function on $\mathcal{A}$. Let $\sigma^2(\cdot)$ and T_n be as given in* (3.2) *and* (2.5). *Then under H_0* (*from* (2.1)), *we have*

$$d\bigl\{\mathcal{L}(T_n), \mathcal{L}\bigl(N(b_{0h}, V)\bigr)\bigr\} \to 0 \tag{3.3}$$

as $n \to \infty$, where

(3.4)

$$b_{0h} = h^{-1/2}K^{(2)}(0)\int_{\mathcal{A}} \frac{\sigma^2(a)w(a)}{\varpi_0(a)}\,da, \qquad V = 2K^{(4)}(0)\int_{\mathcal{A}}\biggl[\frac{\sigma^2(a)w(a)}{\varpi_0(a)}\biggr]^2 da.$$

The full proof is given in the Supplementary Material (Doss et al. (2024)). We provide an outline of the proof here.

Proof outline for Theorem 3.1. We can decompose the statistic T_n as

$$T_n = n\sqrt{h}\int_{\mathcal{A}} \bigl(D_1(a) + D_2(a) + D_3\bigr)^2 w(a)\,da, \tag{3.5}$$

where

$$D_1(a) := \widehat{\theta}_h(a) - \tilde{\theta}_h(a), \qquad D_2(a) := \tilde{\theta}_h(a) - \mathbb{P}_n\bar{\xi}(\boldsymbol{Z}), \qquad D_3 := \mathbb{P}_n\bar{\xi}(\boldsymbol{Z}) - \mathbb{P}_n\widehat{\xi}(\boldsymbol{Z}),$$

and where $\tilde{\theta}_h(a)$ is the local linear estimator regressing the oracle doubly robust mappings $\bar{\xi}_i := \xi(\boldsymbol{Z}_i;\bar{\pi},\bar{\mu})$ on A_i. Expanding the square leads to 6 terms to be analyzed, so we break the proof up into 6 main steps corresponding to each of those terms. In Step 1, we verify that $n\sqrt{h}\int_{\mathcal{A}} D_2(a)^2 w(a)\,da$ is distributed approximately as the given $N(b_{0h}, V)$ limit distribution in the theorem. This follows essentially from Alcalá, Cristóbal and González-Manteiga (1999), which extends the results of Härdle and Mammen (1993) to allow local polynomial estimators. In Steps 2–6, we show that the remainder terms (which are just the terms $n\sqrt{h}\int_{\mathcal{A}} D_3^2 w(a)\,da$, $n\sqrt{h}\int_{\mathcal{A}} D_1(a)^2 w(a)\,da$, $n\sqrt{h}D_3\int_{\mathcal{A}} D_2(a)w(a)\,da$, $n\sqrt{h}\int_{\mathcal{A}} D_1(a)D_3 w(a)\,da$, and $n\sqrt{h}\int_{\mathcal{A}} D_1(a)D_2(a)w(a)\,da$) are all $o_p(1)$ as $n \to \infty$. We now will discuss those remainder terms in slightly more detail, basically in parallel with Steps 2–6.

Let $\eta := (\pi, \mu)$ (and $\widehat{\eta} := (\widehat{\pi}, \widehat{\mu})$, $\bar{\eta} := (\bar{\pi}, \bar{\mu})$). Both $D_1(a)$ and D_3 involve summations over $\widehat{\xi}(\boldsymbol{Z}_i;\widehat{\eta}) - \xi(\boldsymbol{Z}_i;\bar{\eta})$, and in analyzing the remainder terms, we break these summations into sums over $\widehat{\xi}(\boldsymbol{Z}_i;\widehat{\eta}) - \xi(\boldsymbol{Z}_i;\widehat{\eta})$ and $\xi(\boldsymbol{Z}_i;\widehat{\eta}) - \xi(\boldsymbol{Z}_i;\bar{\eta})$. (Recall the definition of $\widehat{\xi}$ given in (2.4), in which dP is replaced by $d\mathbb{P}_n$.) The sums over $\widehat{\xi}(\boldsymbol{Z}_i;\widehat{\eta}) - \xi(\boldsymbol{Z}_i;\widehat{\eta})$ yield terms that can be written as degenerate V-statistics (or, rather, because of the presence of the random $\widehat{\eta}$, terms whose size is governed by degenerate V-processes). We are are able to conclude that these are of very small order of magnitude, but unfortunately the empirical process tools (that we are aware of) for analyzing them requires the imposition of stronger entropy conditions than the normal Donsker-type condition. (Because of the presence of up to order 3

V-processes, we require $J_3(1, \mathcal{F}, L_2) < \infty$ rather than the weaker, more standard Donsker condition, $J_1(1, \mathcal{F}, L_2) < \infty$ (for $\mathcal{F}$ equal to both $\mathcal{F}_\mu$, $\mathcal{F}_\pi$).)

For instance, in Step 2 we write the term D_3 as

$$\mathbb{P}_n\{\xi(\mathbf{Z}; \bar{\pi}, \bar{\mu}) - \widehat{\xi}(\mathbf{Z}; \widehat{\pi}, \widehat{\mu})\} = \mathbb{P}_n\{\xi(\mathbf{Z}; \bar{\pi}, \bar{\mu}) - \xi(\mathbf{Z}; \widehat{\pi}, \widehat{\mu})\} + \mathbb{P}_n\{\xi(\mathbf{Z}; \widehat{\pi}, \widehat{\mu}) - \widehat{\xi}(\mathbf{Z}; \widehat{\pi}, \widehat{\mu})\}.$$

The first summand (of the right-hand side above) is further decomposed into two terms of types that commonly arise in causal inference or semiparametric problems one an empirical process one and one a "second order remainder"; the former is shown to be negligible by an empirical process asymptotic equicontinuity argument and the latter is small by assumptions on $r_n^\infty s_n^\infty$. The second summand in the display above can be written as a degenerate order 2 V-process. We provide an introduction to and discussion of U- and V-processes in Section K of our Supplementary Material (Doss et al. (2024)). Under Assumption E(B)$_2$ (implied by Assumption E(B)$_3$) we show that the V-process is negligible. Note that a degenerate order m U- or V-statistic is usually of order $n^{-m/2}$ (van der Vaart (1998), Chapter 12).

In Step 3, we write $D_1(a)$ as $D_1(a) = d_{1,1}(a) + d_{1,2}(a)$, where

$$d_{1,1}(a) := \mathbb{P}_n\big[W_{ha}(A)\big(\widehat{\xi}(\mathbf{Z}; \widehat{\pi}, \widehat{\mu}) - \xi(\mathbf{Z}; \widehat{\pi}, \widehat{\mu})\big)\big],$$
$$d_{1,2}(a) := \mathbb{P}_n\big[W_{ha}(A)\big(\xi(\mathbf{Z}; \widehat{\pi}, \widehat{\mu}) - \xi(\mathbf{Z}; \bar{\pi}, \bar{\mu})\big)\big],$$

and $W_{ha}(\cdot)$ are "equivalent" kernels for the local polynomial estimator (see the proof for definitions and references or see Fan and Gijbels ((1996), page 63)). The added difficulty now in Step 3 over Step 2 is that these terms are local that is, they depend on h, but thematically the decomposition works the same as in Step 2. The term $d_{1,2}(a)$ is handled by an asymptotic equicontinuity argument and assumptions on nuisance estimators, and the term $d_{1,1}(a)$ by a (nonasymptotic) V-process maximal inequality (see Proposition K.2) for an order 2 V-process.

In Step 4, we consider $\int_{\mathcal{A}} D_2(a) D_3 w(a)\,da$; here D_3 can be factored out and handled by the result of Step 2, and the remaining integral of $D_2(a)$ can be handled in an elementary way by Taylor expansion and the Central Limit Theorem. In Step 5, we consider $\int_{\mathcal{A}} D_1(a) D_3 w(a)\,da$, whose negligibility follows immediately from the analysis in Steps 2 and 3 and the Cauchy–Schwarz inequality.

Finally, in Step 6 an order 3 V-processes arises (and so Assumption E(B)$_3$ is needed) in the analysis of $\int_{\mathcal{A}} D_1(a) D_2(a) w(a)\,da$. This can be essentially simplified into (a sum of) two main terms,

$$\begin{aligned} &\int \frac{1}{\varpi_0^2(a)} \mathbb{P}_n\big[K_{ha}(A)\{\xi(\mathbf{Z}; \widehat{\pi}, \widehat{\mu}) - \xi(\mathbf{Z}; \bar{\pi}, \bar{\mu})\}\big] \frac{1}{n}\sum_{i=1}^n K_h(A_i - a)\bar{\varepsilon}_i w(a)\,da, \\ &\int \frac{1}{\varpi_0^2(a)} \mathbb{P}_n\big[K_{ha}(A)\{\widehat{\xi}(\mathbf{Z}; \widehat{\pi}, \widehat{\mu}) - \xi(\mathbf{Z}; \widehat{\pi}, \widehat{\mu})\}\big] \frac{1}{n}\sum_{i=1}^n K_h(A_i - a)\bar{\varepsilon}_i w(a)\,da, \end{aligned} \tag{3.6}$$

where $\overline{\varepsilon}_i := \xi(\mathbf{Z}_i; \overline{\pi}, \overline{\mu}) - \mathbb{P}\xi(\mathbf{Z}; \overline{\pi}, \overline{\mu})$. The first term in (3.6) is further decomposed, with $\mathbb{P}_n$ written as $(\mathbb{P}_n - \mathbb{P}) + \mathbb{P}$; usually, the $(\mathbb{P}_n - \mathbb{P})$ term yields an empirical process and the $\mathbb{P}$ term yields a second order remainder. Here the empirical process piece can be handled by previous arguments; the "$\mathbb{P}$" term can not be treated as a second order remainder because taking absolute values (as would be commonly done) breaks the mean zero structure of the term and does not yield the correct size (because of the integral over $\mathcal{A}$, the $n^{-1}\sum_{i=1}^n K_h(A_i - a)\bar{\varepsilon}_i w(a)$ term and $\mathbb{P}$ term cannot be analyzed separately). Rather, the term is handled by an empirical process (maximal inequality) argument (in Lemma E.4). Finally, the second term in (3.6) is the order three V-process, requiring $J_3(1, \cdot, L_2) < \infty$ in order to apply a maximal inequality.[4] That completes our proof outline; full details of the proof are given in Section B of the Supplementary Material, Doss et al. (2024).

Next, we state the consistency of our test under alternatives as follows. (The δ_n sequence can in particular be $(n\sqrt{h})^{1/2}$.)

THEOREM 3.2. *Let Assumptions* I1–I3, D1–D4, E(A)1–E(A)4 *hold and let $w(\cdot)$ be a continuously differentiable weight function on $\mathcal{A}$. Let $\sigma^2(\cdot)$ and T_n be as given in* (3.2) *and* (2.5). *Then under the alternative $\theta_0(a) = c_0 + \delta_n (n\sqrt{h})^{-1/2} g(a)$, where $c_0 = \mathbb{P}\xi(\boldsymbol{Z}; \pi_0, \mu_0)$, where $g(\cdot)$ is not the constant* 0, *and where $\int g(a)\varpi(a)w(a)\,da = 0$. Moreover, δ_n is a sequence converging to ∞ such that $\lim_{n\to\infty} n^{1/40}/\delta_n = 0$. Then we have*

$$P(T_n > z_{n,1-\alpha}) \to 1$$

as $n \to \infty$, where we use $z_{n,1-\alpha}$ to denote the upper α quantile of the $N(b_{0h}, V)$ distribution in (3.3).

The proof is given in the Supplementary Material (Doss et al. (2024)). The condition that $\int g(a)\varpi(a)w(a)\,da = 0$ is not a substantive restriction, it is just so that c_0 is “identifiable” in a sense.

Finally, we show that the bootstrap distribution of the statistic can be used to approximate T_n’s unknown distribution (under the null). For our proof to hold, we require an extra entropy condition to accommodate a fourth order V-process that appears in the analysis of the wild bootstrap. Recall the definition of J_4 from (2.8) and the definition of the Dudley metric for weak convergence in (3.1). Let $\mathcal{L}^*(X) := \mathcal{L}(X|\boldsymbol{Z}_1, \dots \boldsymbol{Z}_n)$ denote the conditional law of a random variable X.

A variety of bootstrap definitions are possible (and are included in the simulations) and discussed in Section 2.2. For the following theorem, we let $\widehat{\varepsilon}_i := \widehat{\xi}(\boldsymbol{Z}_i; \widehat{\pi}, \widehat{\mu}) - \sum_{i=1}^n \widehat{\xi}(\boldsymbol{Z}_i; \widehat{\pi}, \widehat{\mu})/n$ be centered at the null estimate, and we take ε_i^* to be the Rademacher choice so equal to $\pm\widehat{\varepsilon}_i$ with probability 1/2 each. Then we proceed as discussed in Section 2.2, with $(\xi_i^* = \mathbb{P}_n\widehat{\xi} + \varepsilon_i^*, A_i)$ as bootstrap observations and defining T_n^* by (2.5) but using the bootstrap observations.

THEOREM 3.3. *Let the assumptions of Theorem* 3.1 *hold. Let T_n^* be the bootstrap test statistic defined in the paragraph preceding this theorem. Then*

$$d(\mathcal{L}^*(T_n^*), \mathcal{L}(N(b_h, V))) \to_p 0$$

as $n \to \infty$.

Next, we consider the bootstrap where $\widehat{\varepsilon}_i := \widehat{\xi}(\boldsymbol{Z}_i; \widehat{\pi}, \widehat{\mu}) - \widehat{\theta}_h(A_i)$ (and the rest of the procedure is as described in the paragraph preceding the previous theorem). We study this case in the next theorem, which requires an extra entropy condition ($J_4 < \infty$).

THEOREM 3.4. *Let the assumptions of Theorem* 3.1 *hold. Further, we assume that $J_4(1, \mathcal{F}, L_2) < \infty$ for $\mathcal{F} = \mathcal{F}_\mu$ and for $\mathcal{F} = \mathcal{F}_\pi$. Let T_n^* be the bootstrap test statistic defined in the paragraph preceding this theorem. Then*

$$d(\mathcal{L}^*(T_n^*), \mathcal{L}(N(b_h, V))) \to_p 0$$

[4] Again, in the analysis of this [and other] term[s] we cannot analyze the multiplicands’ orders of magnitudes separately, because taking absolute values breaks the mean zero structure; that is, to explain, let C_i, D_i be generic random variables and then although we can bound $|\sum_i C_i D_i| \le (\max_i |C_i|) \sum_i |D_i|$, unfortunately then $\sum_i |D_i|$ (with absolute values) is not a sum of mean zero variables, even if the $\{D_i\}$ are mean zero; getting the right order of magnitude requires treating the multiplicands simultaneously as a V-process.

as $n \to \infty$.

The structure of the proofs is analogous to that of Theorem 3.1, although the calculations are somewhat more intricate, and lead to a fourth order V-process which requires the finiteness of J_4 as stated in the theorem. The proof details are given in the Supplementary Material (Doss et al. (2024)).

For more direct understanding of the two bootstraps, we studied both bootstraps in our simulations, and they seem to perform quite similarly (especially in low dimensionality/complexity problems (when we do not use cross-fitting)). The bootstrap studied in Theorem 3.4 is perhaps more common in practice, which is the reason for our presenting that theorem, despite its seemingly requiring stronger conditions for its theoretical justification than the Theorem 3.3 bootstrap. On the one hand, finiteness of J_4 is what arises fundamentally in the study of 4th order V-processes, and on the other perhaps the corresponding error terms are actually quite small (see also Remark 2.3) so that the two bootstraps perform similarly in practice. There is not a major difference in computation time; so either can be recommended for use in low complexity settings. When dimensionality (or complexity) grows, however, we recommend using the Theorem 3.3 bootstrap (together with cross-fitting). We discuss the complexity assumptions more in the next remark.

REMARK 3.1. As mentioned in Remark 2.2, β-Hölder classes with $\beta > mD/2$, $D = d + 1$, satisfy the needed entropy(-type) condition. To allow for more flexibility, (generalized) additive models (Hastie and Tibshirani (1987, 1990), Hastie (2017), Wood (2017)), may be used. Consider fitting the outcome regression, $\mu(a, l)$. One option is a model $\mu(a, l) = f_A(a) + \sum_{i=1}^{d} f_i(l_i)$ where f_A, f_i are Hölder functions, and here having $\beta > m/2$ suffices; if $m = 3$, then $\beta = 2$ works. In this setup one can recommend using any of a variety of smooth fits on the individual components (with the backfitting algorithm, Hastie and Tibshirani (1990)); common choices might be cubic splines or local linear regression. The shape constraint of convexity may also be used on a component while still satisfying the entropy condition (Chen and Samworth (2016)). If $m = 4$, then those choices just barely do not work in that we require $\beta > 2$. Higher order local polynomial regression such as local cubic regression can be recommended, though (see Chapter 7 of Fan and Gijbels (1996)). A user may wish to have a slightly more flexible model that allows interactions between covariates and the treatment, such as $\mu(a, l) = \sum_{i=1}^{d} f_i(l_i, a_i)$ (ignoring the "main effect" functions for simplicity). In this setup, One can again use local cubic regression which is appropriate for requiring the assumption $\beta > 3 = m \times 2/2$ ($m = 3$) or local 5th-order regression for $\beta > 4 = m \times 2/2$ ($m = 4$). One sees that when $m = 4$ the condition is somewhat restrictive.

In our implementations, we have included some machine learning methods (e.g., gradient boosting, random forests) for which we do not formally have verification of the entropy conditions (and our procedure worked well). Understanding entropy conditions for those estimators (rather, for function classes they live within) is an interesting direction for future work.

4. Simulation studies.

4.1. *Simulation for testing constant average treatment effect.* We use simulation to assess the performance of our proposed test in terms of both type I error probability and power. We consider two data generating models, one with a binary response and which is defined similarly as in Kennedy et al. (2017), and another one with a continuous response. In more detail, they are specified as follows.

Model 1: we simulate the covariates from independent standard normal distributions, $\boldsymbol{L} = (L_1, \dots, L_4)^T \sim N(0, \boldsymbol{I}_4)$, and simulate the treatment level from a Beta distribution,

$$(A/20)|\boldsymbol{L} \sim \text{Beta}\big(\lambda(\boldsymbol{L}), 1 - \lambda(\boldsymbol{L})\big),$$
$$\text{logit}\, \lambda(\boldsymbol{L}) = -0.8 + 0.1L_1 + 0.1L_2 - 0.1L_3 + 0.2L_4,$$

and finally the binary outcome is simulated as $Y|\boldsymbol{L}, A \sim \text{Bernoulli}(\mu(\boldsymbol{L}, A))$, where $\text{logit}\, \mu(\boldsymbol{L}, A) = 1 + (0.2, 0.2, 0.3, -0.1)\boldsymbol{L} + \delta A(0.1 - 0.1L_1 + 0.1L_3 - 0.13^2 A^2)$.

Model 2: the covariates are simulated the same as in Model 1. We simulate the treatment level from a Beta distribution,

$$(A/5)|\boldsymbol{L} \sim \text{Beta}\big(\lambda(\boldsymbol{L}), 1 - \lambda(\boldsymbol{L})\big),$$
$$\text{logit}\, \lambda(\boldsymbol{L}) = 0.1L_1 + 0.1L_2 - 0.1L_3 + 0.2L_4,$$

and we simulate the continuous response from conditional normal distributions as $Y|\boldsymbol{L}, A \sim N(\mu(\boldsymbol{L}, A), 0.5^2)$, where

$$\mu(\boldsymbol{L}, A) = (0.2, 0.2, 0.3, -0.1)\boldsymbol{L} + A(-0.1L_1 + 0.1L_3) + \delta \exp\left\{-\frac{(A - 2.5)^2}{(1/2)^2}\right\}.$$

In both models, we have a parameter δ that controls the distance between the true treatment effect and the null hypothesis, that is, treatment effect is constant, with $\delta = 0$ indicating no treatment effect in both models. In Model 1, $\delta = 0$ yields a constant average treatment effect and is a "strong null" meaning all individuals have the same treated and untreated outcomes; in Model 2 when $\delta = 0$ the "weak null" holds meaning conditional average treatment effects are nonconstant but the average treatment effect is constant. Specifically, for Model 1, we let $\delta \in \{0, 0.002, 0.004, 0.006, 0.008, 0.01\}$, and we let $\delta \in \{0, 0.1, 0.2, 0.3, 0, 4, 0.5\}$ for Model 2. We plot the treatment effect curve with $\delta = 0.01$ for Model 1 and the treatment effect curve with $\delta = 0.5$ for Model 2 in Figure 1.

For each data generating model, we test the performance of our method under 4 scenarios: (1) π is correctly specified with a parametric model, μ is incorrectly specified with a parametric model; (2) π is incorrectly specified with a parametric model, μ is correctly specified with a parametric model; (3) both π and μ are correctly specified with a parametric model; (4) both π and μ are estimated with Super Learners (van der Laan, Polley and Hubbard (2007)). In the first two scenarios, the incorrect parametric models are constructed in the same fashion as in Kang and Schafer (2007). The first three scenarios are used to test the double robustness of our method and the last one to show the empirical performance of our method when we use flexible machine learning models to estimate the nuisance functions. After we calculate the

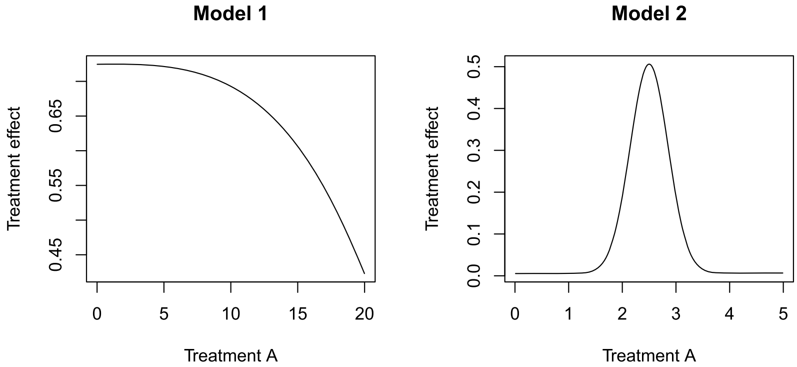

FIG. 1. *Treatment effect curves in Model* 1 *and Model* 2. *Left*: *Model* 1 *with* $\delta = 0.01$. *Right*: *Model* 2 *with* $\delta = 0.5$.

pseudo-outcomes, we use the rule of thumb for bandwidth selection (Fan and Gijbels (1996)) for the local linear estimator. We compare the performance of our method with Westling (2022) in the first three scenarios; with Westling (2022) and a discretized version of TMLE (Gruber and van der Laan (2012)) with treatment dichotomized at the middle point in the last scenario. In our method, we choose the weight function $w(a) \equiv 1$. We implemented all three versions (L_1, L_2 and L_∞) of the methods in Westling (2022) for comparison. Rejection probabilities are estimated with 1000 independent replications of simulation. Finally, we consider sample sizes in {500, 1000, 2000}.

Figures 2 and 3 show the results for Model 1. We can see when at least one of the nuisance functions is correctly estimated, our method and Westling's methods performed similarly in terms of both type I error probability and power. When both nuisance functions are estimated with Super Learners, Westling's methods have slightly larger power than our method but also have slightly larger type I error probabilities. Note that in this case, the discretized version of TMLE outperformed both our method and Westling's method in terms of type I error probability and power. The reason may be the shape of the treatment effect curve in Model 1 is somewhat simple and monotone, and there isn't much information loss if we dichotomize the continuous treatment to form a simpler testing problem. We also note that apparently Westling (2022)'s method is doubly robust on the specific data generating model we use here.

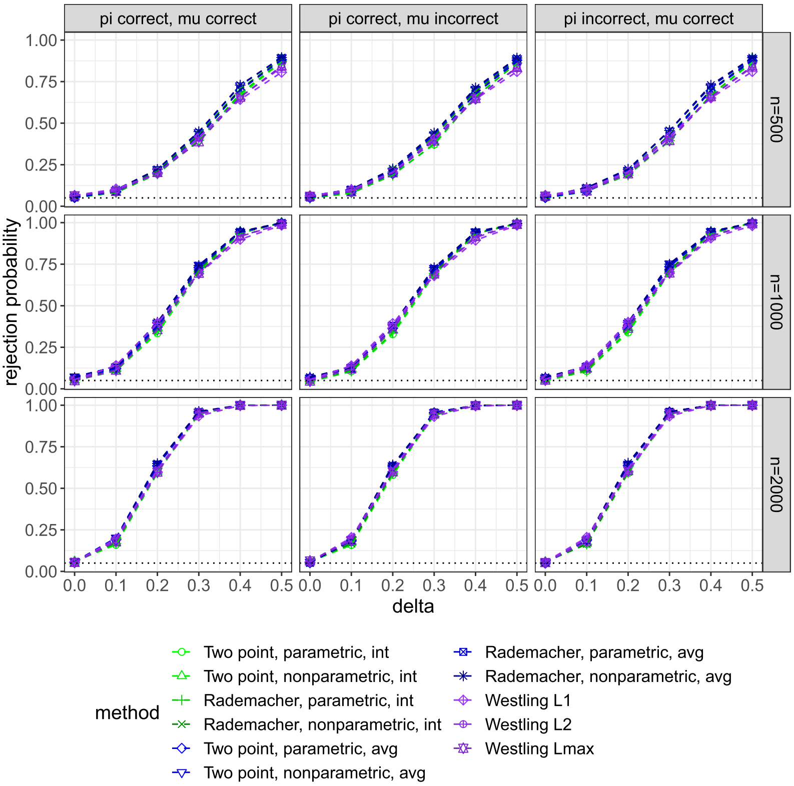

FIG. 2. *Simulation result for Model* 1 *with* π *and* μ *estimated from parametric models. The variations of our method are labeled as follows*: *Two point/Rademacher indicates whether the "Two point" distribution or Rademacher distribution is used to generate the bootstrap residuals*; *parametric/nonparametric indicates whether the residuals are estimated from a parametric treatment effect model* ($\hat{\varepsilon}_i = \widehat{\xi}(\mathbf{Z}_i; \widehat{\pi}, \widehat{\mu}) - \sum_{i=1}^n \widehat{\xi}(\mathbf{Z}_i; \widehat{\pi}, \widehat{\mu})/n$) *or a nonparametric treatment effect model* ($\hat{\varepsilon}_i = \widehat{\xi}(\mathbf{Z}_i; \widehat{\pi}, \widehat{\mu}) - \widehat{\theta}_h(A_i)$); *int means the test uses the original integral form of the test statistic and avg means the test uses the sample average approximation, that is*, $\sqrt{h}\sum_{i=1}^n \{\widehat{\theta}_h(A_i) - \mathbb{P}_n\widehat{\xi}(\mathbf{Z})\}^2$.

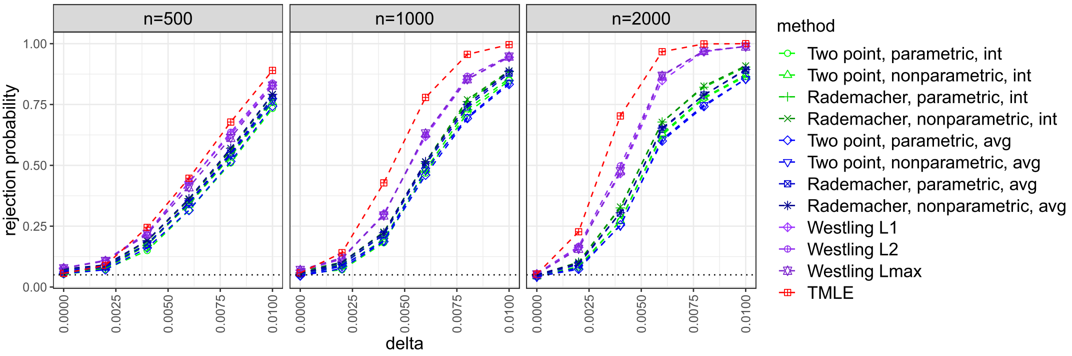

FIG. 3. *Simulation result for Model* 1 *with* π *and* μ *estimated from nonparametric models. See Figure* 2 *for label descriptions.*

Figures 4 and 5 show the results for Model 2 where we have a slightly more jagged and nonmonotone treatment effect curve (from the right-hand side of Figure 1). We observe that in the first three scenarios our methods outperform Westling's methods in terms of power in all cases. Our method does better even with a small sample size and a weak deviation from the null model, compared with Westling's method. Similar observations hold when we use Super Learner to estimate both nuisance functions. Another observation worth noting is that in Model 2 the discretized TMLE fails to detect any deviation from the null model since the treatment effect curve is symmetric, which provides an example in which discretizing a continuous treatment and applying a binary test may lead to a completely incorrect conclusion.

4.2. *Cross-fitted test procedure.* We also consider simulations to analyze how the dimensionality of the confounders $\boldsymbol{L}$ can affect the performance of our test and how cross-fitting could be applied to improve finite sample performance as dimension increases. Due to space constraints, we defer this material to Section J of Doss et al. (2024). The detailed description of the test procedure with cross-fitting can be found in Section J.1. We conduct simulation studies to compare our main proposed test (without cross-fitting) and the cross-fitted test under low dimensional data regimes and under increasing dimensionality regimes, respectively, in Section J.2. The following is a summary of the results from the simulations. When the dimensionality of the covariates is small, both noncross-fitted and cross-fitted tests can achieve the desired type I error probability but the cross-fitted version tends to have lower power; as the dimensionality of the covariates increases, only the cross-fitted version maintains the desired type I error probability. The non cross-fitted version fails even in the moderate (or "proportional") regime, when $d = 100$ and $n \in \{500, 1000\}$.

5. Analysis of data on nursing hours and hospital performance. In this section we apply our test to a real data problem. In Kennedy et al. (2017) and McHugh, Berez and Small (2013), the authors were interested in whether nurse staffing (measured in nurse hours per patient day) affected a hospital's risk of excess readmission penalty after adjusting for hospital characteristics (for more detail of the data and related background of the problem, see McHugh, Berez and Small (2013)). Kennedy et al. (2017) proposed a doubly robust procedure to estimate the probability of readmission penalty against average adjusted nursing hours per patient day, and provided pointwise confidence intervals for the estimated treatment curve. However, their method and analysis did not answer the question of whether nurse staffing significantly affects the probability of excess readmission penalty after adjusting for hospital characteristics. We apply our method to test the null hypothesis: nurse staffing does not affect

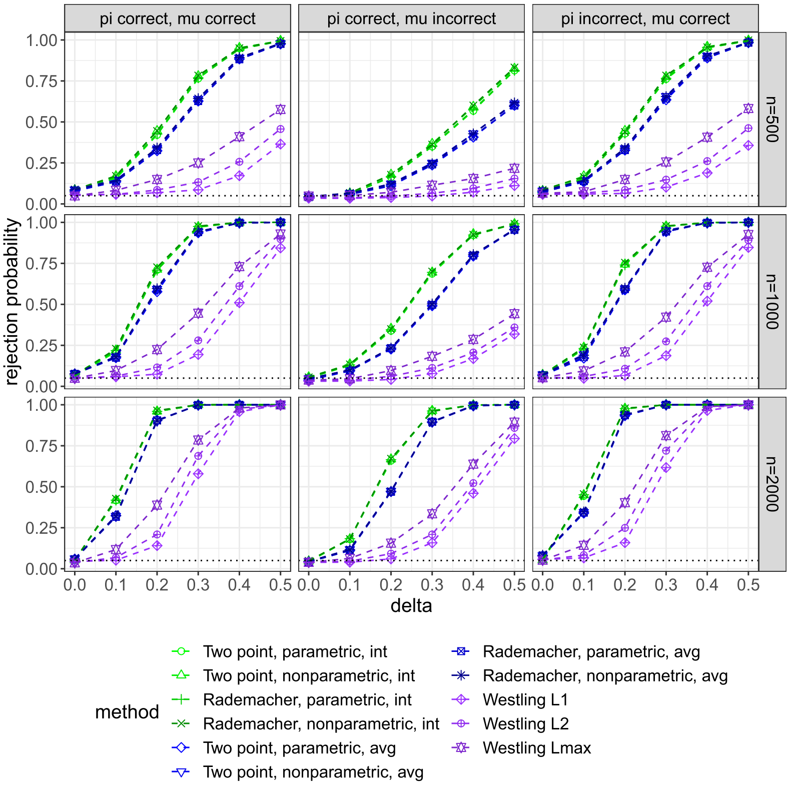

FIG. 4. *Simulation result for Model* 2 *with* π *and* μ *estimated from parametric models. See Figure* 2 *for label descriptions.*

hospital's risk of excess readmission penalty after adjusting for hospital characteristics, with updated data from the year 2018. As a brief summary of the data, the outcome Y indicates whether the hospital was penalized due to excess readmissions and are calculated by the Center for Medicare & Medicaid Services (https://www.cms.gov). The treatment A measures nurse staffing hours and we calculate it as the ratio of registered nurse hours to inpatient days, which is slightly different from Kennedy et al. (2017) and McHugh, Berez and Small (2013), because we don't have access to the hospitals' financial data and thus are not able to calculate adjusted inpatient days. The covariates $\boldsymbol{L}$ include the following nine variables: the number of beds, the teaching intensity, an indicator for not-for-profit status, an indicator for

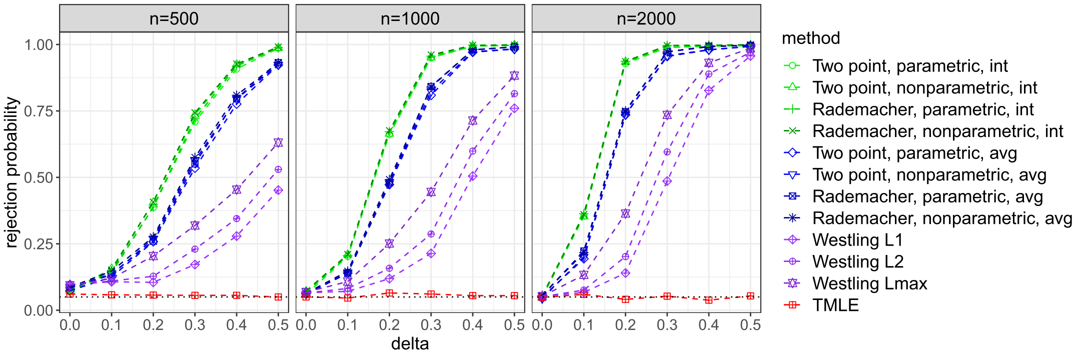

FIG. 5. *Simulation result for Model* 2 *with* π *and* μ *estimated from nonparametric models. See Figure* 2 *for label descriptions.*

whether the location is urban or rural, the proportion of patients on Medicaid, the average patient socioeconomic status, a measure of market competition, an indicator for whether the hospital has a skilled nursing facility (because our measure of nurse staffing hours A will unfortunately include hours worked in such a skilled nursing facility), and whether open heart or organ transplant surgery is performed (which serves as a measurement of whether the hospital is high technology). We omitted patient race proportions and operating margin from the analysis (present in Kennedy et al. (2017) and McHugh, Berez and Small (2013)) because we don't have access to those features. Figure I.3 shows an unadjusted loess fit of the readmission penalty as a function of the average nursing hours and the loess fits of the covariates against the average nursing hours. The curves are not identical to those in Kennedy et al. (2017) since we've used updated data from 2018, but we observe generally similar patterns and nurse staffing hours is correlated with many hospital characteristics. In the analysis, we use Super Learner (van der Laan, Polley and Hubbard (2007)) with the same implementation as in Kennedy et al. (2017) to estimate π_0 and μ_0. We truncate $\widehat{\pi}$ to be 0.01 if it fell below that value. The rule of thumb Fan and Gijbels (1996) is applied for bandwidth selection as in Section 4. Since our test statistic is based on the integrated distance between the nonparametric fit of the treatment effect curve and the parametric fit of the treatment effect curve under the null hypothesis, a byproduct of the test is the estimated treatment effect curve. We plot the estimated treatment effect curve of average nurse staffing in Figure 6 (the solid red curve).

We apply our test, Westling's test, and the discretized version of TMLE to this data set. All versions of our methods and Westling's methods have p-values of 0. (Exact zeros are due to the fact that we use simulated reference distributions.) The discretized TMLE reports a p-value of 0.0017. So all the tests suggest strong statistical evidence against the null hypothesis of constant treatment effect, meaning average nursing hours does have a significant causal impact on a hospital's chance of being penalized for excess readmissions. This interpretation requires that we have included all important confounders in our analysis. If we have not, our test result is interpreted as being based on a partially adjusted estimate of association (rather than the treatment effect curve).

Finally, we test whether an indicator for whether a hospital is in a rural or in an urban setting is a treatment effect modifier. We present the estimated conditional treatment effect curves for rural hospitals and for the urban hospitals in Figure 6 (dashed green for the rural hospitals and dotted blue for the urban hospitals). We observe that for the hospitals in urban areas, the pattern of the effect curve has a shape that is close to concave and is similar to the pattern in the overall average treatment effect curve: after average nursing hours exceeds 10, increasing the average nursing hour results in a decrease in the probability of the readmission

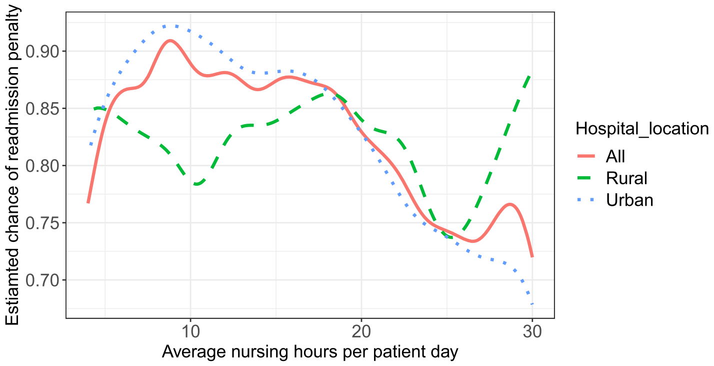

FIG. 6. *Estimated treatment effect of average nursing hours on probability of readmission penalty.*

penalty. The increasing trend of the curve up to 8 average nursing hours seems to be counterintuitive, but it turns out that there are not many hospitals in that range of the data and thus the left tail behavior is likely an artifact due to low sample size in that region. On the other hand, the effect curve for rural hospitals is wavy and does not suggest a clear pattern.

We first apply our main test separately to each of the two groups of hospitals to see whether the two individual treatment effect curves are constant or not. We obtain a p-value of 0.28 for the group of rural hospitals and a p-value of approximately 0 for the urban hospitals. This analysis suggests that the conditional treatment curve for the rural hospitals is not significantly different from constant, so the wavy pattern we see in the estimated curve is likely due to randomness. Next we apply the extended test procedure and obtain a p-value of 0.007, which indicates a significant difference between the two conditional treatment curves. Again, these interpretations require that we have included all important confounders in our analysis. It is somewhat surprising that in this dataset the rural hospitals show no significant effect of nurse staffing on readmission penalty. It is possible that hospital occupancy rates, case mix, financial stability and differing abilities to recruit and retain nurses are important for understanding the effect of nurse staffing on hospital performance, either as confounders or as treatment effect modifiers.

Acknowledgments. Guangwei Weng and Charles R. Doss are jointly first authors on the paper.

Funding. Charles R. Doss is partially supported by NSF grant DMS-1712664, NSF grant DMS-1712706, and NSF grant DMS-2210312.

SUPPLEMENTARY MATERIAL

Supplement: Proofs and technical details (DOI: 10.1214/24-AOS2405SUPP; .pdf). In the supplement, we provide proofs and technical details that were omitted from the main paper.

REFERENCES

Alcalá, J. T., Cristóbal, J. A. and González-Manteiga, W. (1999). Goodness-of-fit test for linear models based on local polynomials. *Statist. Probab. Lett.* **42** 39–46. MR1671753 https://doi.org/10.1016/S0167-7152(98)00184-9

Arcones, M. A. and Giné, E. (1993). Limit theorems for U-processes. *Ann. Probab.* **21** 1494–1542. MR1235426

Benkeser, D., Carone, M., van der Laan, M. J. and Gilbert, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika* **104** 863–880. MR3737309 https://doi.org/10.1093/biomet/asx053

Chen, G., Li, X. and Yu, M. (2022). Policy learning for optimal individualized dose intervals. In *Proceedings of the* 25*th International Conference on Artificial Intelligence and Statistics* (G. Camps-Valls, F. J. R. Ruiz and I. Valera, eds.). *Proceedings of Machine Learning Research* **151** 1671–1693. PMLR.

Chen, G., Zeng, D. and Kosorok, M. R. (2016). Personalized dose finding using outcome weighted learning. *J. Amer. Statist. Assoc.* **111** 1509–1521. MR3601705 https://doi.org/10.1080/01621459.2016.1148611

Chen, Y. and Samworth, R. J. (2016). Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 729–754. MR3534348 https://doi.org/10.1111/rssb.12137

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

Chernozhukov, V., Newey, W. K. and Singh, R. (2022). Debiased machine learning of global and local parameters using regularized Riesz representers. *Econom. J.* **25** 576–601. MR4565439 https://doi.org/10.1093/ectj/utac002

Colangelo, K. and Lee, Y.-Y. (2020). Double debiased machine learning nonparametric inference with continuous treatments. arXiv.org.

COULOMBE, J., MOODIE, E. E. M. and PLATT, R. W. (2021). Estimating the marginal effect of a continuous exposure on an ordinal outcome using data subject to covariate-driven treatment and visit processes. *Stat. Med.* **40** 5746–5764. MR4330577 https://doi.org/10.1002/sim.9151

DAVIDSON, R. and FLACHAIRE, E. (2008). The wild bootstrap, tamed at last. *J. Econometrics* **146** 162–169. MR2459651 https://doi.org/10.1016/j.jeconom.2008.08.003

DETTE, H. and NEUMEYER, N. (2001). Nonparametric analysis of covariance. *Ann. Statist.* **29** 1361–1400. MR1873335 https://doi.org/10.1214/aos/1013203458

DOSS, C. R., WENG, G., WANG, L., MOSCOVICE, I. and CHANTARAT, T. (2024). Supplement to "A nonparametric doubly robust test for a continuous treatment effect." https://doi.org/10.1214/24-AOS2405SUPP

DUDLEY, M. (1999). *Uniform Central Limit Theorems*. *Cambridge Studies in Advanced Mathematics* **63**. Cambridge University Press, Cambridge. MR1720712 https://doi.org/10.1017/CBO9780511665622

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903 https://doi.org/10.1007/978-1-4899-4541-9

EUBANK, R. L. and LARICCIA, V. N. (1992). Asymptotic comparison of Cramér–von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *Ann. Statist.* **20** 2071–2086. MR1193326 https://doi.org/10.1214/aos/1176348903

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587

FOSTER, D. J. and SYRGKANIS, V. (2023). Orthogonal statistical learning. *Ann. Statist.* **51** 879–908. MR4630373 https://doi.org/10.1214/23-AOS2258

GALVAO, A. F. and WANG, L. (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *J. Amer. Statist. Assoc.* **110** 1528–1542. MR3449052 https://doi.org/10.1080/01621459.2014.978005

GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics*, [40]. Cambridge Univ. Press, New York. MR3588285 https://doi.org/10.1017/CBO9781107337862

GRUBER, S. and VAN DER LAAN, M. (2012). tmle: An R package for targeted maximum likelihood estimation. *J. Stat. Softw.* **51** 1–35. https://doi.org/10.18637/jss.v051.i13

HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** 1926–1947. MR1245774 https://doi.org/10.1214/aos/1176349403

HASTIE, T. and TIBSHIRANI, R. (1987). Generalized additive models: some applications. *J. Amer. Statist. Assoc.* **82** 371–386.

HASTIE, T. J. (2017). Generalized additive models. In *Statistical Models in S* 249–307. Routledge, London.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. *Monographs on Statistics and Applied Probability* **43**. CRC Press, London. MR1082147

HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

HIRANO, K. and IMBENS, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. *Wiley Ser. Probab. Stat.* 73–84. Wiley, Chichester. MR2134803 https://doi.org/10.1002/0470090456.ch7

HOROWITZ, J. L. and SPOKOINY, V. G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* **69** 599–631. MR1828537 https://doi.org/10.1111/1468-0262.00207

IMAI, K. and VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866. MR2090918 https://doi.org/10.1198/016214504000001187

IMBENS, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

INGSTER, Y. I. (1993a). Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Math. Methods Statist.* **2** 85–114. MR1257978

INGSTER, Y. I. (1993b). Asymptotically minimax hypothesis testing for nonparametric alternatives. II. *Math. Methods Statist.* **2** 171–189. MR1257983

INGSTER, Y. I. (1993c). Asymptotically minimax hypothesis testing for nonparametric alternatives. III. *Math. Methods Statist.* **2** 249–268. MR1259685

KALLUS, N. and ZHOU, A. (2018). Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics* 1243–1251. PMLR.

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

KENNEDY, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electron. J. Stat.* **17** 3008–3049. MR4667730 https://doi.org/10.1214/23-ejs2157

KENNEDY, E. H., BALAKRISHNAN, S. and WASSERMAN, L. (2022). Minimax rates for heterogeneous causal effect estimation. arXiv.

KENNEDY, E. H., MA, Z., MCHUGH, M. D. and SMALL, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1229–1245. MR3689316 https://doi.org/10.1111/rssb.12212

KREIF, N., GRIEVE, R., DÍAZ, I. and HARRISON, D. (2015). Evaluation of the effect of a continuous treatment: A machine learning approach with an application to treatment for traumatic brain injury. *Health Econ.* **24** 1213–1228. https://doi.org/10.1002/hec.3189

LEE, S., OKUI, R. and WHANG, Y.-J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *J. Appl. Econometrics* **32** 1207–1225. MR3734484 https://doi.org/10.1002/jae.2574

LUEDTKE, A., CARONE, M. and VAN DER LAAN, M. J. (2019). An omnibus non-parametric test of equality in distribution for unknown functions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 75–99. MR3904780 https://doi.org/10.1111/rssb.12299

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Super-learning of an optimal dynamic treatment rule. *Int. J. Biostat.* **12** 305–332. MR3505699 https://doi.org/10.1515/ijb-2015-0052

MCHUGH, M. D., BEREZ, J. and SMALL, D. S. (2013). Hospitals with higher nurse staffing had lower odds of readmissions penalties than hospitals with lower staffing. *Health Aff.* **32** 1740–1747. https://doi.org/10.1377/hlthaff.2013.0613

NEUGEBAUER, R. and VAN DER LAAN, M. (2007). Nonparametric causal effects based on marginal structural models. *J. Statist. Plann. Inference* **137** 419–434. MR2298947 https://doi.org/10.1016/j.jspi.2005.12.008

NEWEY, W. K. and ROBINS, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. arXiv.

NIE, X. and WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. MR4259133 https://doi.org/10.1093/biomet/asaa076

POUET, C. (2001). An asymptotically optimal test for a parametric set of regression functions against a non-parametric alternative. *J. Statist. Plann. Inference* **98** 177–189. MR1860233 https://doi.org/10.1016/S0378-3758(00)00300-1

ROBINS, J., SUED, M., LEI-GOMEZ, Q. and ROTNITZKY, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable [MR2420458]. *Statist. Sci.* **22** 544–559. MR2420460 https://doi.org/10.1214/07-STS227D

ROBINS, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials (Minneapolis, MN,* 1997*). IMA Vol. Math. Appl.* **116** 95–133. Springer, New York. MR1731682 https://doi.org/10.1007/978-1-4612-1284-3_2

ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762 https://doi.org/10.2307/1912705

RUBIN, D. B. (1975). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

SCHULZ, J. and MOODIE, E. E. M. (2021). Doubly robust estimation of optimal dosing strategies. *J. Amer. Statist. Assoc.* **116** 256–268. MR4227692 https://doi.org/10.1080/01621459.2020.1753521

SEMENOVA, V. and CHERNOZHUKOV, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *Econom. J.* **24** 264–289. MR4281225 https://doi.org/10.1093/ectj/utaa027

SHORACK, G. R. (2000). *Probability for Statisticians. Springer Texts in Statistics*. Springer, New York. MR1762415

SU, L., URA, T. and ZHANG, Y. (2019). Non-separable models with high-dimensional data. *J. Econometrics* **212** 646–677. MR4001678 https://doi.org/10.1016/j.jeconom.2019.06.004

VAN DER LAAN, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int. J. Biostat.* **10** 29–57. MR3208072 https://doi.org/10.1515/ijb-2012-0038

VAN DER LAAN, M. J. and DUDOIT, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, Univ. California Berkeley.

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Art. 25, 23. MR2349918 https://doi.org/10.2202/1544-6115.1309

VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With applications to statistics. Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

WANG, J., WONG, R. K. W., YANG, S. and CHAN, K. C. G. (2022). Estimation of partially conditional average treatment effect by double kernel-covariate balancing. *Electron. J. Stat.* **16** 4332–4378. MR4474576 https://doi.org/10.1214/22-ejs2000

WESTLING, T. (2022). Nonparametric tests of the causal null with nondiscrete exposures. *J. Amer. Statist. Assoc.* **117** 1551–1562. MR4480731 https://doi.org/10.1080/01621459.2020.1865168

WESTLING, T., GILBERT, P. and CARONE, M. (2020). Causal isotonic regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 719–747. MR4112782 https://doi.org/10.1111/rssb.12372

WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R: An Introduction with R. Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. Second edition of [MR2206355]. MR3726911

ZIMMERT, M. and LECHNER, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. arXiv preprint. Available at arXiv:1908.08779.