Biometrika (2021), **103**, 1, *pp*. 1–20
Printed in Great Britain

Kernel Methods for Causal Functions: Dose, Heterogeneous, and Incremental Response Curves

BY R. SINGH

Department of Economics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, U.S.A.

rahul.singh@mit.edu

L. XU

Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, U.K.

liyuan.jo.19@ucl.ac.uk

AND A. GRETTON

Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, U.K.

arthur.gretton@gmail.com

SUMMARY

We propose estimators based on kernel ridge regression for nonparametric causal functions such as dose, heterogeneous, and incremental response curves. The treatment and covariates may be discrete or continuous in general spaces. Due to a decomposition property specific to the reproducing kernel Hilbert space, our estimators have simple closed form solutions. We prove uniform consistency with finite sample rates via an original analysis of generalized kernel ridge regression. We extend our main results to counterfactual distributions and to causal functions identified by front and back door criteria. We achieve state-of-the-art performance in nonlinear simulations with many covariates, and conduct a policy evaluation of the US Job Corps training program for disadvantaged youths.

Some key words: Reproducing Kernel Hilbert Space; Continuous Treatment; Uniform Consistency.

1. INTRODUCTION

Program evaluation aims to measure the counterfactual relationship between the treatment D and the outcome Y, which may vary for different subpopulations: if we intervened on the treatment, setting $D = d$, what would be the expected counterfactual outcome $Y^{(d)}$ for individuals with characteristics $V = v$? When the treatment is binary, the causal parameter is a function $\theta_0(v) = E\{Y^{(1)} - Y^{(0)} \mid V = v\}$ called the heterogeneous treatment effect; when the treatment is continuous, it is a function $\theta_0(d, v) = E\{Y^{(d)} \mid V = v\}$ that we call the heterogeneous response curve. Assuming selection on the observable covariates (V, X), the causal function $\theta_0(d, v)$ can be recovered by integrating the regression function $\gamma_0(d, v, x) = E(Y \mid D = d, V = v, X = x)$ according to the conditional distribution $P(x \mid v)$: $\theta_0(d, v) = \int \gamma_0(d, v, x) \mathrm{d}P(x \mid v)$ (Rosenbaum & Rubin, 1983; Robins, 1986), which may be complex when there are many covariates.

The same is true for other causal functions such as dose and incremental response curves, and even counterfactual distributions, albeit with different regressions and reweightings. Therefore nonparametric estimation of a causal function involves three challenging steps: estimating a nonlinear regression, with possibly many covariates; estimating the distribution for reweighting, which may be conditional; and using the nonparametric distribution to integrate the nonparametric regression. For this reason, flexible estimation of nonparametric causal functions, such as $\theta_0(d, v)$, is often deemed too computationally demanding to be practical for program evaluation.

Our key insight is that the definition of the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ resolves technical and practical issues that arise when estimating nonparametric causal functions with a continuous treatment. Suppose the treatment is continuous, fix the values (d, v), and define the functional $F : \gamma \mapsto \int \gamma(d, v, x) \mathrm{d}P(x \mid v)$ so that the heterogeneous response curve evaluated at (d, v) is $\theta_0(d, v) = F(\gamma_0)$. When F is defined over $\mathbb{L}^2$, it not bounded; there does not exist a constant $C < \infty$ such that $F(\gamma) \leq C\|\gamma\|_{\mathbb{L}^2}$ for all $\gamma \in \mathbb{L}^2$ (van der Vaart, 1991; Newey, 1994a). We show that when F is defined over $\mathcal{H} \subset \mathbb{L}^2$, it is bounded; there exists a constant $C < \infty$ such that $F(\gamma) \leq C\|\gamma\|_{\mathcal{H}}$ for all $\gamma \in \mathcal{H}$ by the RKHS definition. Indeed, the RKHS is defined as the Hilbert space of functions for which the functional $\gamma \mapsto \gamma(d, v, x)$ is bounded (Berlinet & Thomas-Agnan, 2004), so under weak regularity conditions, the functional $\gamma \mapsto \int \gamma(d, v, x) \mathrm{d}P(x \mid v)$ is bounded over the RKHS as well. Supplement 6 provides a related discussion of pathwise differentiability (Bickel et al., 1993, ch. 3 and 5).

This insight has practical consequences. By the Riesz representation theorem, since F is a bounded linear functional over a Hilbert space, it admits an inner product representation within that Hilbert space: there exists some $\tilde{\alpha}_0 \in \mathcal{H}$ such that $F(\gamma) = \langle \gamma, \tilde{\alpha}_0 \rangle_{\mathcal{H}}$ for all $\gamma \in \mathcal{H}$. In particular, $\theta_0(d, v) = \langle \gamma_0, \tilde{\alpha}_0 \rangle_{\mathcal{H}}$. The Riesz representation separates the steps of nonparametric causal estimation in the RKHS into three simple steps: estimating the regression γ_0; estimating $\tilde{\alpha}_0$, which turns out to be a generalized regression that embeds $P(x \mid v)$; and taking their product. This decomposition is a specific strength of our framework, since it follows from the RKHS definition, and we use it to derive simple estimators.

Algorithmically, we adapt kernel ridge regression, a classic machine learning algorithm that generalizes splines (Wahba, 1990), to estimate causal functions such as dose, heterogeneous, and incremental response curves. Based on our key insight, we propose nonparametric estimators that are inner products of kernel ridge regressions, which therefore have closed form solutions unlike previous work. They are substantially simpler yet outperform some leading alternatives in nonlinear simulations with many covariates; see Section 8. As extensions, we generalize our new algorithmic techniques to counterfactual distributions, and to causal functions and counterfactual distributions identified by front and back door criteria; see the Supplementary Material.

Statistically, we prove uniform consistency: our estimators converge to causal functions in $\sup$ norm, which is a useful norm for policymakers who may be concerned about each treatment value. Our nonasymptotic rates of convergence combine minimax optimal rates for smooth nonparametric regressions, and they explicitly account for each source of error at any finite sample size. Our rates do not directly depend on the data dimension, but rather the smoothness of nonlinear functions and the spectral decay of covariance operators, generalizing standard Sobolev assumptions. The rates may indirectly depend on dimension; see Section 4 for a discussion. Of independent interest, we provide a technical innovation to justify our main results: relative to previous work, we prove faster rates of convergence in Hilbert–Schmidt norm for conditional expectation operators. We generalize our main results to prove weak convergence for counterfactual distributions. Future research may provide uniform confidence bands.

Empirically, we demonstrate how kernel methods for causal functions are practical tools for empirical economics through a program evaluation of the Job Corps, the largest US job training

program for disadvantaged youths. Our key statistical assumption is that different intensities of job training have smooth effects on counterfactual employment, and those effects are smoothly modified by age. We find that the effect of job training on employment substantially varies by class hours and by age; a targeted policy may be more effective. Our program evaluation confirms earlier findings while also uncovering meaningful heterogeneity.

2. Related work

We view nonparametric causal functions as reweightings of an underlying regression, synthesizing the g formula (Robins, 1986) and partial means (Newey, 1994b) frameworks. We quote identification theorems that assume selection on observables (Rosenbaum & Rubin, 1983; Robins, 1986; Altonji & Matzkin, 2005) then propose simple, global estimators that combine kernel ridge regressions. Previous works that take a global view include van der Laan & Dudoit (2003); van der Laan (2006); Díaz & van der Laan (2013); Luedtke & van der Laan (2016b); Semenova & Chernozhukov (2021); Foster & Syrgkanis (2019); Kennedy (2020), and references therein. A broad literature instead views causal functions as collections of localized treatment effects and proposes local estimators with Nadaraya–Watson smoothing, e.g. Imai & Van Dyk (2004); Rubin & van der Laan (2005, 2006); Galvao & Wang (2015); Luedtke & van der Laan (2016a); Kennedy et al. (2017); Kallus & Zhou (2018); Chernozhukov et al. (2022); Fan et al. (2022); Zimmert & Lechner (2019); Colangelo & Lee (2020); Chernozhukov et al. (2023), and references therein. By taking a global view, we propose simple estimators that can be computed once and evaluated at any value of a continuous treatment, rather than a computationally intensive procedure that must be reimplemented at any treatment value. Section 6 gives comparisons.

Our work appears to be the first to reduce the estimation of dose, heterogeneous, and incremental response curves to kernel ridge regressions. Previous works incorporating the RKHS into nonparametric estimation focus on different causal functions: the nonparametric instrumental variable regression (Carrasco et al., 2007; Darolles et al., 2011; Singh et al., 2019), and the heterogeneous treatment effect conditional on the full vector of covariates (Nie & Wager, 2021). Nie & Wager (2021) propose the R learner to estimate the heterogeneous treatment effect $\theta_0(x) = E\{Y^{(1)} - Y^{(0)} \mid X = x\}$, and review the extensive literature that considers this estimand. The R learner minimizes a loss that contains inverse propensities and different regularization (Nie & Wager, 2021, eq. A24), and it does not appear to have a closed form solution. The authors prove oracle mean square error rates. By contrast, we pursue a more general heterogeneous response curve with a discrete or continuous treatment, conditional on some interpretable subvector V (van der Laan, 2006; Abrevaya et al., 2015): $\theta_0(d, v) = E\{Y^{(d)} \mid V = v\}$. Unlike previous work on nonparametric causal functions in the RKHS, we (i) consider dose, heterogeneous, and incremental response curves; (ii) propose estimators with simple closed form solutions; and (iii) prove uniform consistency, which is important for policy evaluation.

We extend the framework from causal functions to counterfactual distributions. Existing work focuses on distributional generalizations of the average treatment effect (ATE) or average treatment on the treated (ATT) for a binary treatment (Firpo, 2007; Cattaneo, 2010; Chernozhukov et al., 2013), e.g. $\theta_0 = P\{Y^{(1)}\} - P\{Y^{(0)}\}$. Muandet et al. (2021) propose an RKHS approach for distributional ATE and ATT with a binary treatment using inverse propensity scores and an assumption on the smoothness of a density ratio, which differs from our approach. Unlike previous work, we (i) allow the treatment to be continuous; (ii) avoid the inversion of propensity scores and densities; and (iii) study a broad class of counterfactual distributions for the full population, subpopulations, and alternative populations, e.g. $\theta_0(d, v) = P\{Y^{(d)} \mid V = v\}$.

We provide a detailed comparison with kernel methods for binary treatment effects in Section 6. Whereas we study causal functions, these works study causal scalars (Kallus, 2020; Hirshberg et al., 2019; Singh, 2021). We clarify the sense in which our causal function estimators generalize known estimators for treatment effects to new estimators for causal functions. Previous work is inherently tied to the $\mathbb{L}^2$ bounded functional perspective. However, evaluation of a causal function is not a bounded functional over all of $\mathbb{L}^2$, as described in Section 1. Therefore our algorithms extend the conceptual framework of kernel methods for causal inference in a new direction. Our statistical contribution is a new, uniform analysis of response curves that goes beyond pointwise approximation of response curves by local treatment effects.

This paper subsumes our previous draft (Singh et al., 2020, Section 2).

3. CAUSAL FUNCTIONS

A causal function summarizes the expected counterfactual outcome $Y^{(d)}$ given a hypothetical intervention on a continuous treatment that sets $D = d$. The causal inference literature studies a rich variety of causal functions with nuanced interpretations, which we define below. Unless otherwise noted, expectations are with respect to the population distribution P.

DEFINITION 1 (CAUSAL FUNCTIONS). *We define the following.*

1. *Dose response:* $\theta_0^{ATE}(d) = E\{Y^{(d)}\}$ *is the counterfactual mean outcome given the intervention* $D = d$ *for the entire population.*
2. *Dose response with distribution shift:* $\theta_0^{DS}(d, \tilde{P}) = E_{\tilde{P}}\{Y^{(d)}\}$ *is the counterfactual mean outcome given the intervention* $D = d$ *for an alternative population with the distribution* $\tilde{P}$.
3. *Conditional response:* $\theta_0^{ATT}(d, d') = E\{Y^{(d')} \mid D = d\}$ *is the counterfactual mean outcome given the intervention* $D = d'$ *for the subpopulation who received the treatment* $D = d$.
4. *Heterogeneous response:* $\theta_0^{CATE}(d, v) = E\{Y^{(d)} \mid V = v\}$ *is the counterfactual mean outcome given the intervention* $D = d$ *for the subpopulation with the subcovariate value* $V = v$.

Likewise we define incremental functions, e.g. $\theta_0^{\nabla:ATE}(d) = E\{\nabla_d Y^{(d)}\}$ *where* ∇_d *means* $\partial/\partial d$.

The superscript of each causal function corresponds to its familiar parametric analogue. Results for the means of potential outcomes immediately imply results for the differences thereof. See Supplement 2 for counterfactual distributions and Supplement 3 for graphical models.

The dose response curves $\theta_0^{ATE}(d)$ and $\theta_0^{DS}(d, \tilde{P})$ are causal functions for entire populations. The second argument of $\theta_0^{DS}(d, \tilde{P})$ concerns external validity: though our data were drawn from population P, what would be the dose response for a different population $\tilde{P}$? For example, a job training study may be conducted in Virginia, yet we may wish to inform policy in Arkansas, a state with different demographics (Hotz et al., 2005). Such questions are studied under the names of transfer learning, distribution shift, and covariate shift (Quiñonero-Candela et al., 2009; Pearl & Bareinboim, 2014).

Both $\theta_0^{ATE}(d)$ and $\theta_0^{DS}(d, \tilde{P})$ are dose response curves for entire populations, but causal functions may vary for different subpopulations. Towards the goal of personalized interventions, an analyst may ask another nuanced counterfactual question: what would have been the effect of the treatment $D = d'$ for the subpopulation who received the treatment $D = d$? When the treatment is continuous, we may define the conditional response $\theta_0^{ATT}(d, d') = E\{Y^{(d')} \mid D = d\}$.

For $\theta_0^{ATT}(d, d')$, heterogeneity is indexed by the treatment D. Heterogeneity may instead be indexed by some interpretable covariate subvector V, e.g. age, race, or gender, and an analyst may wish to measure effects for subpopulations characterized by different values of V (van der

Laan, 2006; Abrevaya et al., 2015). For simplicity, we write the covariates as (V, X) for this setting, where X are additional identifying covariates besides the interpretable covariates V. While many works focus on the special case where the treatment is binary, our definition of the heterogeneous response curve $\theta_0^{CATE}(d, v) = E\{Y^{(d)} \mid V = v\}$ allows for a continuous treatment.

LEMMA 1 (IDENTIFICATION (ROSENBAUM & RUBIN, 1983; ROBINS, 1986)). *Under standard assumptions of selection on observables and covariate shift in Supplement 1,* $\theta_0^{ATE}(d) = \int \gamma_0(d, x) \mathrm{d}P(x)$, $\theta_0^{DS}(d, \tilde{P}) = \int \gamma_0(d, x) \mathrm{d}\tilde{P}(x)$, $\theta_0^{ATT}(d, d') = \int \gamma_0(d', x) \mathrm{d}P(x \mid d)$, *and* $\theta_0^{CATE}(d, v) = \int \gamma_0(d, v, x) \mathrm{d}P(x \mid v)$, *where* $\gamma_0(d, x) = E(Y \mid D = d, X = x)$ *and* $\gamma_0(d, v, x) = E(Y \mid D = d, V = v, X = x)$. *Likewise we identify incremental functions, e.g.* $\theta_0^{\nabla:ATE}(d) = \int \nabla_d \gamma_0(d, x) \mathrm{d}P(x)$ *(Altonji & Matzkin, 2005).*

Lemma 1 expresses each causal function as an integral of the regression function γ_0 according to a marginal or conditional distribution. As previewed in Section 1, nonparametric estimation of $\theta_0^{CATE}(d, v)$ involves three steps: estimating a nonlinear regression $\gamma_0(d, v, x)$, which may involve many covariates X; estimating the conditional distribution $P(x \mid v)$ for reweighting; and using the latter to integrate the former. In what follows, we summarize the RKHS concepts that we will use to propose original estimators that achieve all three steps of nonparametric estimation in a simple closed form solution with finite sample uniform guarantees.

4. RKHS BACKGROUND

4.1. *Concepts for algorithm derivations in Sections 5 and 6*

A scalar-valued RKHS $\mathcal{H}$ is a Hilbert space of functions $\gamma : \mathcal{W} \to \mathbb{R}$. The RKHS is fully characterized by its feature map, which takes a point w in the original space $\mathcal{W}$ and maps it to a feature $\phi(w)$ in the RKHS $\mathcal{H}$. The closure of $span\{\phi(w)\}_{w \in \mathcal{W}}$ is the RKHS $\mathcal{H}$. In other words, $\{\phi(w)\}_{w \in \mathcal{W}}$ can be viewed as the dictionary of basis functions for the RKHS $\mathcal{H}$. The kernel $k : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ is the inner product of features $\phi(w)$ and $\phi(w')$: $k(w, w') = \langle \phi(w), \phi(w') \rangle_{\mathcal{H}}$. A real-valued kernel k is continuous, symmetric, and positive definite.

The essential property of a function γ in an RKHS $\mathcal{H}$ is the eponymous reproducing property: $\gamma(w) = \langle \gamma, \phi(w) \rangle_{\mathcal{H}}$. In other words, to evaluate γ at w, we take the RKHS inner product between γ and the features $\phi(w)$ for $\mathcal{H}$. We formally define the RKHS inner product and the features below. Our key insight is to interpret the reproducing property as a way to separate the function γ from the features $\phi(w)$ and thereby decouple the three steps of nonparametric causal estimation.

The RKHS is a practical hypothesis space for nonparametric regression. Consider the output $Y \in \mathbb{R}$, the input $W \in \mathcal{W}$, and the goal of estimating the conditional expectation function $\gamma_0(w) = E(Y \mid W = w)$. A kernel ridge regression estimator of γ_0 is

$$\hat{\gamma} = \underset{\gamma \in \mathcal{H}}{\operatorname{argmin}}\, n^{-1} \sum_{i=1}^{n} \{Y_i - \langle \gamma, \phi(W_i) \rangle_{\mathcal{H}}\}^2 + \lambda \|\gamma\|_{\mathcal{H}}^2, \tag{1}$$

where $\lambda > 0$ is a hyperparameter on the ridge penalty $\|\gamma\|_{\mathcal{H}}^2$, which imposes smoothness in estimation. The solution to the optimization problem has a well known closed form (Kimeldorf & Wahba, 1971), which we exploit and generalize throughout this work:

$$\hat{\gamma}(w) = Y^\top (K_{WW} + n\lambda I)^{-1} K_{Ww}. \tag{2}$$

The closed form solution involves the kernel matrix $K_{WW} \in \mathbb{R}^{n \times n}$ with (i, j)th entry $k(W_i, W_j)$, and the kernel vector $K_{Ww} \in \mathbb{R}^n$ with ith entry $k(W_i, w)$. To tune the ridge hy-

perparameter λ, both generalized cross validation and leave-one-out cross validation have closed form solutions, and the former is asymptotically optimal (Craven & Wahba, 1978; Li, 1986).

The feature map takes a value in the original space $w \in \mathcal{W}$ and maps it to a feature in the RKHS $\phi(w) \in \mathcal{H}$. One may generalize this idea, from the embedding of a value w by $\phi(w) \in \mathcal{H}$, to the embedding of a distribution Q by $\mu = E_Q\{\phi(W)\} \in \mathcal{H}$ (Berlinet & Thomas-Agnan, 2004; Smola et al., 2007). Boundedness of the kernel implies existence of the mean embedding. Mean embeddings facilitate the evaluation of expectations of RKHS functions: for $\gamma \in \mathcal{H}$, $E_Q\{\gamma(W)\} = E_Q\left\{\langle \gamma, \phi(W)\rangle_{\mathcal{H}}\right\} = \langle \gamma, \mu\rangle_{\mathcal{H}}$. The final expression foreshadows how we will use the technique of mean embeddings to decouple the nonparametric regression step from the nonparametric reweighting step in the estimation of causal functions. A natural question is whether the embedding $Q \mapsto E_Q\{\phi(W)\}$ is injective, i.e. whether the RKHS representation of the distribution is unique. This is called the characteristic property of the kernel k, and it holds for commonly used RKHSs e.g. the exponentiated quadratic kernel (Sriperumbudur et al., 2010).

The tensor product RKHS is one way to construct an RKHS for functions with multiple arguments. Consider the RKHSs $\mathcal{H}_1$ and $\mathcal{H}_2$ with positive definite kernels $k_1 : \mathcal{W}_1 \times \mathcal{W}_1 \to \mathbb{R}$ and $k_2 : \mathcal{W}_2 \times \mathcal{W}_2 \to \mathbb{R}$, respectively. An element $\gamma_1 \in \mathcal{H}_1$ is a function $\gamma_1 : \mathcal{W}_1 \to \mathbb{R}$ and an element $\gamma_2 \in \mathcal{H}_2$ is a function $\gamma_2 : \mathcal{W}_2 \to \mathbb{R}$. The tensor product RKHS $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ is the RKHS with the product kernel $k : (\mathcal{W}_1 \times \mathcal{W}_2) \times (\mathcal{W}_1 \times \mathcal{W}_2) \to \mathbb{R},\ \{(w_1, w_2), (w_1', w_2')\} \mapsto k_1(w_1, w_1')k_2(w_2, w_2')$. Equivalently, the tensor product RKHS $\mathcal{H}$ has the feature map $\phi(w_1) \otimes \phi(w_2)$ such that $\|\phi(w_1) \otimes \phi(w_2)\|_{\mathcal{H}} = \|\phi(w_1)\|_{\mathcal{H}_1}\|\phi(w_2)\|_{\mathcal{H}_2}$. Formally, tensor product notation means $(a \otimes b)c = a\langle b, c\rangle$. An element of the tensor product RKHS $\gamma \in \mathcal{H}$ is a function $\gamma : \mathcal{W}_1 \times \mathcal{W}_2 \to \mathbb{R}$. We assume that the regression function $\gamma_0(w_1, w_2) = E(Y \mid W = w_1, w_2 = w_2)$ is an element of a tensor product RKHS, i.e. $\gamma_0 \in \mathcal{H}$. As such, the different arguments of γ_0 are decoupled, which we exploit when calculating partial means.

Finally, we introduce the RKHS $\mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)$, which is a space of Hilbert–Schmidt operators from one RKHS to another. If the operator E is an element of $\mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)$, then $E : \mathcal{H}_1 \to \mathcal{H}_2$. Moreover, tensor products form Hilbert–Schmidt operators, e.g. $\|\phi(w_1) \otimes \phi(w_2)\|_{\mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)} = \|\phi(w_1)\|_{\mathcal{H}_1}\|\phi(w_2)\|_{\mathcal{H}_2}$. The space $\mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)$ is an example of a vector-valued RKHS with an appropriately defined kernel and feature map (Micchelli & Pontil, 2005). In the present work, we assume that the conditional expectation operator $E_0 : \gamma_1(\cdot) \mapsto E\{\gamma_1(W_1) \mid W_2 = \cdot\}$ is an element of this RKHS. We estimate E_0 by a kernel ridge regression in $\mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)$, which coincides with estimating the embedding $\mu_{w_1}(w_2) = E\{\phi(W_1) \mid W_2 = w_2\}$ via the kernel ridge regression of $\phi(W_1)$ on $\phi(W_2)$; see Section 5.

Remark 1 (*Takeaways*). For $\gamma \in \mathcal{H}$, $\gamma(w) = \langle \gamma, \phi(w)\rangle_{\mathcal{H}}$, where $\phi(w)$ is called the feature map for the RKHS $\mathcal{H}$. Moreover, $E\{\gamma(W)\} = \langle \gamma, \mu\rangle_{\mathcal{H}}$, where μ is called the mean embedding of the distribution of W. Ridge regression in $\mathcal{H}$ has a closed form solution. We can construct RKHSs with these properties for functions of multiple variables and for operators. See Supplement 5 for further technical details.

4.2. *Concepts for consistency proofs in Section 7*

To prove uniform consistency, we place approximation assumptions which are standard in RKHS learning theory: smoothness and spectral decay. To define these approximation assumptions, we introduce an eigendecomposition. Recall the example of a generic RKHS $\mathcal{H}$ with the kernel $k : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ consisting of functions of the form $\gamma : \mathcal{W} \to \mathbb{R}$. Let ν be any Borel measure on $\mathcal{W}$. Let $\mathbb{L}^2_\nu(\mathcal{W})$ be the space of square integrable functions with respect to the measure ν. Define the integral operator $L : \mathbb{L}^2_\nu(\mathcal{W}) \to \mathbb{L}^2_\nu(\mathcal{W}),\ \gamma \mapsto \int k(\cdot, w)\gamma(w)\mathrm{d}\nu(w)$ with eigenvalues

(η_j) and eigenfunctions $\{(\varphi_j)_\nu\}$, the latter of which are equivalence classes in $\mathbb{L}^2_\nu(\mathcal{W})$. Let φ_j be a continuous element in the equivalence class $(\varphi_j)_\nu$.

The following observations help to interpret this eigendecomposition. Without loss of generality, $\eta_j \geq \eta_{j+1}$, and (η_j) are the eigenvalues of the feature covariance operator $T = E\{\phi(W) \otimes \phi(W)\}$. The eigenfunctions $\{(\varphi_j)_\nu\}$ form an orthonormal basis of $\mathbb{L}^2_\nu(\mathcal{W})$. By a generalized Mercer's Theorem, under regularity conditions described in Supplement 7, $k(w, w') = \sum_{j=1}^{\infty} \eta_j \varphi_j(w)\varphi_j(w')$, where (w, w') are in the support of ν; under regularity conditions, (η_j) and (φ_j) describe the eigendecomposition of the kernel. Since $k(w, w') = \langle \phi(w), \phi(w') \rangle_{\mathcal{H}}$, we can therefore express the feature map $\phi(w)$ as $\{\eta_j^{1/2}\varphi_j(w)\}$ for w in the support of ν.

We have seen how to conduct kernel ridge regression with the RKHS $\mathcal{H}$. To analyze the bias from ridge regularization, we place a smoothness assumption called the source condition on the regression function $\gamma_0(w) = E(Y \mid W = w)$ (Smale & Zhou, 2007; Caponnetto & De Vito, 2007; Carrasco et al., 2007). Formally, we place assumptions of the form

$$\gamma_0 \in \mathcal{H}^c = \left(f = \sum_{j=1}^{\infty} \gamma_j \varphi_j : \sum_{j=1}^{\infty} \frac{\gamma_j^2}{\eta_j^c} < \infty \right) \subset \mathcal{H}, \quad c \in (1, 2]. \tag{3}$$

While $c = 1$ recovers correct specification $\gamma_0 \in \mathcal{H}$, $c \in (1, 2]$ is a stronger condition: γ_0 is a particularly smooth element of $\mathcal{H}$, well approximated by the leading terms in the series $\{(\varphi_j)_\nu\}$. Smoothness delivers uniform consistency. A larger value of c corresponds to a smoother target γ_0 and a faster convergence rate for $\hat{\gamma}$. Rates do not further improve for $c > 2$.

To analyze the variance of kernel ridge regression, we place a spectral decay assumption called the effective dimension of the basis (φ_j) for the RKHS $\mathcal{H}$. To obtain faster convergence rates, we place a direct assumption on the rate at which the eigenvalues (η_j), and hence the importance of the eigenfunctions (φ_j), decay: we assume there exists some constant C such that for all j

$$\eta_j \leq C j^{-b}, \quad b \geq 1. \tag{4}$$

A bounded kernel, which we will assume, implies that b is at least one (Fischer & Steinwart, 2020, Lemma 10). The limit $b \to \infty$ may be interpreted as a finite dimensional RKHS (Caponnetto & De Vito, 2007). For intermediate values of b, the polynomial rate of spectral decay quantifies the effective dimension of the RKHS $\mathcal{H}$ in light of the measure ν. Intuitively, a higher value of b corresponds to a lower effective dimension and a faster convergence rate for $\hat{\gamma}$.

For intuition, we relate the source condition and effective dimension to a familiar notion of smoothness in the Sobolev space, since certain Sobolev spaces are RKHSs. Let $\mathcal{W} \subset [0, 1]^p$. Denote by $\mathbb{H}_2^s$ the Sobolev space with $s > p/2$ derivatives that are square integrable. Suppose $\mathcal{H} = \mathbb{H}_2^s$ is the RKHS used for estimation. Suppose the measure ν supported on $\mathcal{W}$ is absolutely continuous with respect to the uniform distribution and bounded away from zero. If $\gamma_0 \in \mathbb{H}_2^{s_0}$, then $c = s_0/s$ (Pillaud-Vivien et al., 2018; Berthier et al., 2020). Written another way, $(\mathbb{H}_2^s)^c = \mathbb{H}_2^{s_0}$. In this sense, c precisely quantifies the additional smoothness of γ_0 relative to $\mathcal{H}$. Moreover, in this Sobolev space, $b = 2s/p > 1$ (Fischer & Steinwart, 2020). The effective dimension is increasing in the input dimension p and decreasing in the degree of smoothness s. The minimax optimal rate in Sobolev norm is $n^{-(c-1)/\{2(c+1/b)\}} = n^{-(s_0-s)/(2s_0+p)}$, which is achieved by kernel ridge regression with the rate optimal regularization $\lambda = n^{-1/(c+1/b)} = n^{-2s/(2s_0+p)}$. Our analysis applies to Sobolev spaces over $[0, 1]^p$ as a special case; our results are much more general, allowing the treatment and covariates to be in Polish spaces.

Remark 2 (*Takeaways*). Let (φ_j) be the eigenfunctions and (η_j) be the eigenvalues of the kernel. Functions in $\mathbb{L}^2$ can be expressed in terms of (φ_j) with square summable coefficients.

Functions in $\mathcal{H}$ can be expressed in terms of (φ_j) with coefficients that remain square summable after dividing by (η_j). The smoothness assumption is square summability of coefficients after dividing by higher powers of (η_j). The spectral decay assumption is the rate at which (η_j) vanish. See Supplement 7 for further technical details.

5. ALGORITHM

5.1. *Decoupled representation*

Lemma 1 makes precise how each causal function is identified as a partial mean of the form $\int \gamma_0(d,x)\mathrm{d}Q$ for some distribution Q. To facilitate estimation, we now assume that γ_0 is an element of an RKHS. In our construction, we define scalar-valued RKHSs for the treatment D and covariates (V,X), then assume that the regression is an element of the tensor product space. Let $k_{\mathcal{D}} : \mathcal{D}\times\mathcal{D}\to\mathbb{R}$, $k_{\mathcal{V}} : \mathcal{V}\times\mathcal{V}\to\mathbb{R}$, and $k_{\mathcal{X}} : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be measurable positive definite kernels corresponding to scalar-valued RKHSs $\mathcal{H}_{\mathcal{D}}$, $\mathcal{H}_{\mathcal{V}}$, and $\mathcal{H}_{\mathcal{X}}$. Denote the feature maps $\phi_{\mathcal{D}} : \mathcal{D}\to\mathcal{H}_{\mathcal{D}},\ d\mapsto k_{\mathcal{D}}(d,\cdot);\ \phi_{\mathcal{V}} : \mathcal{V}\to\mathcal{H}_{\mathcal{V}},\ v\mapsto k_{\mathcal{V}}(v,\cdot);\ \phi_{\mathcal{X}} : \mathcal{X}\to\mathcal{H}_{\mathcal{X}},\ x\mapsto k_{\mathcal{X}}(x,\cdot)$. To lighten notation, we suppress subscripts when arguments are provided.

For θ_0^{ATE}, θ_0^{DS}, and θ_0^{ATT}, we assume the regression γ_0 is an element of the RKHS $\mathcal{H}$ with the kernel $k(d,x;d',x') = k_{\mathcal{D}}(d,d')k_{\mathcal{X}}(x,x')$. We appeal to the fact that the product of positive definite kernels for $\mathcal{H}_{\mathcal{D}}$ and $\mathcal{H}_{\mathcal{X}}$ defines a new positive definite kernel for $\mathcal{H}$. The product construction provides a rich composite basis; $\mathcal{H}$ has the tensor product feature map $\phi(d)\otimes\phi(x)$ and $\mathcal{H} = \mathcal{H}_{\mathcal{D}}\otimes\mathcal{H}_{\mathcal{X}}$. In this RKHS, $\gamma_0(d,x) = \langle\gamma_0, \phi(d)\otimes\phi(x)\rangle_{\mathcal{H}}$. Likewise for θ_0^{CATE} we assume $\gamma_0\in\mathcal{H} = \mathcal{H}_{\mathcal{D}}\otimes\mathcal{H}_{\mathcal{V}}\otimes\mathcal{H}_{\mathcal{X}}$. We place regularity conditions on this RKHS construction in order to represent causal functions as inner products in $\mathcal{H}$. In anticipation of counterfactual distributions in Supplement 2, we also include conditions for an outcome RKHS in parentheses.

Assumption 1 (*RKHS regularity conditions*). Assume (i) $k_{\mathcal{D}}$, $k_{\mathcal{V}}$, $k_{\mathcal{X}}$ (and $k_{\mathcal{Y}}$) are continuous and bounded, i.e. $\sup_{d\in\mathcal{D}}\|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}}\le\kappa_d$, $\sup_{v\in\mathcal{V}}\|\phi(v)\|_{\mathcal{H}_{\mathcal{V}}}\le\kappa_v$, $\sup_{x\in\mathcal{X}}\|\phi(x)\|_{\mathcal{H}_{\mathcal{X}}}\le\kappa_x$ {and $\sup_{y\in\mathcal{Y}}\|\phi(y)\|_{\mathcal{H}_{\mathcal{Y}}}\le\kappa_y$}; (ii) $\phi(d)$, $\phi(v)$, $\phi(x)$ {and $\phi(y)$} are measurable; (iii) $k_{\mathcal{X}}$ (and $k_{\mathcal{Y}}$) are characteristic. For incremental functions, further assume $\mathcal{D}\subset\mathbb{R}$ is an open set and $\nabla_d\nabla_{d'}k_{\mathcal{D}}(d,d')$ exists and is continuous, hence $\sup_{d\in\mathcal{D}}\|\nabla_d\phi(d)\|_{\mathcal{H}}\le\kappa'_d$.

Commonly used kernels are continuous and bounded. Measurability is a similarly weak condition. The characteristic property ensures injectivity of the mean embeddings.

THEOREM 1 (DECOUPLING VIA KERNEL MEAN EMBEDDINGS). *Suppose the conditions of Lemma 1, Assumption 1, and $\gamma_0\in\mathcal{H}$ hold. Then (i) $\theta_0^{ATE}(d) = \langle\gamma_0,\phi(d)\otimes\mu_x\rangle_{\mathcal{H}}$ where $\mu_x = \int\phi(x)\mathrm{d}P(x)$; (ii) $\theta_0^{DS}(d,\tilde{P}) = \langle\gamma_0,\phi(d)\otimes\nu_x\rangle_{\mathcal{H}}$ where $\nu_x = \int\phi(x)\mathrm{d}\tilde{P}(x)$; (iii) $\theta_0^{ATT}(d,d') = \langle\gamma_0,\phi(d')\otimes\mu_x(d)\rangle_{\mathcal{H}}$ where $\mu_x(d) = \int\phi(x)\mathrm{d}P(x\mid d)$; (iv) $\theta_0^{CATE}(d,v) = \langle\gamma_0,\phi(d)\otimes\phi(v)\otimes\mu_x(v)\rangle_{\mathcal{H}}$ where $\mu_x(v) = \int\phi(x)\mathrm{d}P(x\mid v)$. Likewise for incremental functions, e.g. $\theta_0^{\nabla:ATE}(d) = \langle\gamma_0,\nabla_d\phi(d)\otimes\mu_x\rangle_{\mathcal{H}}$.*

Proof sketch. Consider $\theta_0^{CATE}(d,v)$. Boundedness of the kernel implies Bochner integrability, which allows us to exchange the integral and inner product:

$$\int\gamma_0(d,v,x)\mathrm{d}P(x\mid v) = \int\langle\gamma_0,\phi(d)\otimes\phi(v)\otimes\phi(x)\rangle_{\mathcal{H}}\mathrm{d}P(x\mid v) = \langle\gamma_0,\phi(d)\otimes\mu_x(v)\rangle_{\mathcal{H}}.$$

See Supplement 4 for the full proof. Here, $\mu_x(v) = \int\phi(x)P(x\mid v)$ is the mean embedding of the conditional distribution $P(x\mid v)$. It encodes the distribution $P(x\mid v)$ as a function $\mu_x(v)\in\mathcal{H}_{\mathcal{X}}$ such that the causal function $\theta_0^{CATE}(d,v)$ can be expressed as an inner product in $\mathcal{H}$.

5.2. *Closed form solution*

The representation in Theorem 1 is essential to the algorithm derivation. In particular, the representation cleanly separates the three steps necessary to estimate a causal function: estimating a nonlinear regression, which may involve many covariates; estimating the distribution for reweighting; and using the nonparametric distribution to integrate the nonparametric regression. For example, for $\theta_0^{CATE}(d,v)$, our estimator is $\hat{\theta}^{CATE}(d,v) = \langle \hat{\gamma}, \phi(d) \otimes \phi(v) \otimes \hat{\mu}_x(v) \rangle_{\mathcal{H}}$. The nonlinear regression estimator $\hat{\gamma}$ is a standard kernel ridge regression of Y on $\phi(D) \otimes \phi(V) \otimes \phi(X)$; the reweighting distribution estimator $\hat{\mu}_x(v)$ is a generalized kernel ridge regression of $\phi(X)$ on $\phi(V)$; and the latter can be used to integrate the former by simply multiplying the two. This algorithmic insight is a key innovation of the present work, and the reason why our estimators have simple closed form solutions despite possibly complicated integration.

Algorithm 1 (*Estimation of causal functions*). Denote the empirical kernel matrices $K_{DD}, K_{VV}, K_{XX} \in \mathbb{R}^{n \times n}$ calculated from observations drawn from the population P. Let $\tilde{X}_i$ $(i = 1, ..., \tilde{n})$ be observations drawn from the population $\tilde{P}$. Denote by $\odot$ the elementwise product. Causal function estimators have the closed form solutions

1. $\hat{\theta}^{ATE}(d) = n^{-1} \sum_{i=1}^{n} Y^\top (K_{DD} \odot K_{XX} + n\lambda I)^{-1} (K_{Dd} \odot K_{Xx_i})$;
2. $\hat{\theta}^{DS}(d, \tilde{P}) = \tilde{n}^{-1} \sum_{i=1}^{\tilde{n}} Y^\top (K_{DD} \odot K_{XX} + n\lambda I)^{-1} (K_{Dd} \odot K_{X\tilde{x}_i})$;
3. $\hat{\theta}^{ATT}(d, d') = Y^\top (K_{DD} \odot K_{XX} + n\lambda I)^{-1} [K_{Dd'} \odot \{K_{XX}(K_{DD} + n\lambda_1 I)^{-1} K_{Dd}\}]$;
4. $\hat{\theta}^{CATE}(d,v) = Y^\top (K_{DD} \odot K_{VV} \odot K_{XX} + n\lambda I)^{-1} [K_{Dd} \odot K_{Vv} \odot \{K_{XX}(K_{VV} + n\lambda_2 I)^{-1} K_{Vv}\}]$;

where $(\lambda, \lambda_1, \lambda_2)$ are ridge regression penalty hyperparameters. Likewise for incremental functions, e.g. $\hat{\theta}^{\nabla:ATE}(d) = n^{-1} \sum_{i=1}^{n} Y^\top (K_{DD} \odot K_{XX} + n\lambda I)^{-1} (\nabla_d K_{Dd} \odot K_{Xx_i})$ where $(\nabla_d K_{Dd})_i = \nabla_d k(D_i, d)$.

Proof sketch. Consider $\theta_0^{CATE}(d,v)$. Analogously to (1), the kernel ridge regression estimators of the regression γ_0 and the conditional mean embedding $\mu_x(v)$ are given by $\hat{\gamma} = \operatorname{argmin}_{\gamma \in \mathcal{H}} n^{-1} \sum_{i=1}^{n} \{Y_i - \langle \gamma, \phi(D_i) \otimes \phi(V_i) \otimes \phi(X_i) \rangle_{\mathcal{H}}\}^2 + \lambda \|\gamma\|_{\mathcal{H}}^2$ and $\hat{E} = \operatorname{argmin}_{E \in \mathcal{L}_2(\mathcal{H}_X, \mathcal{H}_V)} n^{-1} \sum_{i=1}^{n} \{\phi(X_i) - E^* \phi(V_i)\}^2 + \lambda_2 \|E\|^2_{\mathcal{L}_2(\mathcal{H}_X, \mathcal{H}_V)}$, where $\hat{\mu}_x(v) = \hat{E}^* \phi(v)$ and E^* is the adjoint of E. Analogously to (2), the closed forms are

$$\hat{\gamma}(d, v, \cdot) = Y^\top (K_{DD} \odot K_{VV} \odot K_{XX} + n\lambda I)^{-1} \{K_{Dd} \odot K_{Vv} \odot K_{X(\cdot)}\},$$
$$[\hat{\mu}_x(v)](\cdot) = K_{(\cdot)X}(K_{VV} + n\lambda_2 I)^{-1} K_{Vv}.$$

To arrive at the main result, match the empty arguments $(\cdot)$ of the kernel ridge regressions. □

See Supplement 4 for the full derivation and a comparison to series estimation. We give theoretical values for $(\lambda, \lambda_1, \lambda_2)$ that optimally balance bias and variance in Theorem 2 below. Supplement 5 gives practical tuning procedures based on generalized and leave-one-out cross validation to empirically balance bias and variance, the former of which is asymptotically optimal.

6. Detailed comparisons to kernel methods for causal scalars

We now connect our kernel methods for causal functions with related kernel methods for treatment effects. Recall the definition $\theta_0^{ATE}(d) = E\{Y^{(d)}\}$. We allow the treatment to be continuous, so θ_0^{ATE} is a causal function called the dose response. In related work, the treatment is

binary, so θ_0^{ATE} is a vector of two causal scalars $\theta_0^{ATE}(1), \theta_0^{ATE}(0)$ whose difference is called the treatment effect.

We clarify three points. (i) There is a sense in which our algorithms generalize known estimators for treatment effects to new estimators for causal functions. (ii) A treatment effect is a bounded functional over $\mathbb{L}^2$ with a balancing weight representation, while a response curve is not. Our key insight is that a response curve is a bounded functional over the RKHS $\mathcal{H}$, which is a subset of $\mathbb{L}^2$. (iii) Our theoretical contribution is a new, uniform analysis of response curves. The analysis goes beyond pointwise approximation of response curves by local treatment effects.

We begin by reviewing the theory of balancing weights, which are widely used in causal inference with a binary treatment. For clarity, in this section we emphasize a fixed treatment value by writing $d^* \in \mathcal{D}$. The following representation is well known.

PROPOSITION 1 (EXISTENCE FOR TREATMENT EFFECTS (HERNÁN & ROBINS, 2020)). *Suppose selection on observables holds as stated in Supplement 1, and that the treatment is binary. Fix $d^* \in \mathcal{D}$. If $\text{pr}(D = d^* \mid X)$ is bounded away from zero almost surely, then there exists a balancing weight $\alpha_0 \in \mathbb{L}^2$ such that for all $\gamma \in \mathbb{L}^2$, $\int \gamma(d^*, x) \mathrm{d}P(x) = \langle \gamma, \alpha_0 \rangle_{\mathbb{L}^2}$. In particular, $\theta_0^{ATE}(d^*) = \int y \alpha_0(d, x) \mathrm{d}P(d, x, y) = \langle \gamma_0, \alpha_0 \rangle_{\mathbb{L}^2}$ and the balancing weight is $\alpha_0(d, x) = 1(d = d^*)/\text{pr}(D = d^* \mid x)$, where $1(\cdot)$ is the indicator function.*

In summary, a treatment effect has two representations: the primal representation of Lemma 1 as a partial mean of the regression $\gamma_0(d, x) = E(Y \mid D = d, X = x)$, and the dual representation of Proposition 1 as a reweighting of the outcome Y using the balancing weight $\alpha_0(d, x) = 1(d = d^*)/\text{pr}(D = d^* \mid x)$. Clearly, the two representations are related by the law of iterated expectations. Moreover, from the closed form of α_0, we require $\text{pr}(D = d^* \mid X) > 0$ for α_0 to exist. This property keenly relies on the treatment being discrete. Indeed, it is well known that a balancing weight representation does not exist for response curves.

PROPOSITION 2 (NON-EXISTENCE FOR RESPONSE CURVES (VAN DER VAART, 1991)). *Suppose selection on observables holds as stated in Supplement 1, and that the treatment is continuous. Fix $d^* \in \mathcal{D}$. Even if the density $f(d^* \mid X)$ is bounded away from zero almost surely, there does not exist a balancing weight $\alpha_0 \in \mathbb{L}^2$ such that for all $\gamma \in \mathbb{L}^2$, $\int \gamma(d^*, x) \mathrm{d}P(x) = \langle \gamma, \alpha_0 \rangle_{\mathbb{L}^2}$. In particular, without further restrictions, there does not exist $\alpha_0 \in \mathbb{L}^2$ such that $\theta_0^{ATE}(d^*) = \int y \alpha_0(d, x) \mathrm{d}P(d, x, y) = \langle \gamma_0, \alpha_0 \rangle_{\mathbb{L}^2}$.*

Whereas a binary treatment effect is a bounded functional over $\mathbb{L}^2$ with a balancing weight representation, a dose response is not a bounded functional over $\mathbb{L}^2$ and does not have a balancing weight representation in the classic sense. From a functional analytic perspective, this discrepancy is the reason why the problems we study are nonparametric whereas previous work on kernel methods for treatment effects are semiparametric. See Supplement 6 for a discussion.

Our key insight is that the dose response is a bounded functional over the RKHS $\mathcal{H}$, which is a subset of $\mathbb{L}^2$. This fact follows from three simple observations: (i) the dose response is a partial mean; (ii) in the RKHS, a partial mean can be reformulated as a kind of evaluation; and (iii) the RKHS $\mathcal{H}$ is the subset of $\mathbb{L}^2$ for which evaluation is a bounded functional. Through this lens, Theorem 1 shows that there can exist a function $\tilde{\alpha}_0 \in \mathcal{H}$ such that $\theta_0^{ATE}(d) = \langle \gamma_0, \tilde{\alpha}_0 \rangle_{\mathcal{H}}$ even when there does not exist a function $\alpha_0 \in \mathbb{L}^2$ such that $\theta_0^{ATE}(d) = \langle \gamma_0, \alpha_0 \rangle_{\mathbb{L}^2}$.

What is the relationship between our kernel methods for causal functions and existing kernel methods for treatment effects? There is a sense in which our dose response estimator, which is the simplest case of our framework, is a relaxation of kernel balancing weight estimators from a binary treatment to a continuous treatment. We formalize this connection as follows.

COROLLARY 1 (RELAXATION OF BALANCING WEIGHT ESTIMATORS). *Suppose the treatment is binary, and take $k_{\mathcal{D}}(d, d') = 1(d = d')$ to be the treatment kernel. Then $\hat{\theta}^{ATE}(d^*) = n^{-1}\sum_{i=1}^n Y_i\hat{\alpha}_i$, where $\hat{\alpha}_i = \hat{\alpha}(D_i, X_i)$ and $\hat{\alpha}$ is a ridge regularized estimator of $\alpha_0 \in \mathbb{L}^2$.*

See Supplement 6 for the proof. The balancing weight estimator $\hat{\alpha}$ minimizes a generalized balancing weight loss with ridge regularization; see Kallus (2020, eq. 8), Hirshberg et al. (2019, eq. 1), and Singh (2021, Definition 3.2) for various formulations. Corollary 1 provides intuition for our tensor product RKHS construction, which ensures that using the binary treatment kernel amounts to subsetting and hence recovers previous algorithms. Our RKHS construction naturally relaxes a binary treatment to a continuous treatment while retaining computational tractability.

As argued in Proposition 2, the balancing weight $\alpha_0 \in \mathbb{L}^2$ does not exist for the dose response. Nonetheless, our key insight in Theorem 1 is that a function $\tilde{\alpha}_0 \in \mathcal{H}$ does exist to serve a similar purpose. By combining the partial mean perspective with the technique of kernel mean embedding, we demonstrate that our framework easily extends to conditional nonparametric causal functions, e.g. the heterogeneous response curve $\theta_0^{CATE}(d, v)$, which are substantially more challenging than unconditional nonparametric causal functions, e.g. the dose response $\theta_0^{ATE}(d)$.

Perhaps the most surprising consequence of our construction is the closed form solution for causal functions. In particular, each closed form solution is a reweighting of the observed outcomes with empirical weights that we characterize even though a population balancing weight in $\mathbb{L}^2$ does not exist. In sum, previous work (Kallus, 2020; Hirshberg et al., 2019; Singh, 2021) on kernel methods for treatment effects is inherently tied to the $\mathbb{L}^2$ population balancing weight perspective; our algorithms apply the conceptual framework of kernel methods to new classes of causal functions. The following corollary reinterprets Algorithm 1 through this lens.

COROLLARY 2 (CLOSED FORM EVEN WHEN BALANCING WEIGHT DOES NOT EXIST). *Suppose the treatment is continuous, with k_D that is continuous and bounded. Then $\hat{\theta}^{ATE}(d) = n^{-1}\sum_{i=1}^n Y_i\hat{\alpha}_i^{ATE}$, $\hat{\theta}^{DS}(d, \tilde{P}) = n^{-1}\sum_{i=1}^n Y_i\hat{\alpha}_i^{DS}$, $\hat{\theta}^{ATT}(d, d') = n^{-1}\sum_{i=1}^n Y_i\hat{\alpha}_i^{ATT}$, and $\hat{\theta}^{CATE}(d, v) = n^{-1}\sum_{i=1}^n Y_i\hat{\alpha}_i^{CATE}$, where the weights have closed form solutions given in Supplement 6. Likewise for incremental functions, e.g. $\hat{\theta}^{\nabla:ATE}(d) = n^{-1}\sum_{i=1}^n Y_i\hat{\alpha}_i^{\nabla:ATE}$.*

Each of our proposed causal function estimators is global. In particular, within Corollary 2, the weights $(\hat{\alpha}_j^{ATE}, \hat{\alpha}_j^{DS}, \hat{\alpha}_j^{ATT}, \hat{\alpha}_j^{CATE})$ $(j = 1, ..., n)$ depend on all of the observations as refracted through the ridge regularized empirical covariance and the kernel evaluations $k(D_i, d)$. This approach departs from a localization approach to causal functions whereby the weight assigned to each observation is determined by Nadaraya–Watson smoothing (Kennedy et al., 2017; Kallus & Zhou, 2018; Colangelo & Lee, 2020; Chernozhukov et al., 2022). In the localization approach, the weight is $k^{NW}\{(D_i - d)/h\}$ where k^{NW} is a Nadaraya–Watson kernel and h is a vanishing bandwidth. By contrast, we consider a fixed kernel and vanishing ridge regularization.

The global perspective has several advantages. Our estimators can be computed once and evaluated at any value of a continuous treatment. By contrast, a localized estimator is a computationally intensive procedure that must be reimplemented at any treatment value. Next, our estimators are constructed from function classes with designed-in smoothness properties, which leads to smoother and therefore more plausible response curves. We compare our smooth estimate with a jagged localizing estimate in the program evaluation of Section 8. Finally, we prove uniform consistency of response curves, whereas localizations of previous results would only lead to pointwise consistency. These uniform guarantees are the focus of Section 7.

Due to space constraints, we continue this discussion in Supplement 6. In particular, we relate the above discussion to pathwise differentiability. We also connect our kernel methods for heterogeneous response curves with global estimators for heterogeneous treatment effects. We

make three similar points. (i) There is a limited sense in which our algorithms adapt principles from heterogeneous treatment effect estimation to heterogeneous response curve estimation. (ii) The pseudo outcome of a heterogeneous treatment effect admits a balancing weight representation in $\mathbb{L}^2$, while that of a heterogeneous response curve does not. We nonetheless achieve a closed form for the latter estimand by appealing to the geometry of $\mathcal{H}$ rather than $\mathbb{L}^2$. (iii) Our theoretical contribution is a new, uniform analysis of heterogeneous response curves. Our analysis complements mean square (Nie & Wager, 2021), excess risk (Foster & Syrgkanis, 2019), and pointwise (Kennedy, 2020) analyses for heterogeneous treatment effects in the RKHS.

7. UNIFORM CONSISTENCY WITH FINITE SAMPLE RATES

Towards a guarantee of uniform consistency, we place regularity conditions on the original spaces. In anticipation of counterfactual distributions in Supplement 2, we also include conditions for the outcome space in parentheses.

Assumption 2 (*Original space regularity conditions*). Assume $\mathcal{D}$, $\mathcal{V}$, $\mathcal{X}$ (and $\mathcal{Y}$) are Polish spaces. Further assume $\mathcal{Y} \subset \mathbb{R}$, $\int y^2 \mathrm{d}P(y) < \infty$, and a moment condition holds: there exist constants σ,τ such that for all $m \geq 2$, $\int |y - \gamma_0(D, X)|^m \mathrm{d}P(y \mid D, X) \leq m! \sigma^2 \tau^{m-2}/2$ almost surely. For θ_0^{CATE}, replace X with (V, X).

A Polish space is a separable and completely metrizable topological space. Random variables supported in a Polish space may be discrete or continuous and may even be infinite dimensional. A bounded outcome Y implies the moment condition.

Next, we assume the regression γ_0 is smooth in the sense of (3), and $\mathcal{H}$ has low effective dimension in the sense of (4). Denote the jth eigenvalue of the integral operator for $\mathcal{H}$ by $\eta_j(\mathcal{H})$. Recall that $\eta_j(\mathcal{H})$ is also the jth eigenvalue of the feature covariance operator.

Assumption 3 (*Smoothness and spectral decay for regression*). Assume $\gamma_0 \in \mathcal{H}^c$ with $c \in (1, 2]$, and $\eta_j(\mathcal{H}) \leq C j^{-b}$ with $b \geq 1$.

See Supplement 7 for alternative ways of writing and interpreting Assumption 3. We place similar smoothness and spectral decay conditions on the conditional mean embeddings $\mu_x(d)$ and $\mu_x(v)$, which are generalized conditional expectation functions. We articulate this assumption abstractly for the conditional mean embedding $\mu_a(b) = \int \phi(a) \mathrm{d}P(a \mid b)$ where $a \in \mathcal{A}_\ell$ and $b \in \mathcal{B}_\ell$. All one has to do is specify $\mathcal{A}_\ell$ and $\mathcal{B}_\ell$ to specialize the assumption. For $\mu_x(d)$, $\mathcal{A}_1 = \mathcal{X}$ and $\mathcal{B}_1 = \mathcal{D}$; for $\mu_x(v)$, $\mathcal{A}_2 = \mathcal{X}$ and $\mathcal{B}_2 = \mathcal{V}$. For fixed $\mathcal{A}_\ell$ and $\mathcal{B}_\ell$, we parametrize smoothness by c_ℓ and spectral decay by b_ℓ.

Formally, define the conditional expectation operator $E_\ell : \mathcal{H}_{\mathcal{A}_\ell} \to \mathcal{H}_{\mathcal{B}_\ell}$, $f(\cdot) \mapsto E\{f(A_\ell) \mid B_\ell = \cdot\}$. By construction, E_ℓ encodes the same information as $\mu_a(b)$ since

$$\{\mu_a(b)\}(\cdot) = \int \phi(a) \mathrm{d}P(a \mid b) = \{E_\ell \phi(\cdot)\}(b) = \{E_\ell^* \phi(b)\}(\cdot), \quad a \in \mathcal{A}_\ell, \quad b \in \mathcal{B}_\ell,$$

where E_ℓ^* is the adjoint of E_ℓ. We denote the space of Hilbert–Schmidt operators between $\mathcal{H}_{\mathcal{A}_\ell}$ and $\mathcal{H}_{\mathcal{B}_\ell}$ by $\mathcal{L}_2(\mathcal{H}_{\mathcal{A}_\ell}, \mathcal{H}_{\mathcal{B}_\ell})$. Grünewälder et al. (2013) and Singh et al. (2019) prove that $\mathcal{L}_2(\mathcal{H}_{\mathcal{A}_\ell}, \mathcal{H}_{\mathcal{B}_\ell})$ is an RKHS in its own right, for which we can assume smoothness in the sense of (3) and spectral decay in the sense of (4).

Assumption 4 (*Smoothness and spectral decay for mean embedding*). Assume the following: $E_\ell \in \mathcal{L}_2(\mathcal{H}_{\mathcal{A}_\ell}, \mathcal{H}_{\mathcal{B}_\ell}^{c_\ell})$ with $c_\ell \in (1, 2]$, and $\eta_\ell(\mathcal{H}_{\mathcal{B}_\ell}) \leq C j^{-b_\ell}$ with $b_\ell \geq 1$.

Just as we place approximation assumptions for γ_0 in terms of $\mathcal{H}$, which provides the features onto which we project Y, we place approximation assumptions for E_ℓ in terms of $\mathcal{H}_{\mathcal{B}_\ell}$, which provides the features $\phi(B_\ell)$ onto which we project $\phi(A_\ell)$. Under these conditions, we arrive at our main theoretical guarantee.

THEOREM 2 (UNIFORM CONSISTENCY OF CAUSAL FUNCTIONS). *Suppose the conditions of Lemma 1 hold, as well as Assumptions 1, 2, and 3. Set* $(\lambda, \lambda_1, \lambda_2) = \{n^{-1/(c+1/b)}, n^{-1/(c_1+1/b_1)}, n^{-1/(c_2+1/b_2)}\}$*, which is rate optimal regularization.*

1. *Then with high probability* $\|\hat{\theta}^{ATE} - \theta_0^{ATE}\|_\infty = O\left[n^{-(c-1)/\{2(c+1/b)\}}\right]$ *and* $\|\hat{\theta}^{DS}(\cdot, \tilde{P}) - \theta_0^{DS}(\cdot, \tilde{P})\|_\infty = O\left[n^{-(c-1)/\{2(c+1/b)\}} + \tilde{n}^{-1/2}\right]$.
2. *If in addition Assumption 4 holds with* $\mathcal{A}_1 = \mathcal{X}$ *and* $\mathcal{B}_1 = \mathcal{D}$*, then with high probability* $\|\hat{\theta}^{ATT} - \theta_0^{ATT}\|_\infty = O\left[n^{-(c-1)/\{2(c+1/b)\}} + n^{-(c_1-1)/\{2(c_1+1/b_1)\}}\right]$.
3. *If in addition Assumption 4 holds with* $\mathcal{A}_2 = \mathcal{X}$ *and* $\mathcal{B}_2 = \mathcal{V}$*, then with high probability* $\|\hat{\theta}^{CATE} - \theta_0^{CATE}\|_\infty = O\left[n^{-(c-1)/\{2(c+1/b)\}} + n^{-(c_2-1)/\{2(c_2+1/b_2)\}}\right]$.

Likewise for incremental functions, e.g. $\|\hat{\theta}^{\nabla:ATE} - \theta_0^{\nabla:ATE}\|_\infty = O\left[n^{-(c-1)/\{2(c+1/b)\}}\right]$.

Explicit constants hidden by the $O(\cdot)$ notation, as well as explicit specializations of Assumption 4, are indicated in Appendices 7 and 8. These rates approach $n^{-1/4}$ when $(c, c_1, c_2) = 2$ and $(b, b_1, b_2) \to \infty$, i.e. when the regressions are smooth and when the effective dimensions are finite. Interestingly, each rate combines minimax optimal rates in RKHS norm: $n^{-(c-1)/\{2(c+1/b)\}}$ for standard nonparametric regression (Fischer & Steinwart, 2020, Theorem 2); $\tilde{n}^{-1/2}$ for unconditional mean embeddings (Tolstikhin et al., 2017, Theorem 1); and, in contemporaneous work, $n^{-(c_\ell-1)/\{2(c_\ell+1/b_\ell)\}}$ for conditional mean embeddings (Li et al., 2022, Theorem 3).

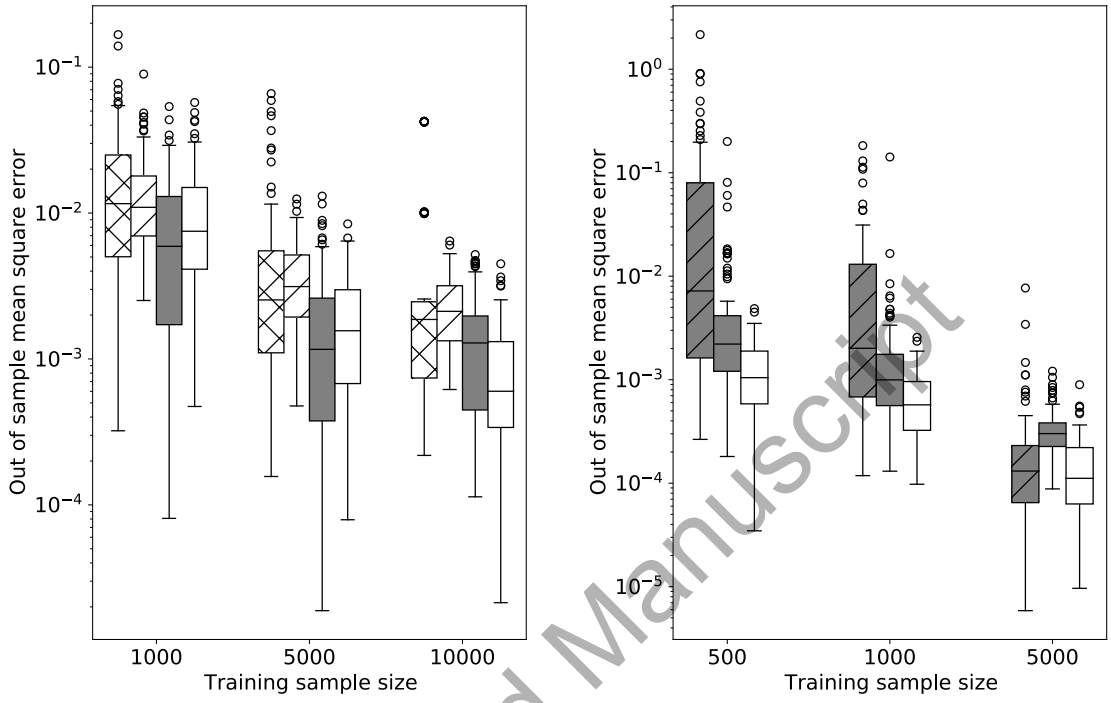
Remark 3 (*Technical innovation*). Our conditional mean embedding rate builds on original analysis of conditional expectation operators in Supplement 8 that is of independent interest. We improve the rate from $n^{-(c_\ell-1)/\{2(c_\ell+1)\}}$ (Singh et al., 2019, Theorem 2) to $n^{-(c_\ell-1)/\{2(c_\ell+1/b_\ell)\}}$. Our consideration of Hilbert–Schmidt norm departs from Park & Muandet (2020) and Talwai et al. (2022), who study surrogate risk and operator norm, respectively. Our assumptions also depart from Singh et al. (2019, Hypothesis 5), Park & Muandet (2020, Theorem 4.5), and Talwai et al. (2022, Assumptions 3 and 4). Instead, Assumption 4 directly generalizes Fischer & Steinwart (2020, Conditions SRC and EVD) from RKHS functions to Hilbert–Schmidt operators.

Overall, rates slower than $n^{-1/4}$ reflect the challenge of a sup norm guarantee, which is stronger than a mean square error guarantee and which is useful for policymakers concerned about each treatment value. For comparison, the minimax optimal Sobolev norm rate for learning an s_0-smooth regression, using $\mathbb{H}_2^s$ over $\mathbb{R}^p$, is $n^{-(c-1)/\{2(c+1/b)\}} = n^{-(s_0-s)/(2s_0+p)}$.

Remark 4 (*Further rate improvements*). We prove uniform rates for an RKHS estimator of the heterogeneous response curve, under smoothness assumptions on the regression function and conditional expectation operators. Supplement 6 connects our estimators with RKHS estimators of heterogeneous treatment effects that have mean square (Nie & Wager, 2021), excess risk (Foster & Syrgkanis, 2019) and pointwise (Kennedy, 2020) convergence rates, under smoothness assumptions on the regression, propensity score, and heterogeneous treatment effect itself. If the heterogeneous treatment effect is smoother than the regression and propensity score, then mean square, excess risk, and pointwise rate improvements are possible (Nie & Wager, 2021; Foster & Syrgkanis, 2019; Kennedy, 2020). Under further assumptions, future research may achieve such rate improvements in sup norm for our RKHS estimator of the heterogeneous response curve.

8. SIMULATIONS AND PROGRAM EVALUATION

8.1. *Simulations*



(a) Dose response curve.

(b) Heterogeneous treatment effect.

Fig. 1: Nonparametric causal function simulations. For the dose response curve, we implement four estimators. From left to right, these are: Kennedy et al. (2017) (`DR1`, checkered white), Colangelo & Lee (2020) (`DR2`, lined white), Semenova & Chernozhukov (2021) (`DR-series`, gray), and our own (`RKHS`, white). For the heterogeneous treatment effect, we implement three estimators. From left to right, these are Abrevaya et al. (2015) (`IPW`, lined gray), Semenova & Chernozhukov (2021) (`DR-series`, gray), and our own (`RKHS`, white). .

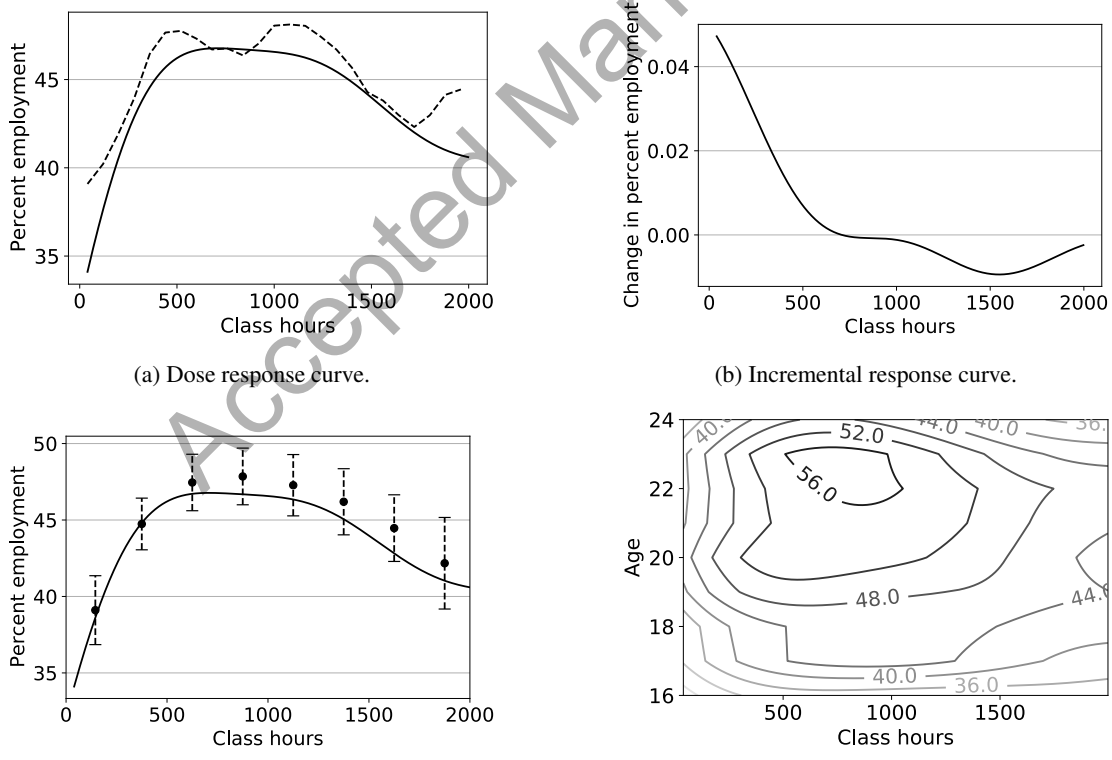
We demonstrate that our nonparametric causal function estimators outperform some leading alternatives in nonlinear simulations with many covariates, despite the relative simplicity of our proposed approach. For each causal function design and sample size, we implement 100 simulations and calculate mean square error with respect to the true causal function. Figure 1 visualizes results. A lower mean square error is desirable. See Supplement 9 for a full exposition of the data generating processes and implementation details.

The dose response curve design (Colangelo & Lee, 2020) considers the causal function $\theta_0^{ATE}(d) = 1.2d + d^2$. A single observation consists of the triple (Y, D, X) for the outcome, treatment, and high dimensional covariates, where $Y, D \in \mathbb{R}$ and $X \in \mathbb{R}^{100}$. In addition to our simple nonparametric estimator (`RKHS`, white), we implement the estimators of Kennedy et al. (2017) (`DR1`, checkered white), Colangelo & Lee (2020) (`DR2`, lined white), and Semenova & Chernozhukov (2021) (`DR-series`, gray). Both `DR1` and `DR2` are local estimators that involve Nadaraya–Watson smoothing with debiased pseudo outcomes, while `DR-series` uses series regression with debiased pseudo outcomes, and we give it the advantage of correct specifica-

tion as a quadratic function. By the Wilcoxon rank sum test, `RKHS` significantly outperforms the alternatives at the sample size 10^4, with a p value less than 10^{-3}, despite its relative simplicity.

Though our approach allows for the heterogeneous response of a continuous treatment, we implement a design for the heterogeneous effect of a binary treatment in order to facilitate comparison with existing methods. The heterogeneous treatment effect design (Abrevaya et al., 2015) considers the causal functions $\theta_0^{CATE}(0, v) = 0$ and $\theta_0^{CATE}(1, v) = v(1 + 2v)^2(v - 1)^2$. Each observation is a tuple (Y, D, V, X) for the outcome, treatment, covariate of interest, and other covariates, where $Y, D, V \in \mathbb{R}$ and $X \in \mathbb{R}^3$. In addition to our simple nonparametric estimator (`RKHS`, white), we implement the estimators of Abrevaya et al. (2015) (`IPW`, lined gray) and Semenova & Chernozhukov (2021) (`DR-series`, gray). The former involves Nadaraya–Watson smoothing around an inverse propensity estimator, and the latter involves correctly specified series regression with a debiased pseudo outcome. The R learner (Nie & Wager, 2021) cannot be implemented since $V \neq X$. The simple `RKHS` approach significantly outperforms the alternatives at sample sizes 500 and 10^3 by the Wilcoxon rank sum test, with p values less than 10^{-5}.

8.2. *Program evaluation: US Job Corps*



(a) Dose response curve.

(b) Incremental response curve.

(c) Discrete treatment effects.

(d) Heterogeneous response curve.

Fig. 2: Effect of job training on employment. We implement our estimators for dose, heterogeneous, and incremental response curves (`RKHS`, solid). For comparison, we also implement the dose response curve estimator of Colangelo & Lee (2020) (`DR2`, dashes) as well as the kernel treatment effects of Singh (2021) (`DR3`, vertical bars).

To demonstrate how kernel methods for causal functions are practical for empirical economics, we conduct a real world program evaluation. Specifically, we estimate dose, heterogeneous, and incremental response curves of the Jobs Corps, the largest job training program for disadvantaged youths in the US, which serves about 50,000 participants annually. Participation is free for individuals who meet low income requirements. Access to the program was randomized from November 1994 to February 1996; see Schochet et al. (2008) for details. Many studies focus on data from this period to evaluate the effect of job training on employment (Flores et al., 2012; Colangelo & Lee, 2020). Though access to the program was randomized, individuals could decide whether to participate and for how many hours. From a causal perspective, we assume that, conditional on the observed covariates, participation hours were as good as random. Statistically, we assume that different intensities of job training have smooth effects on counterfactual employment, and that those effects are smoothly modified by age.

The continuous treatment $D \in \mathbb{R}$ is the total hours spent in academic or vocational classes in the first year after randomization, and the continuous outcome $Y \in \mathbb{R}$ is the proportion of weeks employed in the second year after randomization. The covariates $X \in \mathbb{R}^{40}$ include age, gender, ethnicity, language competency, education, marital status, household size, household income, previous receipt of social aid, family background, health, and health related behavior at base line (Huber et al., 2020). As in Colangelo & Lee (2020), we focus on the $n = 3,906$ observations for which $D \geq 40$, i.e. individuals who completed at least one week of training. We implement various causal parameters in Figure 2: the dose response curve; the incremental response curve; the kernel treatment effects with confidence intervals of Singh (2021); and the heterogeneous response curve with respect to age. For the kernel treatment effects, we discretize treatment into roughly equiprobable bins of class hours. It appears that the heterogeneous response of class hours, a continuous treatment, has not been previously studied in this setting. In Supplement 10, we provide implementation details and verify that our results are robust to the choice of sample.

The dose response curve plateaus and achieves its maximum around $d = 500$, corresponding to 12.5 weeks of classes. Our global estimate (`RKHS`, solid) has the same overall shape but is smoother and slightly lower than the collection of local estimates from Colangelo & Lee (2020) (`DR2`, dashes). The smoothness of our estimator is a consequence of the RKHS assumptions, and we see how it is a virtue for empirical economic research; a smooth dose response curve is more economically plausible in this setting. The first 12.5 weeks of classes confer most of the gain in employment: from 35% employment to more than 47% employment for the average participant. The incremental response curve (`RKHS`, solid) is the derivative of the dose response curve, and it visualizes where the greatest gain happens. The kernel treatment effects of Singh (2021) (`DR3`, vertical bars) corroborate our dose response curve, and their 95% confidence intervals contain the dose response curve of Colangelo & Lee (2020) (`DR2`, dashes) as well as our own (`RKHS`, solid). Finally, the heterogeneous response curve (`RKHS`, solid) shows that age plays a substantial role in the effectiveness of the intervention. For the youngest participants, the intervention has a small effect: employment only increases from 28% to at most 36%. For older participants, the intervention has a large effect: employment increases from 40% to 56%. It seems that 12–14 weeks of classes targeting individuals 21–23 years old may be an effective policy.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes standard identifying assumptions; extensions to counterfactual distributions and graphical models; proofs; further discussion; implementation details; and code.

REFERENCES

ABREVAYA, J., HSU, Y.-C. & LIELI, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics* **33**, 485–505.

ALTONJI, J. G. & MATZKIN, R. L. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* **73**, 1053–1102.

BERLINET, A. & THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media.

BERTHIER, R., BACH, F. & GAILLARD, P. (2020). Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems* **33**, 2576–2586.

BICKEL, P. J., KLAASSEN, C. A., RITOV, Y. & WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, vol. 4. Johns Hopkins University Press Baltimore.

CAPONNETTO, A. & DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics* **7**, 331–368.

CARRASCO, M., FLORENS, J.-P. & RENAULT, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics* **6**, 5633–5751.

CATTANEO, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* **155**, 138–154.

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. & MELLY, B. (2013). Inference on counterfactual distributions. *Econometrica* **81**, 2205–2268.

CHERNOZHUKOV, V., NEWEY, W. K. & SINGH, R. (2022). Debiased machine learning of global and local parameters using regularized Riesz representers. *The Econometrics Journal* .

CHERNOZHUKOV, V., NEWEY, W. K. & SINGH, R. (2023). A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika* **110**, 257–264.

COLANGELO, K. & LEE, Y.-Y. (2020). Double debiased machine learning nonparametric inference with continuous treatments. *arXiv:2004.03036* .

CRAVEN, P. & WAHBA, G. (1978). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377–403.

DAROLLES, S., FAN, Y., FLORENS, J.-P. & RENAULT, E. (2011). Nonparametric instrumental regression. *Econometrica* **79**, 1541–1565.

DÍAZ, I. & VAN DER LAAN, M. J. (2013). Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference* **1**, 171–192.

FAN, Q., HSU, Y.-C., LIELI, R. P. & ZHANG, Y. (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics* **40**, 313–327.

FIRPO, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* **75**, 259–276.

FISCHER, S. & STEINWART, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research* **21**, 205–1.

FLORES, C. A., FLORES-LAGUNES, A., GONZALEZ, A. & NEUMANN, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of Job Corps. *Review of Economics and Statistics* **94**, 153–171.

FOSTER, D. J. & SYRGKANIS, V. (2019). Orthogonal statistical learning. *arXiv:1901.09036* .

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

GRÜNEWÄLDER, S., GRETTON, A. & SHAWE-TAYLOR, J. (2013). Smooth operators. In *International Conference on Machine Learning*.

HERNÁN, M. A. & ROBINS, J. M. (2020). *Causal Inference*. CRC.

HIRSHBERG, D. A., MALEKI, A. & ZUBIZARRETA, J. R. (2019). Minimax linear estimation of the retargeted mean. *arXiv:1901.10296* .

HOTZ, V. J., IMBENS, G. W. & MORTIMER, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics* **125**, 241–270.

HUBER, M., HSU, Y.-C., LEE, Y.-Y. & LETTRY, L. (2020). Job Corps data. Tech. rep., Journal of Applied Econometrics Data Archive. http://qed.econ.queensu.ca/jae/datasets/hsu001/.

IMAI, K. & VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99**, 854–866.

KALLUS, N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research* **21**, 62–1.

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

KENNEDY, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv:2004.14497* .

KENNEDY, E. H., MA, Z., MCHUGH, M. D. & SMALL, D. S. (2017). Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 1229.

KIMELDORF, G. & WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33**, 82–95.

LI, K.-C. (1986). Asymptotic optimality of CL and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics* , 1101–1112.

LI, Z., MEUNIER, D., MOLLENHAUER, M. & GRETTON, A. (2022). Optimal rates for regularized conditional mean embedding learning. *arXiv:2208.01711* .

LUEDTKE, A. R. & VAN DER LAAN, M. J. (2016a). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics* **44**, 713.

LUEDTKE, A. R. & VAN DER LAAN, M. J. (2016b). Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics* **12**, 305–332.

MICCHELLI, C. A. & PONTIL, M. (2005). On learning vector-valued functions. *Neural Computation* **17**, 177–204.

MUANDET, K., KANAGAWA, M., SAENGKYONGAM, S. & MARUKATAT, S. (2021). Counterfactual mean embeddings. *Journal of Machine Learning Research* **22**, 1–71.

NEWEY, W. K. (1994a). The asymptotic variance of semiparametric estimators. *Econometrica* , 1349–1382.

NEWEY, W. K. (1994b). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* , 233–253.

NIE, X. & WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**, 299–319.

PARK, J. & MUANDET, K. (2020). A measure-theoretic approach to kernel conditional mean embeddings. *Advances in Neural Information Processing Systems* **33**, 21247–21259.

PEARL, J. & BAREINBOIM, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science* **29**, 579–595.

PILLAUD-VIVIEN, L., RUDI, A. & BACH, F. (2018). Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*.

QUIÑONERO-CANDELA, J., SUGIYAMA, M., LAWRENCE, N. D. & SCHWAIGHOFER, A. (2009). *Dataset Shift in Machine Learning*. MIT Press.

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7**, 1393–1512.

ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

RUBIN, D. & VAN DER LAAN, M. J. (2005). A general imputation methodology for nonparametric regression with censored data. Tech. rep., UC Berkeley Division of Biostatistics.

RUBIN, D. & VAN DER LAAN, M. J. (2006). Extending marginal structural models through local, penalized, and additive learning. Tech. rep., UC Berkeley Division of Biostatistics.

SCHOCHET, P. Z., BURGHARDT, J. & MCCONNELL, S. (2008). Does Job Corps work? Impact findings from the national Job Corps study. *American Economic Review* **98**, 1864–86.

SEMENOVA, V. & CHERNOZHUKOV, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* **24**, 264–289.

SINGH, R. (2021). Debiased kernel methods. *arXiv:2102.11076* .

SINGH, R., SAHANI, M. & GRETTON, A. (2019). Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*.

SINGH, R., XU, L. & GRETTON, A. (2020). Kernel methods for policy evaluation: Treatment effects, mediation analysis, and off-policy planning. *arXiv:2010.04855* .

SMALE, S. & ZHOU, D.-X. (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation* **26**, 153–172.

SMOLA, A., GRETTON, A., SONG, L. & SCHÖLKOPF, B. (2007). A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*.

SRIPERUMBUDUR, B., FUKUMIZU, K. & LANCKRIET, G. (2010). On the relation between universality, characteristic kernels and RKHS embedding of measures. In *International Conference on Artificial Intelligence and Statistics*.

TALWAI, P., SHAMELI, A. & SIMCHI-LEVI, D. (2022). Sobolev norm learning rates for conditional mean embeddings. In *International Conference on Artificial Intelligence and Statistics*.

TOLSTIKHIN, I., SRIPERUMBUDUR, B. K. & MUANDET, K. (2017). Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research* **18**, 3002–3048.

VAN DER LAAN, M. J. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics* **2**.

VAN DER LAAN, M. J. & 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. Tech. rep., UC Berkeley Division of Biostatistics.

VAN DER VAART, A. (1991). On differentiable functionals. *The Annals of Statistics* **19**, 178–204.

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM.

ZIMMERT, M. & LECHNER, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv:1908.08779* .