HHS Public Access

Author manuscript

*J R Stat Soc Series B Stat Methodol*. Author manuscript; available in PMC 2020 February 01.

Published in final edited form as:

*J R Stat Soc Series B Stat Methodol*. 2019 February ; 81(1): 75–99. doi:10.1111/rssb.12299.

# An omnibus non-parametric test of equality in distribution for unknown functions

**Alexander R. Luedtke**,

Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, aluedtke@fredhutch.org

**Marco Carone**, and

Department of Biostatistics, University of Washington, Seattle, WA, USA

**Mark J. van der Laan**

Division of Biostatistics, University of California, Berkeley, Berkeley, CA, USA

## Abstract

We present a novel family of nonparametric omnibus tests of the hypothesis that two unknown but estimable functions are equal in distribution when applied to the observed data structure. We developed these tests, which represent a generalization of the maximum mean discrepancy tests described in Gretton et al. [2006], using recent developments from the higher-order pathwise differentiability literature. Despite their complex derivation, the associated test statistics can be expressed rather simply as U-statistics. We study the asymptotic behavior of the proposed tests under the null hypothesis and under both fixed and local alternatives. We provide examples to which our tests can be applied and show that they perform well in a simulation study. As an important special case, our proposed tests can be used to determine whether an unknown function, such as the conditional average treatment effect, is equal to zero almost surely.

### Keywords

higher order pathwise differentiability; maximum mean discrepancy; omnibus test; equality in distribution; infinite dimensional parameter

## 1. Introduction

In many scientific problems, it is of interest to determine whether two particular functions are equal to each other. In many settings these functions are unknown and may be viewed as features of a data-generating mechanism from which observations can be collected. As such, these functions can be learned from available data, and estimates of these respective functions can then be compared. To reduce the risk of deriving misleading conclusions due

Supplementary Appendices

Supplementary Appendix A reviews first- and second-order pathwise differentiability. Supplementary Appendix B.1 contains $U$ −process results from Nolan and Pollard [1987, 1988] that are useful in our context, and Supplementary Appendix B.2 contains an empirical process result used to establish the Donsker condition that was assumed under $\mathscr{H}_1$. Supplementary Appendix C contains proofs for the results in the main text.

to model misspecification, it is appealing to employ flexible statistical learning tools to estimate the unknown functions. Unfortunately, inference is usually extremely difficult when such techniques are used, because the resulting estimators tend to be highly irregular. In such cases, conventional techniques for constructing confidence intervals or computing p-values are generally invalid, and a more careful construction, as exemplified by the work presented in this article, is required.

To formulate the problem statistically, suppose that $n$ independent observations $O_1, O_2, \ldots, O_n$ are drawn from a distribution $P_0$ known only to lie in the nonparametric statistical model, denoted by $\mathcal{M}$. Let $\mathcal{O}$ denote the support of $P_0$, and suppose that $P \mapsto R_P$ and $P \mapsto S_P$ are parameters mapping from $\mathcal{M}$ onto the space of univariate bounded real-valued measurable functions defined on $\mathcal{O}$, i.e. $R_P$ and $S_P$ are elements of the space of univariate bounded real-valued measurable functions defined on $\mathcal{O}$. For brevity, we will write $R_0 \triangleq R_{P_0}$ and $S_0 \triangleq S_{P_0}$.

Our objective is to test the null hypothesis

$$\mathcal{H}_0 : R_0(O) \overset{d}{=} S_0(O)$$

versus the complementary alternative $\mathcal{H}_1$ : not $\mathcal{H}_0$, where $O$ follows the distribution $P_0$ and the symbol $\overset{d}{=}$ denotes equality in distribution. We note that $R_0(O) \overset{d}{=} S_0(O)$ if $R_0 \equiv S_0$, i.e. $R_0(O) = S_0(O)$ almost surely, but not conversely. The case where $S_0 \equiv 0$ is of particular interest since then the null simplifies to $\mathcal{H}_0 : R_0 \equiv 0$. Because $P_0$ is unknown, $R_0$ and $S_0$ are not readily available. Nevertheless, the observed data can be used to estimate $P_0$ and hence each of $R_0$ and $S_0$. The approach we propose will apply to functionals within a specified class described later.

Before presenting our general approach, we describe some motivating examples. Consider the data structure $O = (W, A, Y) \sim P$, where $W$ is a collection of covariates, $A$ is binary treatment indicator, and $Y$ is a bounded outcome, i.e., there exists a universal $c$ such that, for all $P \in \mathcal{M}$, $P(|Y| \leq c) = 1$. Note that, in our examples, the condition that $Y$ is bounded cannot easily be relaxed, as the parameter from Gretton et al. [2006] on which we will base our testing procedure requires that the quantities under consideration have compact support.

**Example 1:** Random sample size variant of the two-sample test from Gretton et al. [2006].

If $R_P(o) \triangleq ay$ and $S_P(o) \triangleq (1 - a)y$, the null hypothesis corresponds to $Y|A = 1$ and $Y|A = 0$ sharing the same distribution. This will differ from the setting considered in Gretton et al. [2006] in that, in our setting, the number of subjects with $A = 0$ and $A = 1$ will be treated as random, while the total number of observed subjects is fixed. This is in contrast to Gretton et al. [2006], who studied the case where the number of subjects with $A = 0$ and $A = 1$ were both fixed. This is the simplest example that we will give in this work. In particular, it is our only example in which the functions $R_P$ and $S_P$ do not rely on the (unknown) data generating distribution $P$.

**Example 2:** Testing a null conditional average treatment effect.

If $R_P(o) \triangleq E_P(Y \mid A = 1, W = w) - E_P(Y \mid A = 0, W = w)$ and $S_P \equiv 0$, the null hypothesis corresponds to the absence of a conditional average treatment effect. This definition of $R_P$ corresponds to the so-called blip function introduced by Robins [2004], which plays a critical role in defining optimal personalized treatment strategies [Chakraborty and Moodie, 2013].

**Example 3:** Testing for equality in distribution of regression functions in two populations.

Suppose the setting of the previous example, but where $A$ represents membership to population 0 or 1. If $R_P(o) \triangleq E_P(Y \mid A = 1, W = w)$ and $S_P(o) \triangleq E_P(Y \mid A = 0, W = w)$, the null hypothesis corresponds to the outcome having the conditional mean functions, applied to a random draw of the covariate, having the same distribution in these two populations. We note here that our formulation considers selection of individuals from either population as random rather than fixed so that population-specific sample sizes (as opposed to the total sample size) are themselves random. The same interpretation could also be used for the previous example, now testing if the two regression functions are equivalent.

**Example 4:** Testing a null covariate effect on average response.

Suppose now that the data unit only consists of $O \triangleq (W, Y)$. If $R_P(o) \triangleq E_P(Y \mid W = w)$ and $S_P \equiv 0$, the null hypothesis corresponds to the outcome $Y$ having conditional mean zero in all strata of covariates. This may be interesting when zero has a special importance for the outcome, such as when the outcome is the profit over some period.

**Example 5:** Testing a null variable importance.

Suppose again that $O \triangleq (W, Y)$ and $W \triangleq (W(1), W(2), \ldots, W(K))$. Denote by $W(-k)$ the vector $(W(i) : 1 \leq i \leq K, i \neq k)$. Setting $R_P(o) \triangleq E_P(Y \mid W = w)$ and $S_P(o) \triangleq E_P(Y \mid W(-k) = w(-k))$, the null hypothesis corresponds to $W(k)$ having null variable importance in the presence of $W(-k)$ with respect to the conditional mean of $Y$ given $W$ in the sense that $E_P(Y \mid W) = E_P(Y \mid W(-k))$ almost surely. This is true because if $R_0(W) \overset{d}{=} S_0(W(-k))$, the latter random variables have equal variance and so

$$E_{P_0}\left\{Var_{P_0}\left[R_0(W) \middle| W(-k)\right]\right\} = Var_{P_0}\left[R_0(W)\right] - Var_{P_0}\left\{E_{P_0}\left[R_0(W) \middle| W(-k)\right]\right\}$$
$$= Var_{P_0}\left[R_0(W)\right] - Var_{P_0}\left[S_0(W(-k))\right] = 0,$$

implying that $Var_{P_0}[R_0(W) \mid W(-k)] = 0$ almost surely. Thus, a test of $R_P(O) \stackrel{d}{=} S_P(O)$ is equivalent to a test of almost sure equality between $R_P$ and $S_P$ in this example. We will show in Section 5 that our approach cannot be directly applied to this example, but that a simple extension yields a valid test.

Gretton et al. [2006] investigated the related problem of testing equality between two distributions in a two-sample problem. They proposed estimating the maximum mean discrepancy (hereafter referred to as MMD), a non-negative quantitative summary of the relationship between the two distributions. In particular, the MMD between distributions $P_1$ and $P_2$ for observations $X$ is defined as

$$\sup_{f \in \mathcal{F}} \left( E_{P_1}[f(X)] - E_{P_2}[f(X)] \right). \quad (1)$$

Defining the MMD relies on selecting a function class $\mathcal{F}$. Gretton et al. [2006] propose selecting $\mathcal{F}$ to be the unit ball in a reproducing kernel Hilbert space. If the kernel defining this space is a so-called universal kernel and the support of $X$ under $P_1$ and $P_2$ is compact, then they showed that the MMD is zero if and only if the two distributions are equal. They also observe that the Gaussian kernel is a universal kernel. Gretton et al. also investigated related problems using this technique [see, e.g., Gretton et al., 2009, 2012a, Sejdinovic et al., 2013]. In this work, we also utilize the MMD as a parsimonious summary of equality but consider the more general problem wherein the null hypothesis relies on unknown functions $R_0$ and $S_0$ indexed by the data-generating distribution $P_0$.

Other investigators have proposed omnibus tests of hypotheses of the form $\mathcal{H}_0$ versus $\mathcal{H}_1$ in the literature. In the setting of Example 2 above, the work presented in Racine et al. [2006] and Lavergne et al. [2015] is particularly relevant. The null hypothesis of interest in these papers consists of the equality $E_{P_0}(Y \mid A, W) = E_{P_0}(Y \mid W)$ holding almost surely. If individuals have a nontrivial probability of receiving treatment in all strata of covariates, this null hypothesis is equivalent to $\mathcal{H}_0$. In both these papers, kernel smoothing is used to estimate the required regression functions. Therefore, key smoothness assumptions are needed for their methods to yield valid conclusions. The method we present does not hinge on any particular class of estimators and therefore does not rely on this condition.

To develop our approach, we use techniques from the higher-order pathwise differentiability literature [see, e.g., Pfanzagl, 1985, Robins et al., 2008, van der Vaart, 2014, Carone et al., 2014]. Despite the elegance of the theory presented by these various authors, it has been unclear whether these higher-order methods are truly useful in infinite-dimensional models since most functionals of interest fail to be even second-order pathwise differentiable in such models. This is especially troublesome in problems in which under the null the first-order derivative of the parameter of interest (in an appropriately defined sense) vanishes, since then there seems to be no theoretical basis for adjusting parameter estimates to recover parametric rate asymptotic behavior. At first glance, the MMD parameter seems to provide

one such disappointing example, since its first-order derivative indeed vanishes under the null. The latter fact is a common feature of problems wherein the null value of the parameter is on the boundary of the parameter space. It is also not an entirely surprising phenomenon, at least heuristically, since the MMD achieves its minimum of zero under the null hypothesis. Nevertheless, we are able to show that this parameter is indeed second-order pathwise differentiable under the null hypothesis – this is a rare finding in infinite-dimensional models. As such, we can employ techniques from the recent higher-order pathwise differentiability literature to tackle the problem at hand.

This paper is organized as follows. In Section 2, we formally present our parameter of interest, the squared MMD between two unknown functions, and establish asymptotic representations for this parameter based on its higher-order differentiability, which, as we formally establish, holds even when the MMD involves estimation of unknown nuisance parameters. In Section 3, we discuss estimation of this parameter, discuss the corresponding hypothesis test and study its asymptotic behavior under the null. We study the consistency of our proposed test under fixed and local alternatives in Section 4. We revisit our examples in Section 5 and provide an additional example in which we can still make progress using our techniques even though our regularity conditions fails. In Section 6, we present results from a simulation study to illustrate the finite-sample performance of our test, and we end with concluding remark in Section 7.

Supplementary Appendix A reviews higher-order pathwise differentiability. Supplementary Appendix B gives a summary of the empirical $U$–process results from Nolan and Pollard [1988] that we build upon. All proofs can be found in Supplementary Appendix C.

## 2. Properties of maximum mean discrepancy

### 2.1. Definition

For a distribution $P$ and mappings $T$ and $U$, we define

$$\Phi^{TU}(P) \triangleq \int\int e^{-\left[T_P(o_1) - U_P(o_2)\right]^2} dP(o_1) dP(o_2) \quad (2)$$

and set $\Psi(P) \triangleq \Phi^{RR}(P) - 2\Phi^{RS}(P) + \Phi^{SS}(P)$. The MMD between the distributions of $R_P(O)$ and $S_P(O)$ when $O \sim P$, defined in Eq. 1 using $\mathcal{F}$ to be the unit ball in the RKHS generated by the Gaussian kernel with unit bandwidth, is given by $\sqrt{\Psi(P)}$ and is always well-defined because $\Psi(P)$ is non-negative. Indeed, denoting by $\psi_0$ the true parameter value $\Psi(P_0)$, Theorem 3 of Gretton et al. [2006] establishes that $\psi_0$ equals zero if $\mathcal{H}_0$ holds and is otherwise strictly positive. Though the study in Gretton et al. [2006] is restricted to two-sample problems, their proof of this result is only based upon properties of $\Psi$ and therefore holds regardless of the sample collected. Their proof relies on the fact that two random variables $X$ and $Y$ with compact support are equal in distribution if and only if $E[f(Y)] = E[f(X)]$ for every continuous function $f$, and uses techniques from the theory of Reproducing Kernel Hilbert Spaces [see, e.g., Berlinet and Thomas-Agnan, 2011, for a general

exposition]. We invite interested readers to consult Gretton et al. [2006] – and, in particular, Theorem 3 therein – for additional details. The definition of the MMD we utilize is based on the univariate Gaussian kernel with unit bandwidth, which is appropriate in view of Steinwart [2002]. The results we present in this paper can be generalized to the MMD based on a Gaussian kernel of arbitrary bandwidth $h$ by simply rescaling the mappings $R$ and $S$ to $R/h$ and $S/h$.

### 2.2. First-order differentiability

To develop a test of $\mathscr{H}_0$, we will first construct an estimator $\psi_n$ of $\psi_0$. In order to avoid restrictive model assumptions, we wish to use flexible estimation techniques in estimating $P_0$ and therefore $\psi_0$. To control the operating characteristics of our test, it will be crucial to understand how to generate a parametric-rate estimator of $\psi_0$. For this purpose, it is informative to first investigate the pathwise differentiability of $\Psi$ as a parameter from $\mathscr{M}$ to $\mathbb{R}$.

So far, we have not specified restrictions on the mappings $P \mapsto R_P$ and $P \mapsto S_P$. However, in our developments, we will require these mappings to satisfy certain regularity conditions. Specifically, we will restrict our attention to elements of the class $\mathscr{S}$ of all mappings $T$ for which there exists some measurable function $X^T$ defined on $\mathcal{O}$, e.g. $X^T(o) = X^T(w, a, y) = w$, such that

**(S1)** $T_P$ is a measurable mapping with domain $\left\{X^T(o) : o \in \mathcal{O}\right\}$ and range contained in $[-b, b]$ for some $0 \leq b < \infty$ independent of $P$;

**(S2)** for all submodels $dP_t/dP = 1 + th$ with bounded $h$ with $Ph = 0$, there exists some $\delta > 0$ and a set $\mathcal{O}_1 \subseteq \mathcal{O}$ with $P_0(\mathcal{O}_1) = 1$ such that, for all $(o, t_1) \in \mathcal{O}_1 \times (-\delta, \delta)$, $t \mapsto T_{P_t}(x^T)$ is twice differentiable at $t_1$ with uniformly bounded (in $x^T$) first and second derivatives;

**(S3)** for any $P \in \mathscr{M}$ and submodel $dP_t/dP = 1 + th$ for bounded $h$ with $Ph = 0$, there exists a function $D_P^T : \mathcal{O} \to \mathbb{R}$ uniformly bounded (in $P$ and $o$) such that $\int D_P^T(o) dP(o \mid x^T) = 0$ for almost all $o \in \mathcal{O}$ and

$$\frac{d}{dt} T_{P_t}\left(x^T\right)\Big|_{t=0} = \int D_P^T(o) h(o) dP\left(o \mid x^T\right).$$

Condition (S1) ensures that $T$ is bounded and only relies on a summary measure of an observation $O$. Condition (S2) ensures that we will be able to interchange differentiation and integration when needed. Condition (S3) is a conditional (and weaker) version of pathwise differentiability in that the typical inner product representation only needs to hold for the conditional distribution of $O$ given $X^T$ under $P_0$. We will verify in Section 5 that these conditions hold in the context of the motivating examples presented earlier.

REMARK 1. *As a caution to the reader, we warn that simultaneously satisfying (S1) and (S3) may at times be restrictive. For example, if the observed data unit is* $O \triangleq (W(1), W(2), Y)$, *the parameter*

$$T_P(o) \triangleq E_P[Y \mid W(1) = w(1), W(2) = w(2)] - E_P[Y \mid W(1) = w(1)]$$

*cannot generally satisfy both conditions. In* Section 5, *we discuss this example further and provide a means to tackle this problem using the techniques we have developed. In concluding remarks, we discuss a weakening of our conditions, notably by replacing* $\mathcal{S}$ *by the linear span of elements in* $\mathcal{S}$. *Consideration of this larger class significantly complicates the form of the estimator we propose in* Section 3. □

We are now in a position to discuss the pathwise differentiability of $\Psi$. For any elements $T, U \in \mathcal{S}$, we define

$$\Gamma_P^{TU}(o_1, o_2) \triangleq \left[2\left[T_P(o_1) - U_P(o_2)\right]\left[D_P^U(o_2) - D_P^T(o_1)\right] + 1 - \left\{4\left[T_P(o_1) - U_P(o_2)\right]^2 - 2\right\}D_P^T(o_1)D_P^U(o_2)\right]e^{-\left[T_P(o_1) - U_P(o_2)\right]^2}.$$

and set $\Gamma_P \triangleq \Gamma_P^{RR} - \Gamma_P^{RS} - \Gamma_P^{SR} + \Gamma_P^{SS}$. Note that $\Gamma_P$ is symmetric for any $P \in \mathcal{M}$. For brevity, we will write $\Gamma_0^{TU}$ and $\Gamma_0$ to denote $\Gamma_{P_0}^{TU}$ and $\Gamma_{P_0}$, respectively. The following theorem characterizes the first-order behavior of $\Psi$ at an arbitrary $P \in \mathcal{M}$.

THEOREM 1 (FIRST-ORDER PATHWISE DIFFERENTIABILITY OF $\Psi$ OVER $\mathcal{M}$). *If* $R, S \in \mathcal{S}$, *the parameter* $\Psi : \mathcal{M} \to \mathbb{R}$ *is pathwise differentiable at* $P \in \mathcal{M}$ *with first-order canonical gradient given by* $D_1^\Psi(P)(o) \triangleq 2\left[\int \Gamma_P(o, o_2)dP(o_2) - \Psi(P)\right]$.

Under some conditions, it is straightforward to construct an asymptotically linear estimator of $\psi_0$ with influence function $D_1^\Psi(P_0)$, that is, an estimator $\psi_n$ of $\psi_0$ such that

$$\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} D_1^\Psi(P_0)(O_i) + o_{P_0}\left(n^{-1/2}\right).$$

For example, the one-step Newton-Raphson bias correction procedure [see, e.g., Pfanzagl, 1982] or targeted minimum loss-based estimation [see, e.g., van der Laan and Rose, 2011] can be used for this purpose. If the above representation holds and the variance of $D_1^\Psi(P_0)(O)$ is positive, then $\sqrt{n}(\psi_n - \psi_0) \rightsquigarrow N(0, \sigma_0^2)$, where the symbol $\rightsquigarrow$ denotes convergence in distribution and we write $\sigma_0^2 \triangleq P_0[D_1^\Psi(P_0)^2]$. If $\sigma_0$ is strictly positive and can be consistently estimated, Wald-type confidence intervals for $\psi_0$ with appropriate asymptotic coverage can be constructed.

The situation is more challenging if $\sigma_0 = 0$. In this case, $\sqrt{n}(\psi_n - \psi_0) \to 0$ in probability and typical Wald-type confidence intervals will not be appropriate. Because $D_1^{\Psi}(P_0)(O)$ has mean zero under $P_0$, this happens if and only if $D_1^{\Psi}(P_0) \equiv 0$. The following lemma provides necessary and sufficient conditions under which $\sigma_0 = 0$.

COROLLARY 1 (FIRST-ORDER DEGENERACY UNDER $\mathscr{H}_0$). *If $R, S \in \mathcal{S}$, it will be the case that $\sigma_0 = 0$ if and only if either (i) $\mathscr{H}_0$ holds, or (ii) $R_0(O)$ and $S_0(O)$ are degenerate, i.e. almost surely constant but not necessarily equal, with $D_0^R \equiv D_0^S$.*

The above results rely in part on knowledge of $D_0^R$ and $D_0^S$. It is useful to note that, in some situations, the computation of $D_P^T(o)$ for a given $T \in \mathcal{S}$ and $P \in \mathcal{M}$ can be streamlined. This is the case, for example, if $P \mapsto T_P$ is invariant to fluctuations of the marginal distribution of $X^T$, as it seems (S3) may suggest. Consider obtaining iid samples of increasing size from the conditional distribution of $O$ given $X^T = x^T$ under $P$, so that all individuals have observed $X^T = x^T$. Consider the fluctuation submodel $dP_t(o \mid x^T) \triangleq [1 + th(o)]dP(o \mid x^T)$ for the conditional distribution, where $h$ is uniformly bounded and $\int h(o)dP(o \mid x^T) = 0$. Suppose that (i) $P \mapsto T_P(x^T)$ is differentiable at $t = 0$ with respect to the above submodel and (ii) this derivative satisfies the inner product representation

$$\frac{d}{dt} T_{P_t}(x^T)\Big|_{t=0} = \int \tilde{D}_P^T(o \mid x^T) h(o) dP(o \mid x^T)$$

for some uniformly bounded function $o \mapsto \tilde{D}_P^T(o \mid x^T)$ with $\int \tilde{D}_P^T(o \mid x^T) dP(o \mid x^T) = 0$. If the above holds for all $x^T$, we may take $D_P^T(o) = \tilde{D}_P^T(o \mid x^T)$ for all $o$ with $X^T(o) = x^T$. If $D_P^T$ is uniformly bounded in $P$, (S3) then holds.

In summary, the above discussion suggests that, if $T$ is invariant to fluctuations of the marginal distribution of $X^T$, (S3) can be expected to hold if there exists a regular, asymptotically linear estimator of each $T_P(x^T)$ under iid sampling from the conditional distribution of $O$ given $X^T = x^T$ implied by $P$.

REMARK 2. *If $T$ is invariant to fluctuations of the marginal distribution of $X^T$, one can also expect (S3) to hold if $P \mapsto \int T_P(X^T(o))dP(o)$ is pathwise differentiable with canonical gradient uniformly bounded in $P$ and $o$ in the model in which the marginal distribution of $X$ is known. The canonical gradient in this model is equal to $D_P^T$.* □

## 2.3. Second-order differentiability and asymptotic representation

As indicated above, if $\sigma_0 = 0$, the behavior of $\Psi$ around $P_0$ cannot be adequately characterized by a first-order analysis. For this reason, we must investigate whether $\Psi$ is

second-order differentiable. As we discuss below, under $\mathscr{H}_0$, $\Psi$ is indeed second-order pathwise differentiable at $P_0$ and admits a useful second-order asymptotic representation.

THEOREM 2 (SECOND-ORDER PATHWISE DIFFERENTIABILITY UNDER $\mathscr{H}_0$). *If $R, S \in \mathcal{S}$ and $\mathscr{H}_0$ holds, the parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is second-order pathwise differentiable at $P_0$ with second-order canonical gradient $D_2^{\Psi}(P_0) \triangleq 2\,\Gamma_0$.*

It is easy to confirm that $\Gamma_0$, and thus $D_2^{\Psi}$, is one-degenerate under $\mathscr{H}_0$ in the sense that $\int \Gamma_0(o,o_2)dP_0(o_2) = \int \Gamma_0(o_1, o)dP_0(o_1) = 0$ for all $o$. This is shown as follows. For any $T, U \in \mathcal{S}$, the law of total expectation conditional on $X^U$ and fact that $\int D_0^U(o)dP_0(o \mid x^U) = 0$ yields that

$$\int \Gamma_0^{TU}(o, o_2)dP_0(o_2) = \int \left\{1 - 2\left[T_0(o) - U_0(o_2)\right]D_0^T(o)\right\} e^{-\left[T_0(o) - U_0(o_2)\right]^2} dP_0(o_2),$$

where we have written $\Gamma_0^{TU}$ to denote $\Gamma_{P_0}^{TU}$. Since $\int f(R_0(o))dP_0(o) = \int f(S_0(o))dP_0(o)$ for each measurable function $f$ when $S_0(O) \overset{d}{=} T_0(O)$, this then implies that $\int \Gamma_0^{RS}(o, o_2)dP_0(o_2) = \int \Gamma_0^{RR}(o, o_2)dP(o_2)$ and $\int \Gamma_0^{SR}(o, o_2)dP_0(o_2) = \int \Gamma_0^{SS}(o, o_2)dP_0(o_2)$ under $\mathscr{H}_0$. Hence, it follows that $\int \Gamma_0(o, o_2)dP_0(o_2) = 0$ under $\mathscr{H}_0$ for any $o$.

If second-order pathwise differentiability held in a sufficiently uniform sense over $\mathcal{M}$, we would expect

$$\mathrm{Rem}_P^{\Psi} \triangleq \Psi(P) - \Psi(P_0) - (P - P_0)D_1^{\Psi}(P) + \frac{1}{2}(P - P_0)^2 D_2^{\Psi}(P) \quad (3)$$

to be a third-order remainder term. However, second-order pathwise differentiability has only been established under the null, and in fact, it appears that $\Psi$ may not generally be second-order pathwise differentiable under the alternative. As such, $D_2^{\Psi}$ may not even be defined under the alternative. In writing (3), we either naively set $D_2^{\Psi}(P) \triangleq 2\,\Gamma_P$, which is not appropriately centered to be a candidate second-order gradient, or instead take $D_2^{\Psi}$ to be the centered extension

$$(o_1, o_2) \mapsto 2\left[\Gamma_P(o_1, o_2) - \int \Gamma_P(o_1, o)dP(o) - \int \Gamma_P(o, o_2)dP(o) + P^2\,\Gamma_P\right].$$

Both of these choices yield the same expression above because the product measure $(P - P_0)^2$ is self-centering. The need for an extension renders it a priori unclear whether as $P$

tends to $P_0$ the behavior of $\mathrm{Rem}_P^{\Psi}$ is similar to what is expected under more global second-order pathwise differentiability. Using the fact that $\Psi(P) = P^2\Gamma_P$, we can simplify the expression in (3) to

$$\mathrm{Rem}_P^{\Psi} = P_0^2\,\Gamma_P\ - \psi_0\,. \quad (4)$$

As we discuss below, this remainder term can be bounded in a useful manner, which allows us to determine that it is indeed third-order.

For all $T \in \mathcal{S}$, $P \in \mathcal{M}$ and $o \in \mathcal{O}$, we define

$$\mathrm{Rem}_P^T(o) \triangleq T_P(o) - T_0(o) + \int D_P^T(o_1)\left[dP\left(o_1\middle|x^T\right) - dP_0\left(o_1\middle|x^T\right)\right]$$

as the remainder from the linearization of $T$ based on the conditional gradient $D_P^T$. Typically, $\mathrm{Rem}_P^T(o)$ is a second-order term. Further consideration of this term in the context of our motivating examples is described in Section 5. Furthermore, we define

$$L_P^{RS}(o) \triangleq \max\left\{\left|\mathrm{Rem}_P^R(o)\right|, \left|\mathrm{Rem}_P^S(o)\right|\right\}$$
$$M_P^{RS}(o) \triangleq \max\left\{\left|R_P(o) - R_0(o)\right|, \left|S_P(o) - S_0(o)\right|\right\}.$$

For any given function $f : \mathcal{O} \to \mathbb{R}$, we denote by $\|f\|_{p,P_0} \triangleq \left[\int |f(o)|^p\, dP_0(o)\right]^{1/p}$ the $L^p(P_0)$-norm and use the symbol $\lesssim$ to denote 'less than or equal to up to a positive multiplicative constant'. The following theorem provides an upper bound for the remainder term of interest.

THEOREM 3 (UPPER BOUNDS ON REMAINDER TERM). *For each $P \in \mathcal{M}$, the remainder term, admits the following upper bounds:*

$$\textit{Under } \mathcal{H}_0 : \left|\mathrm{Rem}_P^{\Psi}\right| \lesssim K_{0P} \triangleq \left\|L_P^{RS}\right\|_{2,P_0}\left\|M_P^{RS}\right\|_{2,P_0} + \left\|L_P^{RS}\right\|_{1,P_0}^2 + \left\|M_P^{RS}\right\|_{4,P_0}^4$$
$$\textit{Under } \mathcal{H}_1 : \left|\mathrm{Rem}_P^{\Psi}\right| \lesssim K_{1P} \triangleq \left\|L_P^{RS}\right\|_{1,P_0}\left\|M_P^{RS}\right\|_{2,P_0}^2.$$

To develop a test procedure, we will require an estimator of $P_0$, which will play the role of $P$ in the above expressions. It is helpful to think of parametric model theory when interpreting the above result, with the understanding that certain smoothing methods, such as higher-order kernel smoothing, can achieve near-parametric rates in certain settings. In a parametric model where $P_0$ is estimated with $\hat{P}_n$ (e.g., a maximum likelihood estimator), we could often expect $\| L_{\hat{P}_n}^{RS} \|_{p,P_0}$ and $\| M_{\hat{P}_n}^{RS} \|_{p,P_0}$ to be $O_{P_0}\left(n^{-1}\right)$ and $O_{P_0}\left(n^{-1/2}\right)$, respectively, for $p \geq 1$.

Thus, the above theorem suggests that the approximation error may be $O_{P_0}(n^{-3/2})$ in a parametric model under $\mathscr{H}_0$. In some examples, it is reasonable to expect that $L^{RS}_{\hat{P}_n} \equiv 0$ for a large class of distributions $P$. In such cases, the upper bound on $\mathrm{Rem}^{\Psi}_{\hat{P}_n}$ simplifies to $\| M^{RS}_{\hat{P}_n} \|^4_{4,P_0}$ under $\mathscr{H}_0$, which under a parametric model is often $O_{P_0}(n^{-2})$.

To make these results more concrete, we consider the special case where $R_P$, $S_P$, $D^R_P$, and $D^S_P$ are smooth mappings of regression functions under $P$ conditional on the $d$–dimensional covariate $W$ (e.g., as in Example 3 – see Section 5). Suppose that all of these regression functions under $P_0$ are at least $\ell$-times differentiable. In this case, rates of convergence for the remainder terms are well understood for kernel smoothers using kernels of sufficiently high order. In particular, each regression function converges at rate $n^{-\frac{\ell}{2\ell+d}}$ in $L^2(P_0)$. Under $\mathscr{H}_0$, one could rely on $\| M^{RS}_{\hat{P}_n} \|_{2,P_0}$ being $o_{P_0}(n^{-1/3})$ and $\| L^{RS}_{\hat{P}_n} \|_{2,P_0}$ being $o_{P_0}(n^{-2/3})$. If $L^{RS}_{\hat{P}_n}$ is second-order, this would generally require $\ell > d$, which is more stringent than the usual $\ell > d/2$ requirement for standard first-order estimators. If, on the other hand, $L^{RS}_{\hat{P}_n} \equiv 0$, then we require that $\| M^{RS}_{\hat{P}_n} \|^4_{4,P_0}$ is $o_{P_0}(n^{-1})$ under $\mathscr{H}_0$, which corresponds to requiring $\ell$ slightly greater than $d/2$.

## 3. Proposed test: formulation and inference under the null

### 3.1. Formulation of test

We begin by constructing an estimator of $\psi_0$ from which a test can then be devised. Using the fact that $\Psi(P) = P^2\Gamma_P$, as implied by (4), we note that if $\Gamma_0$ were known, the U-statistic $\mathbb{U}_n \Gamma_0$ would be a natural estimator of $\psi_0$, where $\mathbb{U}_n$ denotes the empirical measure that places equal probability mass on each of the $n(n-1)$ points $(O_i, O_j)$ with $i \neq j$. In practice, $\Gamma_0$ is unknown and must be estimated. This leads to the estimator $\psi_n \triangleq \mathbb{U}_n \Gamma_n$, where we write $\Gamma_n \triangleq \Gamma_{\hat{P}_n}$ for some estimator $\hat{P}_n$ of $P_0$ based on the available data. Since a large value of $\psi_n$ is inconsistent with $\mathscr{H}_0$, we will reject $\mathscr{H}_0$ if and only if $\psi_n > c_n$ for some appropriately chosen cutoff $c_n$.

In the nonparametric model considered, it may be necessary, or at the very least desirable, to utilize a data-adaptive estimator $\hat{P}_n$ of $P_0$ when constructing $\Gamma_n$. Studying the large-sample properties of $\psi_n$ may then seem particularly daunting since at first glance we may be led to believe that the behavior of $\psi_n - \psi_0$ is dominated by $P_0^2(\Gamma_n - \Gamma_0)$. However, this is not the case. As we will see, under some conditions, $\psi_n - \psi_0$ will approximately behave like $(\mathbb{U}_n - P_0^2)\Gamma_0$. Thus, there will be no contribution of $\hat{P}_n$ to the asymptotic behavior of $\psi_n -$

$\psi_0$. Though this result may seem counterintuitive, it arises because $\Psi(P)$ can be expressed as $P^2\Gamma_P$ with $\Gamma_P$ a second-order gradient (or rather an extension thereof) up to a proportionality constant. More concretely, this surprising finding is a direct consequence of (4).

As further support that $\psi_n$ is a natural test statistic, even when a data-adaptive estimator $\hat{P}_n$ of $P_0$ has been used, we note that $\psi_n$ could also have been derived using a second-order one-step Newton-Raphson construction, as described in Robins et al. [2008]. The latter is given by

$$\psi_{n,NR} \triangleq \Psi(\hat{P}_n) + P_n D_1^{\Psi}(\hat{P}_n) + \tfrac{1}{2}\mathbb{U}_n D_2^{\Psi}(\hat{P}_n),$$

where we use the centered extension of $D_2^{\Psi}$ as discussed in Section 2.3. Here and throughout, $P_n$ denotes the empirical distribution. It is straightforward to verify that indeed $\psi_n = \psi_{n,NR}$.

## 3.2. Inference under the null

**3.2.1. Asymptotic behavior—**For each $P \in \mathcal{M}$, we let $\widetilde{\Gamma}_P$ be the $P_0$−centered modification of $\Gamma_P$ given by

$$\widetilde{\Gamma}_P(o_1, o_2) \triangleq \Gamma_P(o_1, o_2) - \int \Gamma_P(o_1, o) dP_0(o) - \int \Gamma_P(o, o_2) dP_0(o) + P_0^2 \Gamma_P$$

and denote $\widetilde{\Gamma}_{P_0}$ by $\widetilde{\Gamma}_0$. While $\widetilde{\Gamma}_0 = \Gamma_0$ under $\mathcal{H}_0$, this is not true more generally. Below, we use $\mathrm{Rem}_n^{\Psi}$ and $\widetilde{\Gamma}_n$ to respectively denote $\mathrm{Rem}_P^{\Psi}$ and $\widetilde{\Gamma}_P$ evaluated at $P = \hat{P}_n$. Straightforward algebraic manipulations allows us to write

$$\begin{aligned} \psi_n - \psi_0 &= \mathbb{U}_n \Gamma_n - \psi_0 = \mathbb{U}_n \Gamma_n - P_0^2 \Gamma_n + P_0^2 \Gamma_n - \psi_0 \\ &= (\mathbb{U}_n - P_0^2)\Gamma_n + \mathrm{Rem}_n^{\Psi} \\ &= \mathbb{U}_n \Gamma_0 + 2(P_n - P_0)P_0 \Gamma_n + \mathbb{U}_n(\widetilde{\Gamma}_n - \Gamma_0) + \mathrm{Rem}_n^{\Psi}. \end{aligned} \tag{5}$$

Our objective is to show that $n(\psi_n - \psi_0)$ behaves like $n\mathbb{U}_n \Gamma_0$ as $n$ gets large under $\mathcal{H}_0$. In view of (5), this will be true, for example, under conditions ensuring that

**C1)** $n(P_n - P_0)P_0 \Gamma_n = o_{P_0}(1)$ (empirical process and consistency conditions);

**C2)** $n\mathbb{U}_n(\widetilde{\Gamma}_n - \Gamma_0) = o_{P_0}(1)$ ($U$−process and consistency conditions);

**C3)** $n\mathrm{Rem}_n^{\Psi} = o_{P_0}(1)$ (consistency and rate conditions).

We have already argued that C3) is reasonable in many examples of interest, including those presented in this paper. Nolan and Pollard [1987, 1988] developed a formal theory that controls terms of the type appearing in C2). In Supplementary Appendix B.1 we restate

specific results from these authors which are useful to study C2). Finally, the following lemma gives sufficient conditions under which C1) holds. We first set $K_{1n} \triangleq \| L^{RS}_{\hat{P}_n} \|_{1,P_0} + \| M^{RS}_{\hat{P}_n} \|^2_{2,P_0}$.

LEMMA 1 (SUFFICIENT CONDITIONS FOR C1)). *Suppose that* $o_1 \mapsto \int \Gamma_n(o_1, o) dP_0(o)/K_{1n}$, *defined to be zero if* $K_{1n} = 0$, *belongs to a* $P_0-$*Donsker class* [van der Vaart and Wellner, 1996] *with probability tending to* 1. *Then, under* $\mathcal{H}_0$,

$$(P_n - P_0) P_0 \Gamma_n = o_{P_0}\left(\frac{K_{1n}}{\sqrt{n}}\right)$$

*and thus C1) holds whenever* $K_{1n} = o_{P_0}(n^{-1/2})$.

The following theorem describes the asymptotic distribution of $n\psi_n$ under the null hypothesis whenever conditions C1), C2) and C3) are satisfied.

THEOREM 4 (ASYMPTOTIC DISTRIBUTION UNDER $\mathcal{H}_0$). *Suppose that C1), C2) and C3) hold. Then, under* $\mathcal{H}_0$,

$$n\psi_n = n\mathbb{U}_n \Gamma_0 + o_{P_0}(1) \rightsquigarrow \sum_{k=1}^{\infty} \lambda_k \left(Z_k^2 - 1\right),$$

*where* $\{\lambda_k\}_{k=1}^{\infty}$ *are the eigenvalues of the integral operator* $h(o) \mapsto \int \Gamma_0(o_1, o) h(o_1) dP_0(o_1)$ *repeated according to their multiplicity, and* $\{Z_k\}_{k=1}^{\infty}$ *is a sequence of independent standard normal random variables. Furthermore, all of these eigenvalues are nonnegative under* $\mathcal{H}_0$.

We note that by employing a sample splitting procedure – namely, estimating $\Gamma_0$ on one portion of the sample and constructing the $U-$statistic based on the remainder of the sample – it is possible to eliminate the $U-$process conditions required for C2). In such a case, satisfaction of C2) only requires convergence of $\tilde{\Gamma}_n$ to $\Gamma_0$ with respect to the $L^2(P_0^2)$-norm. This sample splitting procedure would also allow one to avoid the empirical process conditions in C1): in particular, $o \mapsto P_0\Gamma_n(o, \cdot)$ would need to converge to zero, but no further requirements would then be needed on $\Gamma_n$ for C1) to be satisfied. We also note that the $L^2(P_0)$ consistency of $o \mapsto P_0\Gamma_n(o, \cdot)$ and the $L^2(P_0^2)$ consistency of $\tilde{\Gamma}_n$ are implied by the $L^2(P_0^2)$ consistency of $\Gamma_n$ for $\Gamma_0$, and so when sample splitting is used one could replace C1) and C2) by this single consistency condition.

We note also that, if sample splitting is not used, then one could replace C1) and C2) by this single consistency condition *and* the added assumption that $\Gamma_n$ belongs to a class with a finite uniform entropy integral. See Supplementary Appendix B.2 for a proof that this suffices to imply the needed empirical process conditions for C1). It is also straightforward

to show that controlling the entropy of the class to which $\Gamma_n$ may belong also controls the entropy of the class to which the linear transformation $\widetilde{\Gamma}_n$ of $\Gamma_n$ may belong.

REMARK 3. *In Example 4, sample splitting may prove particularly important when the estimator of* $E_{P_0}(Y \mid W = w)$ *is chosen as the minimizer of an empirical risk since in finite samples the bias induced by using the same residuals* $y - E_{\hat{P}_n}(Y \mid W = w)$ *as those in the definition of* $D^R_{\hat{P}_n}(o)$ *may be significant. Thus, without some form of sample splitting, the finite sample performance of* $\psi_n$ *may be poor even under the conditions stated in* Supplementary Appendix B.1. □

**3.2.2. Estimation of the test cutoff—**As indicated above, our test consists of rejecting $\mathscr{H}_0$ if and only if $\psi_n$ is larger than some cutoff $c_n$. We wish to select $c_n$ to yield a non-conservative test at level $\alpha \in (0,1)$. In view of Theorem 4, denoting by $q_{1-\alpha}$ the $1-\alpha$ quantile of the described limit distribution, the cutoff $c_n$ should be chosen to be $q_{1-\alpha}/n$. We thus reject $\mathscr{H}_0$ if and only if $n\psi_n > q_{1-\alpha}$. As described in the following corollary, $q_{1-\alpha}$ admits a very simple form when $S_P \equiv 0$ for all $P$.

COROLLARY 2 (ASYMPTOTIC DISTRIBUTION UNDER $\mathscr{H}_0$, $S$ DEGENERATE). *Suppose that C1), C2) and C3) hold, that* $S_P \equiv 0$ *for all* $P \in \mathscr{M}$, *and that* $\sigma_R^2 \triangleq Var_{P_0}[D_0^R(O)] > 0$. *Then, under* $\mathscr{H}_0$,

$$\frac{n\psi_n}{2\sigma_R^2} \rightsquigarrow Z^2 - 1,$$

*where Z is a standard normal random variable. It follows then that* $q_{1-\alpha} = 2\sigma_R^2(z_{1-\alpha/2}^2 - 1)$, *where* $z_{1-\alpha/2}$ *is the* $(1-\alpha/2)$ *quantile of the standard normal distribution.*

The above corollary gives an expression for $q_{1-\alpha}$ that can easily be consistently estimated from the data. In particular, one can use $\hat{q}_{1-\alpha} \triangleq 2(z_{1-\alpha/2}^2 - 1)P_n D^R(\hat{P}_n)^2$ as an estimator of $q_{1-\alpha}$, whose consistency can be established under a Glivenko-Cantelli and consistency condition on the estimator of $D_0^R$. However, in general, such a simple expression will not exist. Gretton et al. [2009] proposed estimating the eigenvalues $\nu_k$ of the centered Gram matrix and then computing $\hat{\lambda}_k \triangleq \nu_k/n$. In our context, the eigenvalues $\nu_k$ are those of the $n \times n$ matrix $G \triangleq \left\{G_{ij}\right\}_{1 \le i,j \le n}$ with entries defined as

$$G_{ij} \triangleq \Gamma_n\left(O_i, O_j\right) - \frac{1}{n}\sum_{k=1}^{n}\Gamma_n\left(O_k, O_j\right) - \frac{1}{n}\sum_{\ell=1}^{n}\Gamma_n\left(O_i, O_\ell\right) + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{\ell=1}^{n}\Gamma_n\left(O_k, O_\ell\right). \quad (6)$$

Given these $n$ eigenvalue estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$, one could then simulate from $\sum_{k=1}^{n} \hat{\lambda}_k (Z_k^2 - 1)$ to approximate $\sum_{k=1}^{\infty} \lambda_k (Z_k^2 - 1)$. While this seems to be a plausible approach, a formal study establishing regularity conditions under which this procedure is valid is beyond the scope of this paper. We note that it also does not fall within the scope of results in Gretton et al. [2009] since their kernel does not depend on estimated nuisance parameters. We refer the reader to Franz [2006] for possible sufficient conditions under which this approach may be valid. Though we do not have formal regularity conditions under which this procedure is guaranteed to maintain the type I error level, our simulation results do seem to suggest appropriate control in practice (Section 6).

In practice, it suffices to give a data-dependent asymptotic upper bound on $q_{1-\alpha}$. We will refer to $\hat{q}_{1-\alpha}^{ub}$, which depends on $P_n$, as an asymptotic upper bound of $q_{1-\alpha}$ if

$$\limsup_{n \to \infty} P_0^n \left( n\psi_n > \hat{q}_{1-\alpha}^{ub} \right) \leq 1 - \alpha . \quad (7)$$

If $q_{1-\alpha}$ is consistently estimated, one possible choice of $\hat{q}_{1-\alpha}^{ub}$ is this estimate of $q_{1-\alpha}$ – the inequality above would also become an equality provided the conclusion of Theorem 4 holds. It is easy to derive a data-dependent upper bound with this property using Chebyshev's inequality. To do so, we first note that

$$Var_{P_0} \left[ \sum_{k=1}^{\infty} \lambda_k \left( Z_k^2 - 1 \right) \right] = \sum_{k=1}^{\infty} \lambda_k^2 Var_{P_0} \left( Z_k^2 \right) = 2 \sum_{k=1}^{\infty} \lambda_k^2 = 2 P_0^2 \Gamma_0^2 ,$$

where we have interchanged the variance operation and the limit using the $L^2$ martingale convergence theorem and the last equality holds because $\lambda_k$, $k = 1, 2, \ldots$, are the eigenvalues of the Hilbert-Schmidt integral operator with kernel $\bar{\Gamma}_0$. Under mild regularity conditions, $P_0^2 \Gamma_0^2$ can be consistently estimated using $\mathbb{U}_n \Gamma_n^2$. Provided $P_0^2 \Gamma_0^2 > 0$, we find that

$$\left( 2 \mathbb{U}_n \Gamma_n^2 \right)^{-1/2} n\psi_n \rightsquigarrow \left( 2 P_0^2 \Gamma_0^2 \right)^{-1/2} \sum_{k=1}^{\infty} \lambda_k \left( Z_k^2 - 1 \right), \quad (8)$$

where the limit variate has mean zero and unit variance. The following theorem gives a valid choice of $\hat{q}_{1-\alpha}^{ub}$.

THEOREM 5. *Fix $\alpha \in (0, 1)$ and suppose that C1), C2) and C3) hold. Then, under $\mathcal{H}_0$ and provided $\mathbb{U}_n \Gamma_n^2 \to P_0^2 \Gamma_0^2 > 0$ in probability, $\hat{q}_{1-\alpha}^{ub} \triangleq \left( 2[1 - \alpha] \mathbb{U}_n \Gamma_n^2 / \alpha \right)^{1/2} \geq q_{1-\alpha}$ is a valid upper bound in the sense of* (7).

The proof of the result follows immediately by noting that $P(X > t) \leq (1 + t^2)^{-1}$ for any random variable $X$ with mean zero and unit variance in view of the one-sided Chebyshev's inequality. For $\alpha = 0.05$, the above demonstrates that a conservative cutoff is $6.2 \cdot (\mathbb{U}_n \Gamma_n^2)^{1/2}$. This theorem illustrates concretely that we can obtain a consistent test that controls type I error. In practice, we recommend either using the result of Corollary 2 whenever possible or estimating the eigenvalues of the matrix in (6).

We note that the condition $\sigma_R^2 > 0$ holds in many but not all examples of interest. Fortunately, the plausibility of this assumption can be evaluated analytically. In Section 5, we show that this condition does not hold in Example 5 and provide a way forward despite this.

## 4. Asymptotic behavior under the alternative

### 4.1. Consistency under a fixed alternative

We present two analyses of the asymptotic behavior of our test under a fixed alternative. The first relies on $\hat{P}_n$ providing a good estimate of $P_0$. Under this condition, we give an interpretable limit distribution that provides insight into the behavior of our estimator under the alternative. As we show, surprisingly, $\hat{P}_n$ need not be close to $P_0$ to obtain an asymptotically consistent test, even if the resulting estimate of $\psi_0$ is nowhere near the truth. In the second analysis, we give more general conditions under which our test will be consistent if $\mathscr{H}_1$ holds.

**4.1.1. Nuisance functions have been estimated well—**As we now establish, our test has power against all alternatives $P_0$ except for the fringe cases discussed in Corollary 1 with $\Gamma_0$ one-degenerate. We first note that

$$\psi_n - \psi_0 = \mathbb{U}_n \Gamma_n - \psi_0 = 2(P_n - P_0) P_0 \Gamma_n + \mathbb{U}_n \tilde{\Gamma}_n + \mathrm{Rem}_P^{\Psi}.$$

When scaled by $\sqrt{n}$, the leading term on the right-hand side follows a mean zero normal distribution under regularity conditions. The second summand is typically $O_{P_0}(n^{-1})$ under certain conditions, for example, on the entropy of the class of plausible realizations of the random function $(o_1, o_2) \mapsto \Gamma_n(o_1, o_2)$ [Nolan and Pollard, 1987, 1988]. In view of the second statement in Theorem 3, the third summand is a second-order term that will often be negligible, even after scaling by $\sqrt{n}$. As such, under certain regularity conditions, the leading term in the representation above determines the asymptotic behavior of $\psi_n$, as described in the following theorem.

THEOREM 6 (ASYMPTOTIC DISTRIBUTION UNDER $\mathscr{H}_1$). *Suppose that* $K_{1n} = o_{P_0}(n^{-1/2})$, *that* $\mathbb{U}_n \tilde{\Gamma}_n = o_{P_0}(n^{-1/2})$, *and furthermore, that* $o \mapsto \int \Gamma_n(o_1, o) dP_0(o)$ *belongs to a fixed*

*$P_0$−Donsker class with probability tending to* 1 *while* $\| P_0(\Gamma_n - \Gamma_0) \|_{2,P_0} = o_{P_0}(1)$. *If* $\mathscr{H}_1$ *holds, we have that* $\sqrt{n}(\psi_n - \psi_0) \rightsquigarrow N(0, \tau^2)$, *where* $\tau^2 \triangleq 4Var_{P_0}\left[\int \Gamma_0(O, o)dP_0(o)\right]$.

In view of the results of Section 2, $\tau^2$ coincides with $\sigma_0^2$, the efficiency bound for regular, asymptotically linear estimators in a nonparametric model. Hence, $\psi_n$ is an asymptotically efficient estimator of $\psi_0$ under $\mathscr{H}_1$. Sufficient conditions for $\int \Gamma_n(o_1, o)dP_0(o)$ to belong to a fixed $P_0$−Donsker class with probability approaching one are given in Supplementary Appendix B.2.

The following corollary is trivial in light of Theorem 6. It establishes that the test $n\psi_n > \hat{q}^{ub}_{1-\alpha}$ is consistent against (essentially) all alternatives provided the needed components of the likelihood are estimated sufficiently well.

COROLLARY 3 (CONSISTENCY UNDER A FIXED ALTERNATIVE). *Suppose the conditions of Theorem 6. Furthermore, suppose that* $\tau^2 > 0$ *and* $\hat{q}^{ub}_{1-\alpha} = o_{P_0}(n)$. *Then, under* $\mathscr{H}_1$, *the test* $n\psi_n > \hat{q}^{ub}_{1-\alpha}$ *is consistent in the sense that*

$$\lim_{n \to \infty} P_0^n\left(n\psi_n > \hat{q}^{ub}_{1-\alpha}\right) = 1.$$

The requirement that $\hat{q}^{ub}_{1-\alpha} = o_{P_0}(n)$ is very mild given that $q_{1-\alpha}$ will be finite whenever $R, S \in \mathcal{S}$. As such, we would not expect $\hat{q}^{ub}_{1-\alpha}$ to get arbitrarily large as sample size grows, at least beyond the extent allowed by our corollary. This suggests that most non-trivial upper bounds satisfying (7) will yield a consistent test.

**4.1.2. Nuisance functions have not been estimated well—**We now consider the case where the nuisance functions are not estimated well, in the sense that the consistency conditions of Theorem 6 do not hold. In particular, we argue that failure of these conditions does not necessarily undermine the consistency of our test. Let $\hat{q}^{ub}_{1-\alpha}$ be the estimated cutoff for our test, and suppose that $\hat{q}^{ub}_{1-\alpha} = o_{P_0}(n)$. Suppose also that $P_0^2\,\Gamma_n$ is asymptotically bounded away from zero in the sense that, for some $\delta > 0$, $P_0^n\left(P_0^2\,\Gamma_n > \delta\right)$ tends to one. This condition is reasonable given that $P_0^2\,\Gamma_0 > 0$ if $\mathscr{H}_1$ holds and $\hat{P}_n$ is nevertheless a (possibly inconsistent) estimator of $P_0$. Assuming that $(\mathbb{U}_n - P_0^2)\,\Gamma_n = O_{P_0}(n^{-1/2})$, which is true under entropy conditions on $\Gamma_n$ [Nolan and Pollard, 1987, 1988], we have that

$$P_0^n\left(n\psi_n > \hat{q}^{ub}_{1-\alpha}\right) = P_0^n\left(\sqrt{n}[\mathbb{U}_n - P_0^2]\,\Gamma_n > \frac{\hat{q}^{ub}_{1-\alpha}}{\sqrt{n}} - \sqrt{n}P_0^2\,\Gamma_n\right) \to 1.$$

We have accounted for the random $n^{-1/2}\hat{q}^{ub}_{1-\alpha}$ term as in the proof of Corollary 3. Of course, this result is less satisfying than Theorem 6, which provides a concrete limit distribution.

### 4.2. Consistency under a local alternative

We consider local alternatives of the form

$$dQ_n(o) = \left[1 + n^{-1/2}h_n(o)\right]dP_0(o),$$

where $h_n \to h$ in $\Gamma_0^2(P_0)$ for some non-degenerate $h$ and $P_0$ satisfies the null hypothesis $\mathscr{H}_0$. Suppose that the conditions of Theorem 4 hold. By Theorem 2.1 of Gregory [1977], we have that

$$n\mathbb{U}_n\,\Gamma_0 \overset{Q_n}{\rightsquigarrow} \sum_{k=1}^{\infty} \lambda_k\left[(Z_k + \langle f_k, h\rangle)^2 - 1\right],$$

where $\mathbb{U}_n$ is the $U$–statistic empirical measure from a sample of size $n$ drawn from $Q_n$, $\langle\cdot,\cdot\rangle$ is the inner product in $L^2(P_0)$, $Z_k$ and $\lambda_k$ are as in Theorem 4, and $f_k$ is a normalized eigenfunction corresponding to eigenvalue $\lambda_k$ described in Theorem 4. By the contiguity of $Q_n$, the conditions of Theorem 4 yield that the result above also holds with $\mathbb{U}_n\,\Gamma_0$ replaced by $\mathbb{U}_n\,\Gamma_n$, our estimator applied to a sample of size $n$ drawn from $Q_n$.

If each $\lambda_k$ is non-negative, the limiting distribution under $Q_n$ stochastically dominates the asymptotic distribution under $P_0$, and furthermore, if $\langle f_k, h\rangle \neq 0$ for some $k$ with $\lambda_k > 0$, this dominance is strict. It is straightforward to show that, under the conditions of Theorem 4, the above holds if and only if $\liminf_n \sqrt{n}\,\Psi(Q_n) > 0$, that is, if the sequence of alternatives is not too hard. Suppose that $\hat{q}_{1-\alpha}$, is a consistent estimate of $q_{1-\alpha}$. By Le Cam's third lemma, $\hat{q}_{1-\alpha}$, is consistent for $q_{1-\alpha}$ even when the estimator is computed on samples of size $n$ drawn from $Q_n$ rather than $P_0$. This proves the following theorem.

THEOREM 7 (CONSISTENCY UNDER A LOCAL ALTERNATIVE). *Suppose that the conditions of Theorem 4 hold. Then, under* $\mathscr{H}_0$ *and provided* $\lim\liminf_{n\to\infty}\sqrt{n}\,\Psi(Q_n) > 0$, *the proposed test is locally consistent in the sense that* $\lim_{n\to\infty} Q_n(n\psi_n > \hat{q}_{1-\alpha}) > \alpha$, *where* $\hat{q}_{1-\alpha}$ *is a consistent estimator of* $q_{1-\alpha}$.

## 5. Illustrations

We now return to Examples 2, 3, 4, and 5. We do not return to Example 1 because it has already been well-studied, e.g. the fixed sample size variant was studied in detail in Gretton et al. [2006]. We first show that Examples 2, 3 and 4 satisfy the regularity conditions described in Section 2. Specifically, we show that all involved parameters $R$ and $S$ belong to $\mathscr{S}$ under reasonable conditions. Furthermore, we determine explicit remainder terms for the

asymptotic representation used in each example and describe conditions under which these remainder terms are negligible. For any $T \in \mathcal{S}$, we will use the shorthand notation $\dot{T}_{\tilde{t}}(x^T) \triangleq \frac{d}{dt} T_{P_t}(x^T)\Big|_{t=\tilde{t}}$ for $\tilde{t}$ in a neighborhood of zero.

Example 2 (Continued).

The parameter $S$ with $S_P \equiv 0$ belongs to $\mathcal{S}$ trivially, with $D_P^S \equiv 0$. Condition (S1) holds with $x^R(o) = w$. Condition (S2) holds using that $R_t(w)$ equals

$$\sum_{a=0}^{1} (-1)^{a+1} \int y \left\{ \frac{1 + th_1(w, a, y) + t^2 h_2(w, a, y)}{1 + tE_{P_0}[h_1(w, A, Y)] + t^2 E_{P_0}[h_2(w, A, Y)]} \right\} dP_0(y \mid a, w). \quad (9)$$

Since we must only consider $h_1$ and $h_2$ uniformly bounded, for $t$ sufficiently small, we see that $R_t(w)$ is twice continuously differentiable with uniformly bounded derivatives. Condition (S3) is satisfied by

$$D_P^R(o) \triangleq \frac{2a - 1}{P(A = a \mid W = w)} \{y - E_P[Y \mid A = a, W = w]\}$$

and $D_P^S \equiv 0$. If $\min_a P(A = a \mid W)$ is bounded away from zero with probability 1 uniformly in $P$, it follows that $(P, o) \mapsto D_P^R(o)$ is uniformly bounded.

Clearly, we have that $\mathrm{Rem}_P^S \equiv 0$. We can also verify that $\mathrm{Rem}_P^R(o)$ equals

$$\sum_{\tilde{a}=0}^{1} (-1)^{\tilde{a}} E_{P_0} \left\{ \left[ 1 - \frac{P_0(A = \tilde{a} \mid W)}{P(A = \tilde{a} \mid W)} \right] \times [E_P(Y \mid A, W) - E_{P_0}(Y \mid A, W)] \mid A = \tilde{a}, W = w \right\}.$$

The above remainder is double robust in the sense that it is zero if either the treatment mechanism (i.e., the probability of $A$ given $W$) or the outcome regression (i.e., the expected value of $Y$ given $A$ and $W$) is correctly specified under $P$. In a randomized trial where the treatment mechanism is known and specified correctly in $P$, we have that $\mathrm{Rem}_P^R \equiv 0$ and thus $L_P^{RS} \equiv 0$. More generally, an upper bound for $\mathrm{Rem}_P^R$ can be found using the Cauchy-Schwarz inequality to relate the rate of $\| \mathrm{Rem}_P^R \|_{2, P_0}$ to the product of the $L^2(P_0)$-norm for the difference between each of the treatment mechanism and the outcome regression under $P$ and $P_0$.

Example 3 (Continued).

For (S1) we take $x^R = x^S = w$. Condition (S2) can be verified using an expression similar to that in (9). Condition (S3) is satisfied by

$$D_P^R(o) \triangleq \frac{a}{P(A = a \mid W = w)}[y - E_P(Y \mid A = a, W = w)]$$
$$D_P^S(o) \triangleq \frac{1 - a}{P(A = a \mid W = w)}[y - E_P(Y \mid A = a, W = w)].$$

If $\min_a \mathrm{P}(A = a \mid W)$ is bounded away from zero with probability 1 uniformly in $P$, both $(P, o) \mapsto D_P^R(o)$ and $(P, o) \mapsto D_P^S(o)$ are uniformly bounded.

Similarly to Example 2, we have that $\mathrm{Rem}_P^R(o)$ is equal to

$$E_{P_0}\left\{\left[1 - \frac{P_0(A = 1 \mid W)}{P(A = 1 \mid W)}\right][E_P(Y \mid A, W) - E_{P_0}(Y \mid A, W)] \mid A = 1, W = w\right\}.$$

The remainder $\mathrm{Rem}_P^S(o)$ is equal to the above display but with $A = 1$ replaced by $A = 0$. The discussion about the double robust remainder term from Example 2 applies to these remainders as well.

Example 4 (Continued).

The parameter $S$ is the same as in Example 2. The parameter $R$ satisfies (S1) with $x^R(o) = w$ and (S2) by an identity analogous to that used in Example 2. Condition (S3) is satisfied by $D_P^R(o) \triangleq y - E_P(Y \mid W = w)$. By the bounds on $Y$, $(P, o) \mapsto D_P^R(o)$ is uniformly bounded. Here, the remainder terms are both exactly zero: $\mathrm{Rem}_P^R \equiv \mathrm{Rem}_P^S \equiv 0$. Thus, we have that $L_P^{RS} \equiv 0$ in this example.

The requirement that $Var_{P_0}\left[D_0^R(O)\right] > 0$ in Corollary 2, and more generally that there exist a nonzero eigenvalue $\lambda_j$ for the limit distribution in Theorem 4 to be nondegenerate, may at times present an obstacle to our goal of obtaining asymptotic control of the type I error. This is the case for Example 5, which we now discuss further. Nevertheless, we show that with a little finesse the type I error can still be controlled at the desired level for the given test. In fact, the test we discuss has type I error converging to zero, suggesting it may be noticeably conservative in small to moderate samples.

Example 5 (Continued).

In this example, one can take $x^R = w$ and $x^S = w(-k)$. Furthermore, it is easy to show that

$$D_P^R(o) = Y - E_P[Y \mid W = w]$$
$$D_P^S(o) = Y - E_P[Y \mid W(-k) = w(-k)].$$

The first-order approximations for $R$ and $S$ are exact in this example as the remainder terms $\mathrm{Rem}_P^R$ and $\mathrm{Rem}_P^S$ are both zero. However, we note that if $E_P(Y \mid W) = E_P(Y \mid W(-k))$ almost surely, it follows that $D_P^R \equiv D_P^S$. This implies that $\Gamma_0 \equiv 0$ almost surely under $\mathscr{H}_0$. As such, under the conditions of Theorem 4, all of the eigenvalues in the limit distribution of $n\psi_n$ in Theorem 4 are zero and $n\psi_n \to 0$ in probability. We are then no longer able to control the type I error at level $\alpha$, rendering our proposed test invalid.

Nevertheless, there is a simple albeit unconventional way to repair this example. Let $A$ be a Bernoulli random variable, independent of all other variables, with fixed probability of success $p \in (0,1)$. Replace $S_P$ with $o \mapsto E_P(Y \mid A = 1, W(-k) = w(-k))$ from Example 3, yielding then

$$D_P^S(o) = \frac{a}{p}[y - E_P(Y \mid A, W(-k) = w(-k))].$$

It then follows that $D_0^R \not\equiv D_0^S$ and in particular $\Gamma_0$ is no longer constant. In this case, the limit distribution given in Theorem 4 is non-degenerate. Consistent estimation of $q_{1-\alpha}$ thus yields a test that asymptotically controls type I error. Given that the proposed estimator $\psi_n$ converges to zero faster than $n^{-1}$, the probability of rejecting the null approaches zero as sample size grows. In principle, we could have chosen any positive cutoff given that $n\psi_n \to 0$ in probability, but choosing a more principled cutoff seems judicious.

Because $p$ is known, the remainder term $\mathrm{Rem}_P^S$ is equal to zero. Furthermore, in view of the independence between $A$ and all other variables, one can estimate $E_{P_0}(Y \mid A = 0, W(-k))$ by regressing $Y$ on $W(-k)$ using all of the data without including the covariate $A$.

In future work, it may also be worth checking to see if the parameter is third-order differentiable under the null, and if so whether or not this allows us to construct an $\alpha$−level test without resorting to an artificial source of randomness.

## 6. Simulation studies

In simulation studies, we have explored the performance of our proposed test in the context of Examples 2, 3 and 4, and have also compared our method to the approach of Racine et al. [2006] for which software is readily available – see, e.g., the R package np [Hayfield and Racine, 2008]. We evaluate the performance of computing the eigenvalues of the Gram matrix defined in (6) for Example 3 in two different scenarios. We report the results of our simulation studies in this section.

In all simulation settings, we consider an adaptive bandwidth selection procedure that is a variant of the median heuristic that has been employed in the classical MMD setting where $P \mapsto R_P$ and $P \mapsto S_P$ do not depend on $P$ [Gretton et al., 2012a]. In that case, the median heuristic selects the bandwidth to be equal to the median of the $2n \times 2n$ Euclidean distance matrix of $\{R(O_i) : i = 1, \ldots, n\} \cup \{S(O_i) : i = 1, \ldots, n\}$, where the subscript of $R$ and $S$ on a

distribution $P$ has been omitted to emphasize the lack of this dependence in the classical MMD setting. In our case, we choose the bandwidth to be equal to the median of the Euclidean distance matrix between scalar or vector-valued observations (see Concluding Remark b in Section 7 for the extension to vector-valued unknown functions) in

$$\{R_{\hat{P}_n}(O_i) + D^R_{\hat{P}_n}(O_i) : i = 1, \ldots, n\} \cup \{S_{\hat{P}_n}(O_i) + D^S_{\hat{P}_n}(O_i) : i = 1, \ldots, n\}.$$

This extension is natural in that $R_{\hat{P}_n} + D^R_{\hat{P}_n}$ and $S_{\hat{P}_n} + D^S_{\hat{P}_n}$ are reminiscent of one-step estimators [Pfanzagl, 1982] of the unknown $R_0$ and $S_0$, which should help this procedure account for the uncertainty in $R_{\hat{P}_n}$ and $S_{\hat{P}_n}$. Except where specified, every MMD result presented in this section uses thins mediann heuristic to select the bandwidth. We also compare this procedure to a fixed choice of bandwidth in two of our settings.

### 6.1. Simulation scenario 1

We use an observed data structure $(W, A, Y)$, where $W \triangleq (W_1, W_2, \ldots, W_5)$ is drawn from a standard 5-dimensional normal distribution, $A$ is drawn according to a Bernoulli(0.5) distribution, and $Y = \mu(A, W) + 5\xi(A, W)$, where the different forms of the conditional mean function $\mu(a, w)$ are given in Table 1, and $\xi(a, w)$ is a random variate following a Beta distribution with shape parameters $\alpha = 3\text{expit}(aw_2)$ and $\beta = 2\text{expit}[(1 - a)w_1]$ shifted to have mean zero, where $\text{expit}(x) = 1/(1 + \exp(-x))$.

We performed tests of the null in which $\mu(1, W)$ is equal to $\mu(0, W)$ almost surely and in distribution, as presented in Examples 2 and 3, respectively. Our estimate $\hat{P}_n$ of $P_0$ was constructed using the knowledge that $P_0(A = 1 \mid W) = 1/2$, as would be available, for example, in the context of a randomized trial. The conditional mean function $\mu(a, w)$ was estimated using the ensemble learning algorithm Super Learner [van der Laan et al., 2007], as implemented in the SuperLearner package [Polley and van der Laan, 2013]. This algorithm was implemented using 10-fold cross-validation to determine the best convex combination of regression function candidates minimizing mean-squared error using a candidate library consisting of SL.rpart, SL.glm.interaction, SL.glm, SL.earth, and SL.nnet. We used the results of Corollary 2 to evaluate significance for Example 2, and the eigenvalue approach presented in Section 3.2.2 to evaluate significance for Example 3, where we used all of the positive eigenvalues for $n = 125$ and the largest 200 positive eigenvalues for $n > 125$ using the rARPACK package [Qiu et al., 2014].

To evaluate the performance of the adaptive bandwidth selection procedure in the context of a test of the equality in distribution of two unknown functions applied to $w$, namely $\mu(1, \cdot)$ and $\mu(0, \cdot)$, we also ran our procedure at fixed bandwidths with values $2^k$, $k = -2, 1, 0, 1, 2$. In the context of a test of the almost sure equality of $\mu(1, W)$ and $\mu(0, W)$, we compare our adaptive bandwidth selection procedure to fixing the bandwidth at one. The performance of the adaptive bandwidth selection procedure is evaluated in more detail for a null hypothesis of the almost sure equality of two unknown functions in our third simulation setting.

We ran 1,000 Monte Carlo simulations with samples of size 125, 250, 500, 1000, and 2000, except for the np package, which we only ran for 500 Monte Carlo simulations. For Example 2 we compared our approach with that of Racine et al. [2006] using the npsigtest function from the np package. This requires first selecting a bandwidth, which we did using the npregbw function, specifying that we wanted a local linear estimator and the bandwidth to be selected using the cv.aic method [Hayfield and Racine, 2008].

Figure 1 displays the empirical null rejection probability of our test of equality in distribution of $\mu(1, W)$ and $\mu(0, W)$ for simulation scenarios 1a, 1b and 1c. In particular, we observe that our method is able to properly control type I error for Simulation 1a when testing the hypothesis that $\mu(1, W)$ is equal in distribution to $\mu(0, W)$. Type I error is also properly controlled in Simulation 1b, though the control of the fixed bandwidth procedures appears to be conservative at the larger sample sizes. We also note that the adaptive bandwidth yielded similar performance to the best considered fixed bandwidth of 1. Our selection procedure generally picked values with an average of around 1.5 – at large sample sizes, there was little variability around this average bandwidth, while at smaller sample sizes the selected bandwidths generally fell between 1.25 and 1.75. The adaptive procedure always controlled type I error at or near the nominal level and had power increasing with sample size and comparable to that of a fixed bandwidth of 1. Choosing the largest fixed bandwidth, namely 4, yielded no power at the alternative in Simulation 1c. Choosing the smallest fixed bandwidth, namely 1/4, yielded inflated type I error levels at one of the null distributions, namely Simulation 1b.

Figure 2 displays the empirical coverage of our approach as well as that resulting from use of the np package. At smaller sample sizes, our method does not appear to control type I error near the nominal level. This is likely because we use an asymptotic result to compute the cutoff, even when the sample size is small. Nevertheless, as sample size grows, the type I error of our test approaches the nominal level. We note that choosing the fixed unit bandwidth outperforms the median heuristic bandwidth selection procedure in this setting, especially in terms of power Simulation 1b. We note that in Racine et al. [2006], unlike in our proposal, the bootstrap was used to evaluate the significance of the proposed test. It will be interesting to see if applying a bootstrap procedure at smaller sample sizes improves our small-sample results. At larger sample sizes, it appears that the method of Racine et al. outperforms our approach in terms of power in simulation scenarios 1b and 1c. At smaller sample sizes (125, 250, 500), our method achieves higher power than that of Racine et al., but at the expense of double the type I error of that of Racine et al.: therefore, it appears that the method of Racine et al. outperforms our approach in Simulations 1a, 1b, and 1c when testing the null hypothesis that $\mu(1, W) - \mu(0, W)$ is almost surely equal to zero. Nonetheless, we note that the generality of our approach allows us to apply our test in more settings than a test using the method of Racine et al.. For example, we are not aware of any other test devised to test the equality in distribution of $\mu(1, W)$ and $\mu(0, W)$ (Figure 1).

### 6.2. Simulation scenario 2: comparison with Racine et al. [2006]

We reproduced a simulation study from Section 4.1 of Racine et al. [2006] at sample size $n = 100$. In particular, we let $Y = 1 + \beta A(1 + W_2^2) + W_1 + W_2 + \epsilon$, where $A$, $W_1$, and $W_2$ are

drawn independently from Bernoulli(0.5), Bernoulli(0.5), and $N(0,1)$ distributions, respectively. The error term $\epsilon$ is unobserved and drawn from a $N(0,1)$ distribution independently of all observed variables. The parameter $\beta$ was varied over values −0.5, −0.4, …, 0.4, 0.5 to achieve a range of distributions. The goal is to test whether $E_0(Y \mid A, W) = E_0(Y \mid W)$ almost surely, or equivalently, that $\mu(1, W) - \mu(0, W) = 0$ almost surely.

Due to computational constraints, we only ran the 'Bootstrap I test' to evaluate significance of the method of Racine et al. [2006]. As the authors report, this method is anticonservative relative to their 'Bootstrap II test' and indeed achieves lower power (but proper type I error control) in their simulations.

Except for two minor modifications, our implementation of the method in Example 2 is similar to that as for Simulation 1. For a fair comparison with Racine et al. [2006], in this simulation study, we estimated $P_0(A = 1 \mid W)$ rather than treating it as known. We did this using the same Super Learner library and the 'family=binomial' setting to account for the fact that $A$ is binary. We also scaled the function $\mu(1, w) - \mu(0, w)$ by a factor of 5 to ensure most of the probability mass of $R_0$ falls between −1 and 1 (around 99% when $\beta = 0$). We note that even with scaling the variable $Y$ is not bounded as our regularity conditions require. Nonetheless, an evaluation of our method under violations of our assumptions can itself be informative.

Figure 3 displays the empirical null rejection probability of our test as well as that of Racine et al. [2006]. In this setup, used by the authors themselves to showcase their test procedure, our method performs comparably to their proposal, with slightly lower type I error (closer to nominal) and slightly lower power.

### 6.3. Simulation scenario 3: higher dimensions

We also explored the performance of our method as extended to tackle higher-dimensional hypotheses, as discussed in Section 7. To do this, we used the same distribution as for Simulation 1 but with $Y$ now a 20-dimensional random variable. Our objective here was to test $\mu(1, W) - \mu(0, W)$ is equal to (0,0, …, 0) in probability, where $\mu(a, w) \triangleq (\mu_1(a, w), \mu_2(a, w), \ldots, \mu_{20}(a, w))$ with $\mu_j(a, w) \triangleq E_0(Y_j \mid A = a, W = w)$. Conditional on $A$ and $W$, the coordinates of $Y$ are independent. We varied the number of coordinates that represent signal and noise. For signal coordinate $j$, given $A$ and $W$, $20\,Y_j$ was drawn from the same conditional distribution as $Y$ give $A$ and $W$ in Simulation 1c. For noise coordinate $j$, given $A$ and $W$, $20\,Y_j$ was drawn from the same conditional distribution as $Y$ given $A$ and $W$ in Simulation 1a.

Relative to Simulation 1, we have scaled each coordinate of the outcome to be one twentieth the size of the outcome in Simulation 1. Apart from the adaptive bandwidth selection procedure discussed at the beginning of this section, we considered defining the MMD with a Gaussian kernel with bandwidths of 1/4, 1/2, 1, and 2. Alternatively, this could be viewed as considering bandwidths 5, 10, 20, and 40 if the outcome had not been scaled by 1/20.

We ran the same Super Learner to estimate $\mu(1, w)$ as in Simulation 1, and we again treated the probability of treatment given covariates as known. We evaluated significance by

estimating all of the positive eigenvalues of the centered Gram matrix for $n = 125$ and the largest 200 positive eigenvalues of the centered Gram matrix for $n > 125$.

In Figure 4, the empirical null rejection probability is displayed for our proposed MMD method. We did not include the results for sample size 125 in the figure because type I error control was too poor. For example, for zero signal coordinates, the probability of rejection was 0.24 for bandwidth 1 and 0.33 for bandwidth 1/2. The adaptive bandwidth method performs comparably to the procedure that *a priori* fixes the bandwidth at 1/2. This observation is consistent with the fact that, across all signal levels and sample sizes, the selected bandwidth was closely concentrated about 1/2 for all Monte Carlo repetitions: the minimal selected bandwidth was 0.49 and the maximal selected bandwidth was 0.59. Among the considered fixed bandwidths, 1/2 or 1/4 seem to be the best at the largest sample sizes (1000, 2000), with the tradeoff between the two being that a bandwidth of 1/4 increases power (substantially for a signal of 5) at the cost of inflating type I error. At smaller sample sizes, the bandwidth of 1/4 yields unacceptably inflated type I error (0.4 at $n = 250$ and 0.15 at $n = 500$). Our adaptive bandwidth procedure appears to control type I error well at moderate to large sample sizes (i.e., $n \geq 500$). This simulation shows that, overall, our method indeed has increasing power as sample size grows or as the number of coordinates $j$ for which $\mu_j(1, W) - \mu_j(0, W)$ not equal to zero in probability increases. The only sample size and signal number at which our adaptive bandwidth procedure appears to be outperformed by a fixed bandwidth is at 2000: a fixed bandwidth of 1/4 attains nominal coverage at this sample size but dramatically outperforms the adaptive bandwidth when the signal is 5. This discrepancy disappears when the signal is 10.

## 7. Concluding remarks

We have presented a novel approach to test whether two unknown functions are equal in distribution. Our proposal explicitly allows, and indeed encourages, the use of flexible, data-adaptive techniques for estimating these unknown functions as an intermediate step. Our approach is centered upon the notion of maximum mean discrepancy, as introduced in Gretton et al. [2006], since the MMD provides an elegant means of contrasting the distributions of these two unknown quantities. In their original paper, these authors showed that the MMD, which in their context tests whether two probability distributions are equal using $n$ random draws from each distribution, can be estimated using a $U$- or $V$−statistic. Under the null hypothesis, this $U$- or $V$−statistic is degenerate and converges to the true parameter value quickly. Under the alternative, it converges at the standard $n^{-1/2}$ rate. Because this parameter is a mean over a product distribution from which the data were observed, it is not surprising that a $U$- or $V$−statistic yields a good estimate of the MMD. What is surprising is that we were able to construct an estimator with these same rates even when the null hypothesis involves unknown functions that can only be estimated at slower rates. To accomplish this, we used recent developments from the higher-order pathwise differentiability literature. Our simulation studies indicate that our asymptotic results are meaningful in finite samples, and that in specific examples for which other methods exist, our methods generally perform at least as well as these established, tailor-made methods. Of course, the great appeal of our proposal is that it applies to a much wider class of problems.

In our simulation study, we adapted the median heuristic for selecting the Gaussian kernel bandwidth to our setting in which $R_0$ and $S_0$ are unknown. In some settings, this bandwidth selection procedure performed well in our simulation compared to specifying a fixed bandwidth in our simulation, though we did note settings where the adaptive procedure underperformed relative to using a fixed bandwidth. An advantage of this adaptive bandwidth selection procedure compared to selecting a fixed bandwidth (e.g., the unit bandwidth) is that it yields a procedure that is invariant to a rescaling of the unknown functions $R_0$ and $S_0$. For the classical MMD setting in which $R_0$ and $S_0$ are known functions, Gretton et al. [2012b] showed that other bandwidth selection procedures can outperform the median heuristic. One such procedure involves selecting the bandwidth to maximize an estimate of the power of the test of the null hypothesis of equality in distribution of $R_0(O)$ and $S_0(O)$, subject to a constraint on the estimated type I error. Extending these procedures to our setting where $R_0$ and $S_0$ are unknown is an important area for future research.

We conclude with several possible extensions of our method that may increase further its applicability and appeal.

**(a)** Although this condition is satisfied in all but one of our examples, requiring $R$ and $S$ to be in $\mathcal{S}$ can be somewhat restrictive. Nevertheless, it appears that this condition may be weakened by instead requiring membership to $\mathcal{S}^*$, the class of all parameters $T$ for which there exist some $M < \infty$ and elements $T^1, T^2, \ldots, T^M$ in $\mathcal{S}$ such that $T = \sum_{m=1}^{M} T^m$. While the results in our paper can be established in a similar manner for functions in this generalized class, the expressions for the involved gradients are quite a bit more complicated. Specifically, we find that, for $T, U \in \mathcal{S}^*$ with $T = \sum_{m=1}^{M} T^m$ and $U = \sum_{\ell=1}^{L} U^\ell$, the quantity $\Gamma_P^{TU}(o_1, o_2)$ equals

$$e^{-[T_P(o_1) - U_P(o_2)]^2} + \sum_{\ell=1}^{L} E_P\left[2[T_P(o_1) - U_P(O)]e^{-[T_P(o_1) - U_P(O)]^2} \middle| X^{U^\ell} = x_2^{U^\ell}\right] D_P^{U^\ell}(o_2) - \sum_{m=1}^{M} E_P\left[2[T_P(O) - U_P(o_2)]e^{-[T_P(O) - U_P(o_2)]^2} \middle| X^{T^m} = x_1^{T^m}\right] D_P^{T^m}(o_1) - \sum_{\ell=1}^{L}\sum_{m=1}^{M} E_{P^2}\left[\left\{4[T_P(O_1) - U_P(O_2)]^2 - 2\right\}e^{-[T_P(O_1) - U_P(O_2)]^2} \middle| X_1^{T^\ell} = x_1^{T^\ell}, X_2^{U^m} = x_2^{U^m}\right] D_P^{T^\ell}(o_1) D_P^{U^m}(o_2).$$

In particular, we note the need for conditional expectations with respect to $X^{R^m}$ and $X^{S^m}$ in the definition of $\Gamma$, which could render the implementation of our method more difficult. While we believe this extension is promising, its practicality remains to be investigated.

**(b)** While our paper focuses on univariate hypotheses, our results can be generalized to higher dimensions. Suppose that $P \mapsto R_P$ and $P \mapsto S_P$ are $\mathbb{R}^d$-valued functions on $\mathcal{O}$. The class $\mathcal{S}_d$ of allowed such parameters can be defined similarly as $\mathcal{S}$, with all original conditions applying componentwise. The MMD for the vector-valued parameters $R$ and $S$ using the Gaussian kernel is given by $\Psi_d(P) \triangleq \Phi_d^{RR}(P) - 2\Phi_d^{RS}(P) + \Phi_d^{SS}(P)$, where for any $T, U \in \mathcal{S}_d$ we set

$$\Phi_d^{TU}(P) \triangleq \int\int e^{-\left\|T_P(o_1) - U_P(o_2)\right\|^2} dP(o_1) dP(o_2).$$

It is not difficult to show then that, for any $T, U \in \mathcal{S}_d(P_0)$, $\Gamma_{d,P}^{TU}(o_1, o_2)$ is given by

$$\left[2[T_P(o_1) - U_P(o_2)]'[D_P^U(o_2) - D_P^T(o_1)] + 1 - 2D_P^T(o_1)'\left\{2[T_P(o_1) - U_P(o_2)][T_P(o_1) - U_P(o_2)]' - \mathrm{Id}\right\}D_P^U(o_2)\right] \times e^{-\left\|T_P(o_1) - U_P(o_2)\right\|^2},$$

where Id denotes the $d$–dimensional identity matrix and $A'$ denotes the transpose of a given vector $A$. Using these objects, the method and results presented in this paper can be replicated in higher dimensions rather easily.

**(c)** Our results can be used to develop confidence sets for infinite dimensional parameters by test inversion. Consider a parameter $T$ satisfying our conditions. Then one can test if $R_0 \triangleq T_0 - f$ is equal in distribution to zero for any fixed function $f$ that does not rely on $P$. Under the conditions given in this paper, a $1 - \alpha$ confidence set for $T_0$ is given by all functions $f$ for which we do not reject $\mathcal{H}_0$ at level $\alpha$. The blip function from Example 2 is a particularly interesting example, since a confidence set for this parameter can be mapped into a confidence set for the sign of the blip function, i.e. the optimal individualized treatment strategy [Robins, 2004]. We would hope that the omnibus nature of the test implies that the confidence set does not contain functions $f$ that are "far away" from $T_0$, contrary to a test which has no power against certain alternatives. Formalization of this claim is an area of future research.

**(d)** To improve upon our proposal for nonparametrically testing variable importance via the conditional mean function, as discussed in Section 5, it may be fruitful to consider the related Hilbert Schmidt independence criterion [Gretton et al., 2005]. Higher-order pathwise differentiability may prove useful to estimate and infer about this discrepancy measure.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgement

The authors thank Noah Simon for helpful discussions. Alex Luedtke was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. Marco Carone was supported by a Genentech Endowed Professorship at the University of Washington. Mark van der Laan was supported by NIH grant R01 AI074345–06.

## References

Berlinet A and Thomas-Agnan C. Reproducing kernel Hilbert spaces in probability and statistics. Springer Science & Business Media, 2011.

Carone M, Díaz I, and van der Laan MJ. Higher-order Targeted Minimum Loss-based Estimation. Technical report, Division of Biostatistics, University of California, Berkeley, 2014.

Chakraborty B and Moodie EE. Statistical Methods for Dynamic Treatment Regimes. Springer, Berlin Heidelberg New York, 2013.

Franz C. Discrete approximation of integral operators. Proceedings of the American Mathematical Society, 134(8):2437–2446, 2006.

Gregory GG. Large sample theory for U-statistics and tests of fit. The annals of statistics, pages 110–123, 1977.

Gretton A, Bousquet O, Smola A, and Schölkopf B. Measuring statistical dependence with Hilbert-Schmidt norms In Algorithmic learning theory, pages 63–77. Springer, 2005.

Gretton A, Borgwardt MM, Rasch M, Schölkopf B, and Smola AJ. A kernel method for the two-sample-problem. In Advances in neural information processing systems, pages 513–520, 2006.

Gretton A, Fukumizu K, Harchaoui Z, and Sriperumbudur BK. A fast, consistent kernel two-sample test. In Advances in neural information processing systems, pages 673–681, 2009.

Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, and Smola A. A kernel two-sample test. The Journal of Machine Learning Research, 13(1):723–773, 2012a.

Gretton A, Sejdinovic D, Strathmann H, Balakrishnan S, Pontil M, Fukumizu Kenji, and Sriperumbudur BK. Optimal kernel choice for large-scale two-sample tests. In Advances in neural information processing systems, pages 1205–1213, 2012b.

Hayfield T and Racine JS. Nonparametric Econometrics: The np Package. Journal of Statistical Software, 27(5), 2008 URL http://www.jstatsoft.org/v27/i05/.

Lavergne P, Maistre S, and Patilea V. A significance test for covariates in nonparametric regression. Electronic Journal of Statistics, 9:643–678, 2015.

Nolan D and Pollard D. U-processes: rates of convergence. The Annals of Statistics, 15 (2):780–799, 1987.

Nolan D and Pollard D. Functional Limit Theorems for $U$-Processes. The Annals of Probability, 16(3):1291–1298, 1988 ISSN 0091–1798. doi: 10.1214/aop/1176991691.

Pfanzagl J. No Title. Springer, Berlin Heidelberg New York, 1982.

Pfanzagl J. Asymptotic expansions for general statistical models, volume 31 Springer-Verlag, 1985.

Polley E and van der Laan MJ. SuperLearner: super learner prediction, 2013 URL http://cran.r-project.org/package=SuperLearner.

Qiu Y, Mei J, and authors of the ARPACK library. See file AUTHORS for details. rARPACK: R wrapper of ARPACK for large scale eigenvalue/vector problems, on both dense and sparse matrices, 2014 URL http://cran.r-project.org/package=rARPACK.

Racine JS, Hart J, and Li Q. Testing the significance of categorical predictor variables in nonparametric regression models. Econometric Reviews, 25(4):523–544, 2006.

Robins JM. Optimal structural nested models for optimal sequential decisions. In Lin DY and Heagerty P, editors, Proceedings of the Second Seattle Symposium in Biostatistics, volume 179, pages 189–326, 2004.

Robins JM, Li L, Tchetgen E, and van der Vaart AW. Higher order influence functions and minimax estimation of non-linear functionals In Essays in Honor of David A. Freedman, IMS, Collections Probability and Statistics, pages 335–421. Springer New York, 2008.

Sejdinovic D, Sriperumbudur B, Gretton A, and Fukumizu K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. The Annals of Statistics, 41 (5):2263–2291, 2013.

Steinwart I. On the influence of the kernel on the consistency of support vector machines. The Journal of Machine Learning Research, 2:67–93, 2002.

van der Laan MJ and Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer, New York, New York, 2011.

van der Laan MJ, Polley E, and Hubbard A. Super Learner. Stat Appl Genet Mol, 6 (1):Article 25, 2007 ISSN 1.

van der Vaart AW. Higher order tangent spaces and influence functions. Statistical Science, 29(4):679–686, 2014.

van der Vaart AW and Wellner JA. Weak convergence and empirical processes. Springer, Berlin Heidelberg New York, 1996.

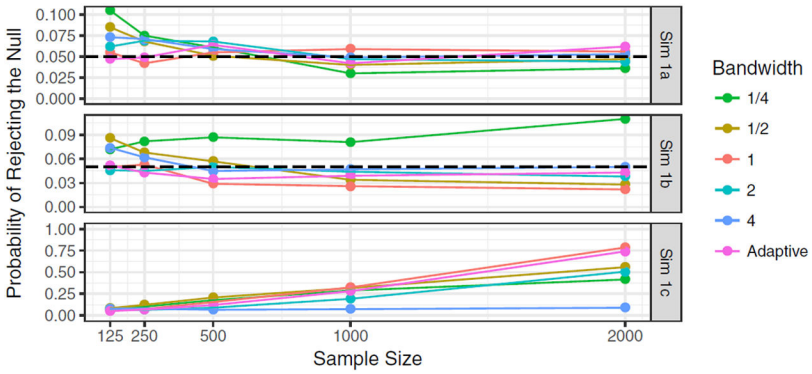


**Fig. 1.**

Empirical probability of rejecting the null when testing the null hypothesis that $\mu(1, W)$ is equal in distribution to $\mu(0, W)$ (Example 3) in Simulation 1. Table 1 indicates that the null is true in Simulations 1a and 1b, and the alternative is true in Simulation 1c.

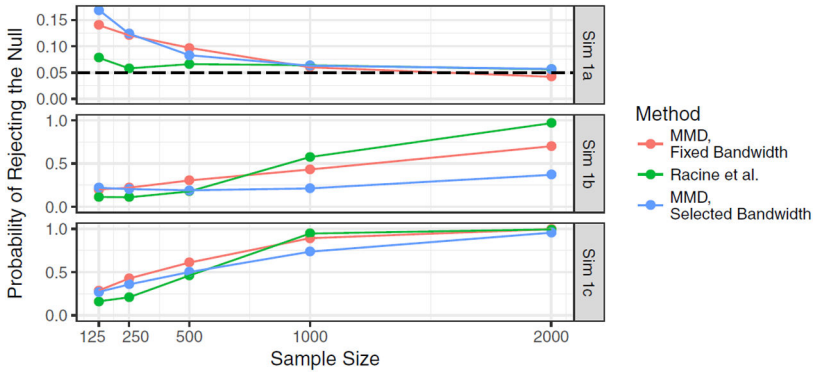


**Fig. 2.**

Empirical probability of rejecting the null when testing the null hypothesis that $\mu(1, W) - \mu(0, W)$ is almost surely equal to zero (Example 2) in Simulation 1. Table 1 indicates that the null is true in Simulation 1a, and the alternative is true in Simulations 1b and 1c.

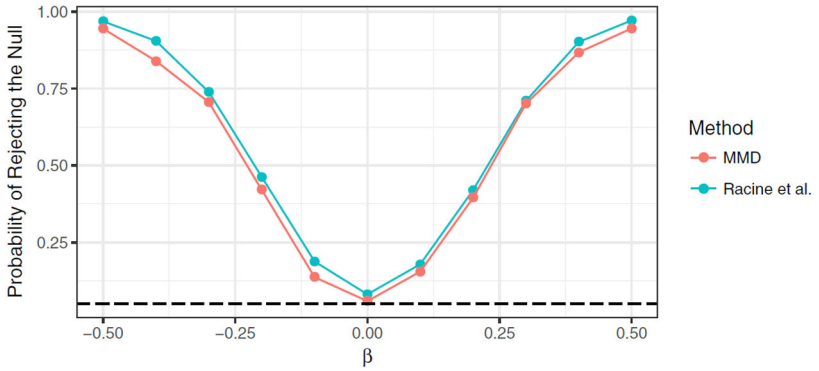


**Fig. 3.**

Empirical probability of rejecting the null when testing the null hypothesis that $\mu(1, W) - \mu(0, W)$ is almost surely equal to zero in Simulation 2.

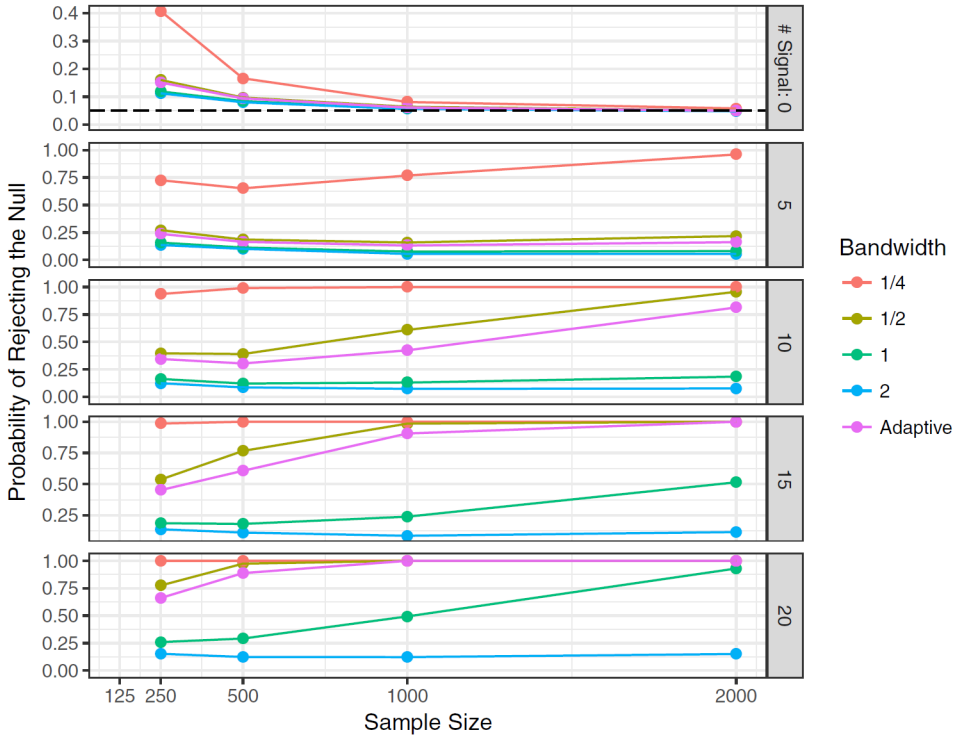


**Fig. 4.**

Probability of rejecting the null when testing the null hypothesis that $\mu(1, W) - \mu(0, W)$ is almost surely equal to zero in Simulation 3.

**Table 1.**

Conditional mean function in each of three simulation settings within simulation scenario 1. Here, $m(a, w) \triangleq 0.2(w_1^2 + w_2 - 2w_3w_4)$, and the third and fourth columns indicate, respectively, whether $\mu(1, W)$ and $\mu(0, W)$ are equal in distribution or almost surely.

| | $\mu(a, w)$ | $\stackrel{d}{=}$ | $\stackrel{a.s.}{=}$ |
|---|---|---|---|
| Simulation 1a | $m(a, w)$ | × | × |
| Simulation 1b | $m(a, w) + 0.4[aw_3 + (1 - a)w_4]$ | × | |
| Simulation 1c | $m(a, w) + 0.8aw_3$ | | |