Kernel Methods for Policy Evaluation:
Treatment Effects, Mediation Analysis, and Off-Policy Planning

Rahul Singh
MIT Economics
rahul.singh@mit.edu

Liyuan Xu
Gatsby Unit, UCL
liyuan.jo.19@ucl.ac.uk

Arthur Gretton
Gatsby Unit, UCL
arthur.gretton@gmail.com

Abstract

We propose a novel framework for non-parametric policy evaluation in static and dynamic settings. Under the assumption of selection on observables, we consider treatment effects of the population, of sub-populations, and of alternative populations that may have alternative covariate distributions. We further consider the decomposition of a total effect into a direct effect and an indirect effect (as mediated by a particular mechanism). Under the assumption of sequential selection on observables, we consider the effects of sequences of treatments. Across settings, we allow for treatments that may be discrete, continuous, or even text. Across settings, we allow for estimation of not only counterfactual mean outcomes but also counterfactual distributions of outcomes. We unify analyses across settings by showing that all of these causal learning problems reduce to the re-weighting of a prediction, i.e. causal adjustment. We implement the re-weighting as an inner product in a function space called a reproducing kernel Hilbert space (RKHS), with a closed form solution that can be computed in one line of code. We prove uniform consistency and provide finite sample rates of convergence. We evaluate our estimators in simulations devised by other authors. We use our new estimators to evaluate continuous and heterogeneous treatment effects of the US Jobs Corps training program for disadvantaged youth.

1 Introduction

1.1 Motivation

The goal of the treatment effect literature is to determine the counterfactual relationship between treatment D and outcome Y: if we *intervened* on D, what would be the counterfactual outcome Y? *Selection on observables* is the widely-used assumption that the assignment of treatment D is as good as random after conditioning on covariates X. The counterfactual relationship between treatment D and outcome Y is not the prediction of Y given (D, X), however it can be recovered by appropriately re-weighting the prediction. Additional nuanced questions can be asked. Is the causal relationship between treatment D and outcome Y learned in one population externally valid in another? For the sub-population who received treatment $D = d$, what would have been their outcome had they instead received treatment $D = d'$? Are treatment effects heterogeneous with respect to some interpretable, low-dimensional covariate V? The goal of mediation analysis is to determine how much of the total effect of the treatment D on the outcome Y is mediated by a particular mechanism M. The goal of off-policy planning is to evaluate the effect of a sequence of treatments $D_{1:T}$ on outcome Y, even when that sequence was not actually implemented. Such empirical questions are ubiquitious in policy evaluation across economics, statistics, and epidemiology.

34th Conference on Neural Information Processing Systems: Workshop on Machine Learning for Economic Policy (ML for Economic Policy 2020), Vancouver, Canada.

1.2 Contribution

Conceptual. We unify these various learning problems into one general learning problem. The generality of the framework is threefold. First, we consider a variety of learning problems: treatment effects (Section 2), mediation analysis (Appendix B), off-policy planning (Appendix C), and graphical effects (Appendix D). Second, we consider the fully non-parametric setting. In semi-parametric causal inference, treatment D is restricted to be binary. We consider non-parametric causal inference, allowing the treatment D to be discrete, continuous, or even text. Third, we provide a unified approach for estimating not only counterfactual mean outcomes but also counterfactual distributions of outcomes (Appendix E).

Algorithmic. We impose additional structure on the general learning problem: we assume that the true causal relationship is a function in a reproducing kernel Hilbert space (RKHS), which is a popular non-parametric setting in machine learning (Appendices F, G). Guided by this additional structure, we propose a family of novel, global estimators with closed form solutions. Our estimators can be computed in one line of code. Their only hyperparameters are ridge regression penalties, which are easily tuned using the closed form solution for leave-one-out cross validation, and kernel hyperparameters, which have well-known heuristics. We evaluate our estimators in simulations devised by other authors (Section 3).

Statistical. We prove uniform consistency: our estimators converge to the true causal relationships in sup-norm, if the true causal relationship is an element of the RKHS. Moreover, we provide *finite sample* rates of convergence, explicitly accounting for the sources of error in any finite sample size. The rates do not depend on the data dimension (which may be infinite), but rather the smoothness of the true causal relationship. At this point, the analysis of uniform confidence bands remains an open question. As such, an indirect contribution is to pose this question for future econometric research.

Empirical. Our new estimators provide new insight into important questions in labor economics. Models of the labor market predict continuous and heterogeneous effects of labor market interventions. We estimate such effects for the US Job Corps, a job training program for disadvantaged youth (Section 3). In our policy analysis, we find that the effect of job training on employment substantially varies by class-hours and by age; a targeted policy will be more effective. We also find that job training reduces arrests via social mechanisms besides employment.

1.3 Causal adjustment

We describe the general learning problem that we term *causal adjustment*. This problem is different than prediction, but it will involve prediction as an intermediate step. In the simplest case, suppose an analyst is interested in the counterfactual mean outcome given treatment $D = d$, i.e. the target parameter is $\theta_0^{ATE}(d) := \mathbb{E}[Y^{(d)}]$, where $Y^{(d)}$ is the potential outcome given the intervention $D = d$. Though we observe outcome Y, we seek to infer means (and distributions) of counterfactual outcomes $\{Y^d\}$ for different populations. In the main text we use potential outcomes, and in Appendix D we present corresponding results for causal directed acyclic graphs (DAGs).

Suppose the analyst observes outcome Y, treatment D, and covariates X. We denote the prediction $\gamma_0(d, x) := \mathbb{E}[Y|D = d, X = x]$. Intuitively, an appropriate formula for the target parameter is $\theta_0^{ATE}(d) = \int \gamma_0(d, x)\mathbb{Q}$, where $\mathbb{Q}$ is a re-weighting that adjusts for confounding due to X. As we will see, the appropriate adjustment for confounding $\mathbb{Q}$ depends on the target parameter and causal setting. It may be as simple as the marginal distribution $\mathbb{Q} = \mathbb{P}(x)$, as when estimating treatment effects for the full population under selection on observables. It may be a complicated product of conditional distributions, as in off-policy planning under sequential selection on observables.

We impose that the prediction γ_0 is an element of a function space called a reproducing kernel Hilbert space (RKHS) over treatment D and covariates X. For a sufficiently rich RKHS, the complicated distribution $\mathbb{Q}$ can also be represented as a vector μ in the RKHS. The injective mapping $\mathbb{Q} \mapsto \mu$ is called the kernel mean embedding, and it preserves all information in $\mathbb{Q}$. As such, the target parameter can be represented as an inner product in the RKHS: $\theta_0^{ATE}(d) = \langle \gamma_0, \mu \rangle$. We estimate $\hat{\gamma}$ and $\hat{\mu}$ by kernel ridge regressions, leading to an estimator of the form $\hat{\theta}^{ATE}(d) = \langle \hat{\gamma}, \hat{\mu} \rangle$. More generally, we use this technique to provide estimators of the causal parameters defined in Table 1 as well as their distributional generalizations.

1.4 Related work

In the causal inference literature, we build on canonical identification theorems for treatment effects [121], mediation analysis [82], and off-policy planning [118]. Our non-parametric framework for causal inference views target parameters as re-weightings of a prediction. This perspective generalizes the widely-accepted semi-parametric framework [26], which views target parameters as scalar summaries of a prediction (and localizations thereof [92, 28, 32]).

We restrict attention to settings where treatment assignment is *conditionally exogenous*. An important literature in econometrics instead considers settings in which treatment assignment is *endogenous*: treatment D reflects the optimal choice of a rational agent with expectations about their potential outcomes $\{Y^{(d)}\}$. [71] presents a unified perspective of policy evaluation with endogenous (binary) treatment, reducing many learning problems to different re-weightings of an underlying philosophical quantity called the marginal treatment effect. We leave to future research the extension of our methods to such settings.

Our non-parametric approach differs from existing, non-parametric approaches in econometrics. *De-biased machine learning* (DML) presents semi-parametric estimators using black-box machine learning of the prediction and an explicit [25] or implicit [10, 29] bias correction. Recent work extends DML to non-parametric learning problems by introducing an additional step of Nadaraya-Watson local smoothing or series regression [92, 126, 28, 42, 150, 32]. Impressively, DML approaches require minimal assumptions. DML analyses do require sufficiently fast rates of black-box prediction performance, which may in turn require additional structure on the learning problem such as approximate sparsity.

Innovative work on random forests permits estimation of conditional average treatment effect (CATE), requiring Lipschitz continuity or sparsity [145, 110]. Drawing inspiration from these works, we too adapt machine learning methods to causal inference and impose additional structure on the learning problem. Whereas the definition of CATE in these works is conditional on the entire covariate vector X, we pursue a more general definition of CATE conditional on some interpretable subvector $V \subset X$. We most directly build on [129], which presents an RKHS approach for non-parametric instrumental variable regression. Our work is complementary in that we consider a much broader variety of causal parameters. Appendix A provides a detailed comparison of the present work with existing work.

2 Treatment effect

2.1 Learning problem

A treatment effect is a statement about the counterfactual outcome $Y^{(d)}$ given a hypothetical intervention on treatment $D = d$. The causal inference literature aims to measure a rich variety of treatment effects with nuanced interpretation. We define these treatment effects below, which are also discussed in [72, 73].

Definition 2.1 (Treatment effects)**.** *We define the following treatment effects*

1. $\theta_0^{ATE}(d) := \mathbb{E}[Y^{(d)}]$ *is the counterfactual mean outcome given intervention* $D = d$ *for the entire population*
2. $\theta_0^{DS}(d, \tilde{\mathbb{P}}) := \mathbb{E}_{\tilde{\mathbb{P}}}[Y^{(d)}]$ *is the counterfactual mean outcome given intervention* $D = d$ *for an alternative population with data distribution* $\tilde{\mathbb{P}}$
3. $\theta_0^{ATT}(d, d') := \mathbb{E}[Y^{(d')}|D = d]$ *is the counterfactual mean outcome given intervention* $D = d'$ *for the sub-population who actually received treatment* $D = d$
4. $\theta_0^{CATE}(d, v) := \mathbb{E}[Y^{(d)}|V = v]$ *is the counterfactual mean outcome given intervention* $D = d$ *for the sub-population with covariate value* $V = v$

$\theta_0^{ATE}(d)$ generalizes the notion of average treatment effect (ATE). Average treatment effect of a binary treatment $D \in \{0, 1\}$ is $\mathbb{E}[Y^{(1)} - Y^{(0)}]$. The analyst is essentially estimating a 2-vector of counterfactual mean outcomes $(\mathbb{E}[Y^{(0)}], \mathbb{E}[Y^{(1)}])$, where the length of the vector is the cardinality of the support of treatment D. For treatment that is more generally discrete, continuous, or even text,

the vector $\theta_0^{ATE}(d) := \mathbb{E}[Y^{(d)}]$ may be infinite-dimensional. The fact that the target parameter is infinite-dimensional is what makes this problem fully non-parametric rather than semi-parametric. This quantity is also called the continuous treatment effect or dose-response curve.

The next target parameter $\theta_0^{DS}(d, \tilde{\mathbb{P}})$ gets to the heart of external validity: though our data were drawn from population $\mathbb{P}$, what would be the treatment effect for a different population $\tilde{\mathbb{P}}$? For example, a study may be conducted in Virginia, and we may want to use those results to inform policy in Arkansas, a state with a different demographic composition [77]. Questions of this nature are widely studied in machine learning under the names of transfer learning, distribution shift, and covariate shift. We contend such questions are equally important in social science.

$\theta_0^{ATE}(d)$ and $\theta_0^{DS}(d, \tilde{\mathbb{P}})$ capture the concept of average treatment effect for the entire population, but treatment effects may be heterogeneous for different sub-populations. Towards the goal of personalized or targeted interventions, an analyst may ask another nuanced counterfactual question: what would have been effect of treatment $D = d'$ for the sub-population who actually received treatment $D = d$? In semi-parametrics, average treatment on the treated (ATT) of a binary treatment $D \in \{0, 1\}$ is $\mathbb{E}[Y^{(1)} - Y^{(0)}|D = 1]$. Each term is a special case of the more general quantity $\theta_0^{ATT}(d, d') := \mathbb{E}[Y^{(d')}|D = d]$, which allows for discrete, continuous, or even text treatment values.

In $\theta_0^{ATT}(d, d')$, heterogeneity is indexed by treatment D. Heterogeneity may instead be indexed by some covariate subset V. An analyst may therefore prefer to measure heterogeneous treatment effects for sub-populations characterized by different values of some interpretable sub-vector V, e.g. age, race, or gender. For simplicity, we will write covariates as (V, X) for this setting. In semi-parametrics, conditional average treatment effect (CATE) of a binary treatment $D \in \{0, 1\}$ is $\mathbb{E}[Y^{(1)} - Y^{(0)}|V = v]$. $\theta_0^{CATE}(d, v) := \mathbb{E}[Y^{(d)}|V = v]$ generalizes this idea, allowing for treatment D that may be discrete, continuous, or text. This quantity is also called the heterogeneous treatment effect.

Clearly, an analyst may ask more even nuanced questions, combining elements of the treatment effects defined above. The analysis remains the same. An analyst may also ask about counterfactual *distributions* of outcomes rather than counterfactual *mean* outcomes. We reserve distributional analysis for Appendix E. Pleasingly, the analysis remains the same, via conditional mean embeddings.

2.2 Identification

In the seminal work [121], the authors state sufficient conditions under which treatment effects–philosophical quantities defined in terms of potential outcomes $\{Y^{(d)}\}$–can be measured from empirical quantities such as outcomes Y, treatments D, and covariates (V, X). Colloquially, this collection of sufficient conditions is known as *selection on observables*.

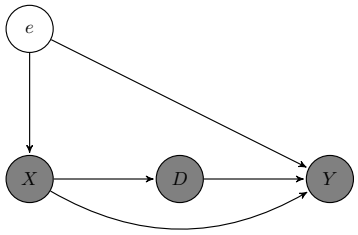


Figure 1: Selection on observables DAG

Assumption 2.1 (Selection on observables)**.** *Assume*

1. *No interference: if $D = d$ then $Y = Y^{(d)}$*
2. *Conditional exchangeability: $\{Y^{(d)}\} \perp\!\!\!\perp D|X$*
3. *Overlap: if $f(x) > 0$ then $f(d|x) > 0$*

where $f(x)$ and $f(d|x)$ are densities. For θ_0^{CATE}, replace X with (V, X).

No interference is also called the stable unit treatment value assumption (SUTVA). It rules out network effects, also called spillovers. Conditional exchangeability states that conditional on covariates X, treatment assignment is as good as random. Overlap ensures that there is no covariate stratum $X = x$ such that treatment has a restricted support. The overlap condition resembles the support condition of [85].

To handle θ_0^{DS}, we also make a standard assumption in transfer learning.

Assumption 2.2 (Distribution shift)**.** *The difference in population distributions $\mathbb{P}$ and $\tilde{\mathbb{P}}$ is only in the distribution of treatments and covariates.*

$$\mathbb{P}(Y, D, X) = \mathbb{P}(Y|D, X)\mathbb{P}(D, X), \quad \tilde{\mathbb{P}}(Y, D, X) = \mathbb{P}(Y|D, X)\tilde{\mathbb{P}}(D, X)$$

An immediate consequence is that the prediction function $\gamma_0(d,x) = \mathbb{E}[Y|D=d, X=x]$ remains the same across the different populations $\mathbb{P}$ and $\tilde{\mathbb{P}}$. Formally, the theorem that uses these assumptions to express treatment effects in terms of data is known as an identification result. We quote a classic identification result below. Define the predictions

$$\gamma_0(d,x) := \mathbb{E}[Y|D=d, X=x], \quad \gamma_0(d,v,x) := \mathbb{E}[Y|D=d, V=v, X=x]$$

Theorem 2.1 (Identification of treatment effects [121])**.** *If Assumption 2.1 holds then*

1. $\theta_0^{ATE}(d) = \int \gamma_0(d,x)\mathbb{P}(x)$
2. *If in addition Assumption 2.2 holds, then* $\theta_0^{DS}(d,\tilde{\mathbb{P}}) = \int \gamma_0(d,x)\tilde{\mathbb{P}}(x)$
3. $\theta_0^{ATT}(d,d') = \int \gamma_0(d',x)\mathbb{P}(x|d)$
4. $\theta_0^{CATE}(d,v) = \int \gamma_0(d,v,x)\mathbb{P}(x|v)$

2.3 Algorithm

Theorem 2.1 makes precise how each treatment effect is identified as a quantity of the form $\int \gamma_0(d,x)\mathbb{Q}$ for some distribution $\mathbb{Q}$. We now assume that γ_0 is an element of a function space called a reproducing kernel Hilbert space (RKHS), which is a dense subset of L^2. See Appendices F, G for background on the RKHS, which is a canonical setting for machine learning.

In our construction, we define scalar-valued RKHSs for treatments D and covariates V and X, then assume that the prediction 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 measureable 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

$$\phi_{\mathcal{D}} : \mathcal{D} \to \mathcal{H}_{\mathcal{D}}, \quad d \mapsto k_{\mathcal{D}}(d,\cdot) \quad \phi_{\mathcal{V}} : \mathcal{V} \to \mathcal{H}_{\mathcal{V}}, \quad v \mapsto k_{\mathcal{V}}(v,\cdot) \quad \phi_{\mathcal{X}} : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}, \quad x \mapsto k_{\mathcal{X}}(x,\cdot)$$

To lighten notation, we will suppress subscripts when arguments are provided, e.g. we will write $\phi(d) = \phi_{\mathcal{D}}(d)$.

For θ_0^{ATE}, θ_0^{DS}, and θ_0^{ATT}, we assume the prediction γ_0 is an element of the RKHS with tensor product feature map $\phi(d) \otimes \phi(x)$, i.e. $\gamma_0 \in \mathcal{H} := \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}}$. In this construction, we appeal to the fact that the product of positive definite kernels corresponding to $\mathcal{H}_{\mathcal{D}}$ and $\mathcal{H}_{\mathcal{X}}$ defines a new positive definite kernel corresponding to $\mathcal{H}$. We choose the product construction because it is provides a rich composite basis. Therefore by the reproducing property, $\gamma_0(d,x) = \langle \gamma_0, \phi(d) \otimes \phi(x) \rangle_{\mathcal{H}}$. Likewise for θ_0^{CATE} we assume $\gamma_0 \in \mathcal{H} := \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{V}} \otimes \mathcal{H}_{\mathcal{X}}$. With this RKHS construction, we obtain a representation of the target parameters as inner products in the space $\mathcal{H}$.

Theorem 2.2 (Representation of treatment effects)**.** *If* $\gamma_0 \in \mathcal{H}$ *then*

1. $\theta_0^{ATE}(d) = \langle \gamma_0, \phi(d) \otimes \mu_x \rangle_{\mathcal{H}}$ *where* $\mu_x := \int \phi(x)\mathbb{P}(x)$
2. $\theta_0^{DS}(d,\tilde{\mathbb{P}}) = \langle \gamma_0, \phi(d) \otimes \nu_x \rangle_{\mathcal{H}}$ *where* $\nu_x := \int \phi(x)\tilde{\mathbb{P}}(x)$
3. $\theta_0^{ATT}(d,d') = \langle \gamma_0, \phi(d') \otimes \mu_x(d) \rangle_{\mathcal{H}}$ *where* $\mu_x(d) := \int \phi(x)\mathbb{P}(x|d)$
4. $\theta_0^{CATE}(d,v) = \langle \gamma_0, \phi(d) \otimes \phi(v) \otimes \mu_x(v) \rangle_{\mathcal{H}}$ *where* $\mu_x(v) := \int \phi(x)\mathbb{P}(x|v)$

The quantity $\mu_x := \int \phi(x)\mathbb{P}(x)$ is called the mean embedding of $\mathbb{P}(x)$. It encodes the distribution $\mathbb{P}(x)$ as a vector $\mu_x \in \mathcal{H}_{\mathcal{X}}$. For some quantities, we have a conditional mean embedding rather than a mean embedding. For example in $\theta_0^{ATT}(d,d')$, we require $\mu_x(d)$. In such cases, we will estimate the conditional mean embedding by an appropriately defined kernel ridge regression to get $\hat{\mu}_x(\cdot)$. This is the key innovation of the present work, and the reason why our estimators have closed form solutions despite complicated causal adjustment. See Appendix H for details.

While these representations appear abstract, they are in fact eminently useful for defining estimators with closed form solutions that can be computed in one line of code (after computing kernel matrices). For example for $\theta_0^{ATE}(d)$, our estimator will be $\hat{\theta}^{ATE}(d) = \langle \hat{\gamma}, \phi(d) \otimes \hat{\mu}_x \rangle_{\mathcal{H}}$. The estimator $\hat{\gamma}$ is a standard kernel ridge regression, with known closed form. The estimator $\hat{\mu}_x$ is an empirical mean.

Algorithm 2.1 (Estimation of treatment effects)**.** *Denote the empirical kernel matrices*

$$K_{DD} \in \mathbb{R}^{n \times n}, \quad K_{VV} \in \mathbb{R}^{n \times n} \quad K_{XX} \in \mathbb{R}^{n \times n}$$

K_{DD}, K_{VV}, and K_{XX} are calculated from observations drawn from population $\mathbb{P}$. We denote by $\{\tilde{x}_i\}_{i \in [\tilde{n}]}$ observations drawn from population $\tilde{\mathbb{P}}$. Denote by $\odot$ the element-wise product. Treatment effect estimators have the closed form solutions

1. $\hat{\theta}^{ATE}(d) = \frac{1}{n} \sum_{i=1}^{n} Y^T (K_{DD} \odot K_{XX} + n\lambda)^{-1} (K_{Dd} \odot K_{Xx_i})$
2. $\hat{\theta}^{DS}(d, \tilde{\mathbb{P}}) = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} Y^T (K_{DD} \odot K_{XX} + n\lambda)^{-1} (K_{Dd} \odot K_{X\tilde{x}_i})$
3. $\hat{\theta}^{ATT}(d, d') = Y^T (K_{DD} \odot K_{XX} + n\lambda)^{-1} (K_{Dd'} \odot [K_{XX}(K_{DD} + n\lambda_1)^{-1} K_{Dd}])$
4. $\hat{\theta}^{CATE}(d, v) = Y^T (K_{DD} \odot K_{VV} \odot K_{XX} + n\lambda)^{-1} (K_{Dd} \odot K_{Vv} \odot K_{XX}(K_{VV} + n\lambda_2)^{-1} K_{Vv})$

where $(\lambda, \lambda_1, \lambda_2)$ are ridge regression penalty hyper-parameters.

We give theoretical values for $(\lambda, \lambda_1, \lambda_2)$ in Appendix I that balance bias and variance. We give a practical tuning procedure in Appendix J based on leave-one-out cross validation.

2.4 Consistency

Towards a guarantee of uniform consistency, we place regularity conditions on the original spaces and scalar-valued RKHSs

Assumption 2.3 (Original space regularity conditions)**.** *Assume*

1. *$\mathcal{D}$, $\mathcal{V}$, and $\mathcal{X}$ are Polish spaces, i.e. separable and completely metrizable topological spaces*
2. *Y is bounded, i.e. $\exists C < \infty$ such that $|Y| \leq C$ almost surely*

A Polish space may be low-, high-, or infinite-dimensional. Random variables with support in a Polish space may be discrete, continuous, or even text. For simplicity of argument, we require that outcome $Y \in \mathbb{R}$ is bounded.

Assumption 2.4 (RKHS regularity conditions)**.** *Assume*

1. *$k_{\mathcal{D}}$, $k_{\mathcal{V}}$, and $k_{\mathcal{X}}$ are continuous and bounded:*

$$\sup_{d \in \mathcal{D}} \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \leq \kappa_d, \quad \sup_{v \in \mathcal{V}} \|\phi(v)\|_{\mathcal{H}_{\mathcal{V}}} \leq \kappa_v, \quad \sup_{x \in \mathcal{X}} \|\phi(x)\|_{\mathcal{H}_{\mathcal{X}}} \leq \kappa_x$$

2. *$\phi(d)$, $\phi(v)$, and $\phi(x)$ are measurable*
3. *$k_{\mathcal{X}}$ is characteristic*

Commonly used kernels are continuous and bounded. Measurability is a similarly weak condition. The characteristic property ensures injectivity of the mean embeddings, and hence uniqueness of the RKHS representation [134].

Next, we assume the prediction γ_0 is smooth.

Assumption 2.5 (Smoothness of prediction)**.** *Assume*

1. *the prediction is well-specified, i.e. $\gamma_0 \in \mathcal{H}$*
2. *the prediction is a particularly smooth element of $\mathcal{H}$. Formally, define the covariance operator T for $\mathcal{H}$. We assume $\exists g \in \mathcal{H}$ s.t. $\gamma_0 = T^{\frac{c-1}{2}} g$, $c \in (1, 2]$, and $\|g\|_{\mathcal{H}}^2 \leq \zeta$*

For θ_0^{ATE}, θ_0^{DS}, and θ_0^{ATT}, $T := \mathbb{E}[\{\phi(D) \otimes \phi(X)\} \otimes \{\phi(D) \otimes \phi(X)\}]$. For θ_0^{CATE}, $T := \mathbb{E}[\{\phi(D) \otimes \phi(V) \otimes \phi(X)\} \otimes \{\phi(D) \otimes \phi(V) \otimes \phi(X)\}]$.

We discuss this assumption in Appendix F. For learning problems with conditional mean embeddings, we place further smoothness conditions on the corresponding conditional expectation operators.

Assumption 2.6 (Smoothness for θ_0^{ATT})**.** *Assume*

1. *the conditional expectation operator E_1 is well-specified as a Hilbert-Schmidt operator between RKHSs, i.e. $E_1 \in \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D})$, where $E_1 : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{D},\ f(\cdot) \mapsto \mathbb{E}[f(X)|D = \cdot]$*
2. *the conditional expectation operator is a particularly smooth element of $\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D})$. Formally, define the covariance operator $T_1 := \mathbb{E}[\phi(D) \otimes \phi(D)]$ for $\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D})$. We assume $\exists G_1 \in \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D})$ s.t. $E_1 = (T_1)^{\frac{c_1-1}{2}} \circ G_1$, $c_1 \in (1, 2]$, and $\|G_1\|^2_{\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D})} \leq \zeta_1$*

Assumption 2.7 (Smoothness for θ_0^{CATE})**.** *Assume*

1. *the conditional expectation operator E_2 is well-specified as a Hilbert-Schmidt operator between RKHSs, i.e. $E_2 \in \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{V})$, where $E_2 : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{V},\ f(\cdot) \mapsto \mathbb{E}[f(X)|V = \cdot]$*
2. *the conditional expectation operator is a particularly smooth element of $\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{V})$. Formally, define the covariance operator $T_2 := \mathbb{E}[\phi(V) \otimes \phi(V)]$ for $\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{V})$. We assume $\exists G_2 \in \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{V})$ s.t. $E_2 = (T_2)^{\frac{c_2-1}{2}} \circ G_2$, $c_2 \in (1, 2]$, and $\|G_2\|^2_{\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{V})} \leq \zeta_2$*

Likewise, we discuss these assumptions in Appendix F. Under these conditions, we arrive at our first main result.

Theorem 2.3 (Consistency)**.** *Suppose Assumptions 2.1, 2.3, 2.4, and 2.5 hold.*

1. *Then*
$$\|\hat{\theta}^{ATE} - \theta_0^{ATE}\|_\infty = O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}}\right)$$
2. *If in addition Assumption 2.2 holds, then*
$$\|\hat{\theta}^{DS}(\cdot, \tilde{\mathbb{P}}) - \theta_0^{DS}(\cdot, \tilde{\mathbb{P}})\|_\infty = O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + \tilde{n}^{-\frac{1}{2}}\right)$$
3. *If in addition Assumption 2.6 holds, then*
$$\|\hat{\theta}^{ATT} - \theta_0^{ATT}\|_\infty = O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + n^{-\frac{1}{2}\frac{c_1-1}{c_1+1}}\right)$$
4. *If in addition Assumption 2.7 holds, then*
$$\|\hat{\theta}^{CATE} - \theta_0^{CATE}\|_\infty = O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + n^{-\frac{1}{2}\frac{c_2-1}{c_2+1}}\right)$$

Exact finite sample rates are given in Appendix I. These rates are at best $n^{-\frac{1}{6}}$ by setting $(c, c_1, c_2) = 2$. The slow rates reflect the challenge of a sup-norm guarantee, which is much stronger than a prediction guarantee, and which encodes caution about worst-case scenarios when informing policy decisions. Our analysis is agnostic about the spectral decay; further assumptions on spectral decay (interpretable as effective dimension) will lead to faster rates of prediction.

3 Experiments

3.1 Simulation

We evaluate the empirical performance of our estimators on various designs with varying sample sizes. The continuous treatment effect design [32] involves learning the counterfactual function $\theta_0^{ATE}(d) = 1.2d + d^2$. A single observation consists of the triple (Y, D, X) for outcome, treatment, and covariates where $Y, D \in \mathbb{R}$ and $X \in \mathbb{R}^{100}$. In addition to our estimator (`MeanEmb`), we implement the estimators of [92] (`DML1`) and [32] (`DML2`), which involve Nadaraya-Watson smoothing around debiased machine learning (DML). A DML estimator for a treatment effect consists of two components: the *plug-in* component and the *inverse-propensity* component, which are estimators in their own right. We implement the plug-in (`PlugIn`) and inverse-propensity (`IPW`) components of [32] as well. For each design, sample size, and algorithm, we implement 20 simulations and calculate MSE with respect to the true counterfactual function. Figure 2 visualizes results. See Appendix K for implementation details and additional designs.

3.2 Application: US Job Corps

We estimate the effects of the US Jobs Corps program, a job-training program for disadvantaged youth that operated in the mid-1990s. We initially consider the effect of training on employment similar to [32]. In this setting, the outcome $Y \in \mathbb{R}$ is the proportion of weeks employed, and the treatment $D \in \mathbb{R}$ is total hours spent in academic or vocational classes. 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. As in [32], we focus on the $n = 2,989$ observations for which $D > 40$ and $Y > 0$, i.e. individuals who completed at least one week of training and who found employment. Whereas [32] present estimates for θ_0^{ATE}, we present results for θ_0^{ATE} and θ_0^{CATE}. $\theta_0^{ATE}(d)$ is the continuous treatment effect of class-hours d on employment, while $\theta_0^{CATE}(d, v)$ is the heterogeneous treatment effect of class-hours d on employment for individuals with age v. Figure 3 visualizes results.

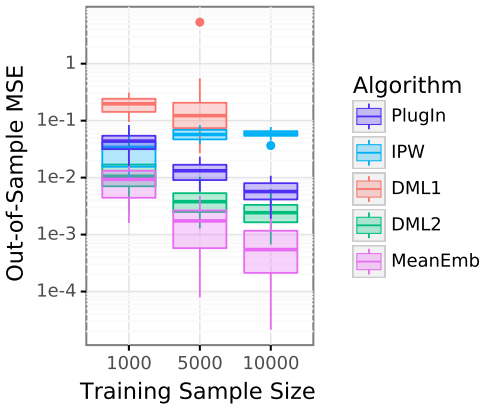
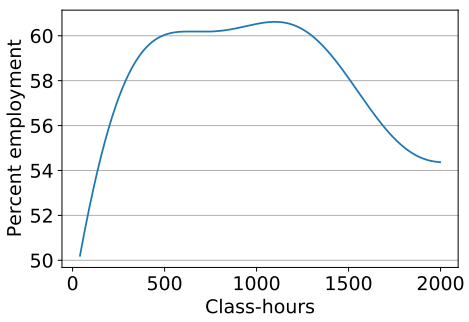
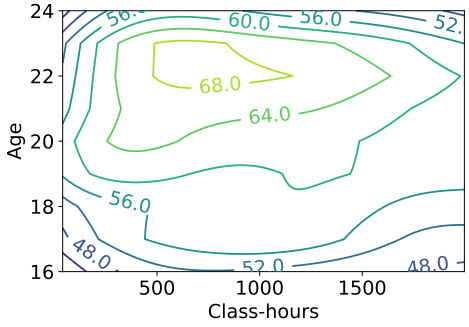


Figure 2: Continuous treatment effect simulation

The continuous treatment effect plateaus at $d = 500$ and achieves its maximum at $d = 1,200$, corresponding to 12.5 and 30 weeks of classes, respectively. From a policy perspective, the first 12.5 weeks of classes confer the greatest gain in employment: from 50% employment to more than 60% employment for the average participant. Our estimate has the same shape but is smoother than that of [32]. The heterogeneous treatment effect 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 48% to at most 52%. For older participants, the intervention has a large effect: employment increases from 52% to 68%. Our policy recommendation is therefore 12-14 weeks of classes targeting individuals who are 21-23 years old. In Appendix K, we investigate the effect of class-hours on arrests as mediated through employment.



(a) Continuous treatment effect



(b) Heterogeneous treatment effect

Figure 3: Effect of job training on employment

4 Conclusion

We propose a family of novel estimators for non-parametric policy evaluation in static and dynamic settings: treatment effects, mediation analysis, off-policy planning, graphical effects, and distributional generalization thereof. Our estimators are easily implemented and uniformly consistent. As a contribution to the policy evaluation literature, we show how to answer nuanced causal questions under the assumption that the true causal relationships are smooth. As a contribution to the kernel methods literature, we show how the RKHS is well-suited to causal inference. Our estimators perform well in simulations devised by other authors. We use our new estimators to evaluate a large scale labor market intervention and to inform policy recommendations. Our results suggest that RKHS methods may be an effective bridge between econometrics and machine learning.

Broader impacts, relevance, and originality

The goal of this project is to provide policymakers with a new toolkit for policy evaluation. We use the new toolkit to perform a policy evaluation of the US Jobs Corps program, which provided job training for disadvantaged youth. In this sense, the beneficiaries of social and economic policies will benefit from this research. Importantly, we articulate the exact assumptions that must hold for the policy evaluations to be valid. These assumptions do allow for biases in the data, of a certain type: selection bias with respect to observable confounders.

These assumptions would be violated in the presence of selection bias with respect to unobservable confounders. In that case, our statistical guarantees no longer hold. A potential danger is that practitioners apply our methods nonetheless, thereby misleading policy decisions. The populations whom the social and economic policies are meant to benefit could be inadvertently harmed instead. For this reason, we urge critical and rigorous examination of whether the key assumptions hold.

The submitted work has not been published. It was not made available until October 2020. It is original work.

Acknowledgments and disclosure of funding

We are grateful to Alberto Abadie, Anna Mikusheva, and Whitney Newey for helpful comments. We are grateful to Ying-Ying Lee, Robert Lieli, and Vira Semenova for sharing code.

References

[1] Alberto Abadie and Matias D Cattaneo. Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503, 2018.

[2] Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

[3] Jaap H Abbring. The event-history approach to program evaluation. *Advances in Econometrics*, 21:33–55, 2008.

[4] Jaap H Abbring and James J Heckman. Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. *Handbook of Econometrics*, 6:5145–5303, 2007.

[5] Jaap H Abbring and Gerard J Van den Berg. The nonparametric identification of treatment effects in duration models. *Econometrica*, 71(5):1491–1517, 2003.

[6] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[7] Hunt Allcott. Site selection bias in program evaluation. *The Quarterly Journal of Economics*, 130(3):1117–1165, 2015.

[8] Mathilde Almlund, Angela Lee Duckworth, James Heckman, and Tim Kautz. Personality, psychology, and economics. In *Handbook of the Economics of Education*, volume 4, pages 1–181. Elsevier, 2011.

[9] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press, 2008.

[10] Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

[11] Francis R Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *ICML*, 2012.

[12] Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986.

[13] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

[14] Marianne P Bitler, Jonah B Gelbach, and Hilary W Hoynes. What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012, 2006.

[15] Richard W Blundell and James L Powell. Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679, 2004.

[16] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[17] Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4(4):377–408, 2006.

[18] Marine Carrasco. A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398, 2012.

[19] Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, 6:5633–5751, 2007.

[20] Marine Carrasco and Barbara Rossi. In-sample inference and forecasting in misspecified factor models. *Journal of Business & Economic Statistics*, 34(3):313–338, 2016.

[21] Marine Carrasco and Guy Tchuente. Regularized LIML for many instruments. *Journal of Econometrics*, 186(2):427–442, 2015.

[22] Matias D Cattaneo. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138–154, 2010.

[23] Gary Chamberlain. Panel data. *Handbook of Econometrics*, 2:1247–1318, 1984.

[24] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Conference on Uncertainty in Artificial Intelligence*, pages 109–116, 2010.

[25] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), 2018.

[26] Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv:1608.00033*, 2016.

[27] Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[28] Victor Chernozhukov, Whitney Newey, and Rahul Singh. De-biased machine learning of global and local parameters using regularized Riesz representers. *arXiv:1802.08667*, 2018.

[29] Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *arXiv:1809.05224*, 2018.

[30] Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. The difficult task of distribution generalization in nonlinear models. *arXiv:2006.07433*, 2020.

[31] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems*, pages 4412–4420, 2016.

[32] Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv:2004.03036*, 2020.

[33] Thomas D Cook, Donald Thomas Campbell, and Arles Day. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, volume 351. Houghton Mifflin Boston, 1979.

[34] Felipe Cucker and Steve Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428, 2002.

[35] Flavio Cunha, James J Heckman, and Susanne M Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.

[36] Janet Currie and Douglas Almond. Human capital development before age five. In *Handbook of Labor Economics*, volume 4, pages 1315–1486. Elsevier, 2011.

[37] Rhian M Daniel, Bianca L De Stavola, and Simon N Cousens. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal*, 11(4):479–517, 2011.

[38] Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

[39] Mert Demirer, Vasilis Syrgkanis, Greg Lewis, and Victor Chernozhukov. Semi-parametric efficient policy learning with continuous actions. *arXiv:1905.10116*, 2019.

[40] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 1996.

[41] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *arXiv:2006.07201*, 2020.

[42] Qingliang Fan, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. Estimation of conditional average treatment effects with high-dimensional data. *arXiv:1908.02399*, 2019.

[43] Helmut Farbmacher, Martin Huber, Henrika Langen, and Martin Spindler. Causal mediation analysis with double machine learning. *arXiv:2002.12710*, 2020.

[44] Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276, 2007.

[45] Bernd Fitzenberger, Aderonke Osikominu, and Robert Völter. Get training or wait? Long-run employment effects of training programs for the unemployed in West Germany. *Annales d'Economie et de Statistique*, pages 321–355, 2008.

[46] Jean-Pierre Florens, James J Heckman, Costas Meghir, and Edward Vytlacil. Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica*, 76(5):1191–1206, 2008.

[47] Jean-Pierre Florens and Michel Mouchart. A note on noncausality. *Econometrica*, pages 583–591, 1982.

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

[49] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv:1901.09036*, 2019.

[50] Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Bharath K Sriperumbudur. Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems*, pages 473–480, 2009.

[51] Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14(1):3753–3783, 2013.

[52] Antonio F Galvao and Liang Wang. Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, 110(512):1528–1542, 2015.

[53] Thomas Gärtner, Peter Flach, and Stefan Wrobel. On graph kernels: Hardness results and efficient alternatives. In *Learning Theory and Kernel Machines*, pages 129–143. Springer, 2003.

[54] Richard D Gill and James M Robins. Causal inference for complex longitudinal data: The continuous case. *Annals of Statistics*, pages 1785–1811, 2001.

[55] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, pages 424–438, 1969.

[56] Arthur Gretton. RKHS in machine learning: Testing statistical dependence. Technical report, UCL Gatsby Unit, 2018.

[57] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[58] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, 3(4):5, 2009.

[59] Steffen Grünewälder, Arthur Gretton, and John Shawe-Taylor. Smooth operators. In *International Conference on Machine Learning*, pages 1184–1192, 2013.

[60] Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *International Conference on Machine Learning*, volume 5, 2012.

[61] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. A distribution-free theory of nonparametric regression, 2002.

[62] Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, pages 1–12, 1943.

[63] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

[64] James Heckman, Rodrigo Pinto, and Peter Savelyev. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, 103(6):2052–86, 2013.

[65] James J Heckman. Sample selection bias as a specification error. *Econometrica*, pages 153–161, 1979.

[66] James J Heckman and Bo E Honore. The empirical content of the Roy model. *Econometrica*, pages 1121–1149, 1990.

[67] James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

[68] James J Heckman and Salvador Navarro. Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics*, 136(2):341–396, 2007.

[69] James J Heckman and Rodrigo Pinto. Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs. *Econometric Reviews*, 34(1-2):6–31, 2015.

[70] James J Heckman and Edward Vytlacil. Policy-relevant treatment effects. *American Economic Review*, 91(2):107–111, 2001.

[71] James J Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738, 2005.

[72] James J Heckman and Edward J Vytlacil. Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. *Handbook of Econometrics*, 6:4779–4874, 2007.

[73] James J Heckman and Edward J Vytlacil. Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *Handbook of Econometrics*, 6:4875–5143, 2007.

[74] Miguel A Hernán and James M Robins. Causal inference, 2010.

[75] Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

[76] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[77] V Joseph Hotz, Guido W Imbens, and Julie H Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1-2):241–270, 2005.

[78] V Joseph Hotz and Robert A Miller. An empirical analysis of life cycle fertility and female labor supply. *Econometrica*, pages 91–118, 1988.

[79] V Joseph Hotz and Robert A Miller. Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, 60(3):497–529, 1993.

[80] Martin Huber, Yu-Chin Hsu, Ying-Ying Lee, and Layal Lettry. Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics*, 2020.

[81] Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309, 2010.

[82] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, pages 51–71, 2010.

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

[84] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

[85] Guido W Imbens and Whitney K Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

[86] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[87] Martin Emil Jakobsen and Jonas Peters. Distributional robustness of K-class estimators and the PULSE. *arXiv:2005.03353*, 2020.

[88] Charles M Judd and David A Kenny. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5(5):602–619, 1981.

[89] Nathan Kallus and Masatoshi Uehara. Doubly robust off-policy value and gradient estimation for deterministic policies. *arXiv:2006.03900*, 2020.

[90] Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International Conference on Artificial Intelligence and Statistics*, pages 1243–1251, 2018.

[91] Masahiro Kato, Masatoshi Uehara, and Shota Yasui. Off-policy evaluation and learning for external validity under a covariate shift. *arXiv:2002.11642*, 2020.

[92] Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229, 2017.

[93] Eric I Knudsen, James J Heckman, Judy L Cameron, and Jack P Shonkoff. Economic, neurobiological, and behavioral perspectives on building America's future workforce. *Proceedings of the National Academy of Sciences*, 103(27):10155–10162, 2006.

[94] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[95] Michael Lechner and Ruth Miquel. Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics*, 39(1):111–137, 2010.

[96] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444, 2002.

[97] Judith J Lok. Statistical modeling of causal effects in continuous time. *The Annals of Statistics*, 36(3):1464–1507, 2008.

[98] Thierry Magnac and David Thesmar. Identifying dynamic discrete decision processes. *Econometrica*, 70(2):801–816, 2002.

[99] Charles F Manski. Dynamic choice in social settings: Learning from the experiences of others. *Journal of Econometrics*, 58(1-2):121–136, 1993.

[100] Rosa L Matzkin. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica*, pages 239–270, 1992.

[101] Shahar Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4(Oct):759–771, 2003.

[102] Breed D Meyer. Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, 13(2):151–161, 1995.

[103] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

[104] Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *arXiv:1805.08845*, 2018.

[105] Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

[106] Jerzy Neyman and Karolina Iwaszkiewicz. Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 2(2):107–180, 1935.

[107] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv:1712.04912*, 2017.

[108] Yu Nishiyama, Abdeslam Boularias, Arthur Gretton, and Kenji Fukumizu. Hilbert space embeddings of POMDPs. In *Conference on Uncertainty in Artificial Intelligence*, 2012.

[109] Andriy Norets and Xun Tang. Semiparametric inference in dynamic binary choice models. *Review of Economic Studies*, 81(3):1229–1262, 2014.

[110] Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941, 2019.

[111] Judea Pearl. Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

[112] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[113] Judea Pearl. Direct and indirect effects. In *Conference on Uncertainty in Artificial Intelligence*, page 411–420. Morgan Kaufmann Publishers Inc., 2001.

[114] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[115] Judea Pearl and James M Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Conference on Uncertainty in Artificial Intelligence*, volume 95, pages 444–453. Citeseer, 1995.

[116] Torsten Persson and Guido Enrico Tabellini. *Political Economics: Explaining Economic Policy*. MIT press, 2002.

[117] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference*. The MIT Press, 2017.

[118] James Robins. 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(9-12):1393–1512, 1986.

[119] James M Robins. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, pages 69–117. Springer, 1997.

[120] James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.

[121] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[122] Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meets causality. *arXiv:1801.06229*, 2018.

[123] Andrew Donald Roy. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146, 1951.

[124] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.

[125] John Rust. Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica*, pages 999–1033, 1987.

[126] Vira Semenova and Victor Chernozhukov. Estimation and inference about conditional average treatment effect and other structural functions. *arXiv:1702.06240*, 2017.

[127] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085, 2017.

[128] Carl-Johann Simon-Gabriel, Alessandro Barp, and Lester Mackey. Metrizing weak convergence with maximum mean discrepancies. *arXiv:2006.09268*, 2020.

[129] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pages 4595–4607, 2019.

[130] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

[131] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31, 2007.

[132] Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

[133] Bharath Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

[134] Bharath Sriperumbudur, Kenji Fukumizu, and Gert Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In *International Conference on Artificial Intelligence and Statistics*, pages 773–780, 2010.

[135] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

[136] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

[137] Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

[138] James H Stock. Nonparametric policy analysis. *Journal of the American Statistical Association*, 84(406):567–575, 1989.

[139] Liangjun Su, Takuya Ura, and Yichong Zhang. Non-separable models with high-dimensional data. *Journal of Econometrics*, 212(2):646–677, 2019.

[140] Eric J Tchetgen Tchetgen and Ilya Shpitser. Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, 40(3):1816, 2012.

[141] Ilya Tolstikhin, Bharath K Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

[142] Sara A Van de Geer. *Applications of Empirical Process Theory*, volume 91. Cambridge University Press, 2000.

[143] Tyler VanderWeele. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015.

[144] Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, and Francesca Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(May):883–904, 2005.

[145] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[146] Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[147] Max Welling. Herding dynamical weights to learn. In *International Conference on Machine Learning*, pages 1121–1128, 2009.

[148] Houssam Zenati, Alberto Bietti, Matthieu Martin, Eustache Diemert, and Julien Mairal. Counterfactual learning of continuous stochastic policies. *arXiv:2004.11722*, 2020.

[149] Tong Zhang. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems*, pages 471–478, 2003.

[150] Michael Zimmert and Michael Lechner. Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv:1908.08779*, 2019.

Contents

A Related work

Table 1: Causal parameters in the present work

Setting	Causal parameter	Expression
Treatment	average treatment effect (ATE)	$\theta_0^{ATE}(d) := \mathbb{E}[Y^{(d)}]$
	ATE with dist. shift	$\theta_0^{DS}(d, \tilde{\mathbb{P}}) := \mathbb{E}_{\tilde{\mathbb{P}}}[Y^{(d)}]$
	average treatment on the treated (ATT)	$\theta_0^{ATT}(d, d') := \mathbb{E}[Y^{(d')}\|D = d]$
	conditional ATE	$\theta_0^{CATE}(d, v) := \mathbb{E}[Y^{(d)}\|V = v]$
Mediation	total effect	$\theta_0^{TE}(d, d') := \mathbb{E}[Y^{(d', M^{(d')})} - Y^{(d, M^{(d)})}]$
	direct effect	$\theta_0^{DE}(d, d') := \mathbb{E}[Y^{(d', M^{(d)})} - Y^{(d, M^{(d)})}]$
	indirect effect	$\theta_0^{IE}(d, d') := \mathbb{E}[Y^{(d', M^{(d')})} - Y^{(d', M^{(d)})}]$
Planning	sequential ATE	$\theta_0^{SATE}(d_{1:T}) := \mathbb{E}[Y^{(d_{1:T})}]$
	sequential ATE with dist. shift	$\theta_0^{SDS}(d_{1:T}, \tilde{\mathbb{P}}) := \mathbb{E}_{\tilde{\mathbb{P}}}[Y^{(d_{1:T})}]$

Table 2: Existing semi-parametric and non-parametric methods

Setting	Causal parameter	Treatment type	Existing approach
Treatment	mean	discrete	DML, random forest
		continuous	Nadaraya-Watson, series
		text	*
Treatment	distribution	discrete	GMM, RKHS
		continuous	*
		text	*
Mediation	mean	discrete	DML
		continuous	*
		text	*
Mediation	distribution	discrete	*
		continuous	*
		text	*
Planning	mean	discrete	Monte Carlo
		continuous	*
		text	*
Planning	distribution	discrete	Monte Carlo
		continuous	*
		text	*

A.1 Identification

A.1.1 Exogenous treatment

In causal inference, identification refers to the study of which causal parameters can be empirically measured. In Table 1, we list various causal parameters. In the present work, we appeal to existing identification theorems under variations of the assumption *selection on observables*. The identification theorems give rise to empirical quantities for which we present novel algorithms. This framework is known as the Neyman-Rubin causal model [106, 124]. For treatment effects, we appeal to the classic identification result of [121]. For mediation analysis, we appeal to the classic identification result of [82]. For off-policy planning, we appeal to the classic identification result of [118]. In Appendix D, we present alternative identification results in terms of causal directed acyclic graphs (DAGs). For a review of the potential outcomes framework, we recommend [74, 9, 86, 117, 1]. For a review of the causal DAG framework, we recommend [114].

A.1.2 Endogenous treatment

In the present work, we restrict attention to settings in which treatment assignment is conditionally exogenous. An important literature in econometrics instead considers settings in which treatment assignment is endogenous: treatment D reflects the optimal choice of a rational agent with expectations about their potential outcomes $\{Y^{(d)}\}$. When treatment is discrete, such causal settings are known as *selection*, *choice*, or *Roy* models [123, 65, 66, 100, 5]. [72, 73, 4] provide an overview of such settings and the corresponding target parameters.

[71] presents a unified perspective on econometric policy evaluation on binary treatment D that may be endogenous. In particular, [71, Table 1A and Table 1B] demonstrate how a broad variety of causal quantities can be viewed as re-weightings of an underlying function called the *marginal treatment effect*:

$$\Delta(x, v) := \mathbb{E}[Y^{(1)} - Y^{(0)} | X = x, V = v]$$

where $V \sim \mathcal{U}[0, 1]$ is unobserved heterogeneity in the model for treatment D. There are several key differences between [71] and the present work. [71] focus on binary treatment D, while we allow for treatment that is discrete, continuous, or text. Whereas $\Delta(x, v)$ is a philosophical quantity in terms of potential outcomes and unobserved heterogeneity, $\gamma_0(d, x)$ is an empirical quantity; it is simply a prediction. [71] present a unified framework, while we present both a framework and estimators. [71] allows for generic models of endogenous treatment assignment via the existence of V, whereas we focus models of conditionally exogenous treatment assignment: *selection on observables* and its variants. We leave to future work the extension of our methods to these additional settings.

The endogenous treatment literature has developed approaches for continuous treatment that bear formal resemblance to our approaches. As summarized by [73], the general model is

$$Y = \alpha(D, X, U), \quad D = \beta(Z, V), \quad (X, Z) \perp\!\!\!\perp (U, V)$$

where (U, V) admit interpretation as non-separable noise. [15] consider as the target parameter the *average structural function*, defined as $\theta(d, x) := \mathbb{E}[Y^{(d)} | X = x] = \int \alpha(d, x, u) \mathbb{P}(u)$. Note that $\theta(d, x)$ can be viewed as a generalization of the *average response probability* in [23]. [46] refer to $\frac{\partial}{\partial d}\theta(d, x)$ as the average treatment effect. Interestingly, [85] show that if β is monotonic in V and if some additional conditional independences hold, then $\theta(d, x) := \mathbb{E}[Y^{(d)} | X = x] = \int \mathbb{E}[Y | D = d, X = x, V = v] \mathbb{P}(v)$. V is called the *control variable*, and though it is unobserved, [85] show that it is identified up to normalization as $V = F(Y | X, Z)$. There are several key differences between [85] and the current work. Whereas the target parameters in [85] are $\mathbb{E}[Y^{(d)} | X = x]$ and $\mathbb{P}[Y^{(d)} | X = x]$, the target parameters in the present work are the superset summarized in Table 1. Whereas the prediction $\mathbb{E}[Y | D = d, X = x, V = v]$ involves $V = F(Y | X, Z)$, the prediction $\gamma_0 := \mathbb{E}[Y | D = d, X = x]$ involves only variables that are directly observed. [85] present series estimators, while we present RKHS estimators. [85] allows for generic models of endogenous treatment assignment via the existence of V, whereas we focus models of conditionally exogenous treatment assignment: *selection on observables* and its variants. It is straightforward to extend our method to the control variable setting. Like [85], we too will require a full support assumption.

A.2 Kernel methods

A.2.1 General

The term *kernel methods* can refer to two classes of algorithms: Nadaraya-Watson estimators predicated on local smoothing; and penalized loss-minimizing estimators predicated on the theory of reproducing kernel Hilbert spaces (RKHSs). Nadaraya-Watson and series estimators are popular approaches in econometrics. In the present work we introduce RKHS estimators, which are popular in machine learning. For a general introduction to RKHS methods, we recommend the textbooks [146, 136, 13]. For uniform consistency of kernel ridge regression, we draw inspiration from [130].

The machine learning literature on mean embeddings is vast; see [103] for a comprehensive review. [131, 132] propose the embeddings of marginal and conditional distributions, respectively. [51] propose the use of conditional mean embeddings for non-parametric Bayesian inference, and in doing so propose algorithmic techniques for closed-form estimators called kernel sum rule, kernel chain rule, and kernel Bayes rule. Though our goal is non-parametric causal inference rather than non-parametric Bayesian inference, our algorithmic techniques are similar. [108] presents a method

for value iteration in kernelized partially observed Markov decision processes (POMDPs). As a non-parametric approach to control, it bears resemblance to our approach to off-policy planning. We present a different estimator, provide causal interpretation, and prove consistency.

Importantly for the present work, [31, 129] draw the formal connection between conditional expectation operators and conditional mean embeddings. Using mean embeddings, we represent the distance between two probability distributions by the RKHS norm of the difference of their mean embeddings, a metric called maximum mean discrepancy (MMD) that is widely used in the training of generative adversarial networks (GANs) [57]. [133, 128] prove that convergence in distribution is equivalent to convergence of mean embeddings in RKHS norm (i.e. MMD metric) under weak conditions, which we state in Appendix E. In doing so, [133, 128] provide an important formal result that elucidates *why* kernel mean embedding is the natural algorithmic and analytical approach for complex distributional causal adjustment. In previous work, [141] prove minimax rates for unconditional mean embeddings in RKHS norm and [129, Corollary 1] prove rates for conditional mean embeddings in RKHS norm.

A.2.2 Causal

Previous work has also incorporated RKHS constructions into non-parametric causal inference. In [19, 38, 129, 41], the authors suggest a role for the RKHS in ill-posed inverse problems, e.g. non-parametric instrumental variable regression, albeit with various constructions. In [107, 49], the authors analyze mean square error for a loss-minimizing RKHS estimator of heterogeneous treatment effects in the case of binary treatment and conditioning on the entire covariate. Drawing inspiration from these works, we present a different estimator that allows for more general treatment and for conditioning on a subset of covariates. In [148], the authors analyze mean square error for a generalized linear model RKHS estimator of continuous treatment effect. We present a different estimator and analyze sup-norm consistency.

Closest to the present work are papers that incorporate the technique of kernel mean embeddings into causal inference. Two existing papers that pursue this approach are [129, 104]. In [129], the authors consider consider the non-parametric instrumental variable regression problem and analyze mean square error. We frequently appeal to fundamental technical lemmas in [129] on the uniform consistency of conditional mean embeddings. In [104], the authors present a semi-parametric approach for estimating counterfactual outcome distributions with binary treatment. We generalize this learning problem in Appendix E, which allows for treatments that may be discrete, continuous, and even text [96, 53, 50]. Whereas [104] is semi-parametric and focuses on two learning problems, the present work is non-parametric and presents a unified perspective of many learning problems. We build on the algorithmic insight of [104] to estimate distributions over outcomes by introducing an RKHS for outcomes.

A.3 Treatment effect

A.3.1 Average treatment effect

Average treatment effect (ATE) is the quantity $\mathbb{E}[Y^{(1)} - Y^{(0)}]$, i.e. the mean potential outcome with treatment $D = 1$ minus the mean potential outcome with treatment $D = 0$. This quantity is defined with respect to a binary treatment D in semi-parametric causal inference, as reviewed in [84]. [63, 67, 2] analyze semi-parametric estimators based on propensity score matching, while [76] analyze a semi-parametric estimator based on inverse propensity weighting. The *de-biased machine learning* literature presents semi-parametric estimators of ATE using black-box machine learning of the prediction γ_0 and an explicit bias correction [25]. Recently, [10, 29] present semi-parametric approaches to estimating ATE using black-box machine learning and implicit bias correction.

A natural generalization of ATE allows for treatment D that is continuous. The target parameter is $\mathbb{E}[Y^{(d)}]$, interpretable as the mean potential outcome with treatment $D = d$. The learning problem is now non-parametric. [83, 75] present a non-parametric estimator of continuous treatment effect based on generalized propensity score matching, while [52] presents a non-parametric estimator of continuous treatment effect based on generalized inverse propensity weighting. Another class of estimators uses Nadaraya-Watson smoothing around DML at a point [92, 90, 139, 32]. These works prove pointwise consistency and pointwise asymptotic normality. Recent work also uses series regression around a de-biased pseudo-outcome, with not only pointwise but also uniform guarantees [126].

A different generalization of ATE considers the quantity $\mathbb{P}(Y^{(1)}) - \mathbb{P}(Y^{(0)})$, i.e. the distribution of potential outcome with treatment $D = 1$ minus the distribution of potential outcome with treatment $D = 0$. This quantity is especially important in assessing the impact of welfare reform on earnings, transfers, and income in labor economics [14]. Note that this quantity differs from $\mathbb{P}(Y^{(1)} - Y^{(0)})$, which is studied in the evaluation of policy regime $D = 1$ versus policy regime $D = 0$ in political economy [116, 4]. [44, 22, 104] present methods for estimating $\mathbb{P}(Y^{(1)}) - \mathbb{P}(Y^{(0)})$, using inverse propensity weighting of check functions, moment functions, and kernel mean embeddings, respectively. In the present work, we provide estimators for these variations of ATE, allowing for treatment D that is discrete, continuous, or text, and allowing for either means or distributions of potential outcomes as the target. We estimate distributions of potential outcomes by modeling the mean embeddings then sampling from the mean embeddings by kernel herding [24].

A.3.2 Distribution shift

[33, 102, 72, 7] examine *external validity* in economic research. For example, the ATE of job training in Baltimore may not be the same as the ATE of job training in San Diego, so careful empirical research using a natural experiment in Baltimore may not be an accurate guide for policymakers in San Diego [77]. In the present work, we consider which assumptions and measurements are necessary to calculate ATE in a city where the natural experiment did not take place. [77] provide an empirical exploration of such issues. The learning problem may be viewed as a generalization of the the so-called *policy relevant treatment effect* defined in [70, 72]. It may also be viewed as a potential outcome refinement of the *policy effect* introduced by [138], for which [28, 29] provide DML estimators with implicit bias correction.

Questions of this nature are widely studied in machine learning under the names of distribution shift, covariate shift, and transfer learning. [58] present an RKHS approach for prediction under distribution shift. [91] present a DML approach for counterfactual prediction under distribution shift that uses Nadaraya-Watson smoothing and density estimation. We present a novel RKHS approach for counterfactual means and distributions of potential outcomes, allowing for treatment that may be discrete, continuous, or text. Note that our potential outcome formulation of the distribution shift learning problem differs from [122, 87, 30].

A.3.3 Average treatment on the treated

Average treatment on the treated (ATT) is the quantity $\mathbb{E}[Y^{(1)} - Y^{(0)}|D = 1]$, i.e. the mean potential outcome with treatment $D = 1$ minus the mean potential outcome with treatment $D = 0$ for the sub-population who received treatment. This quantity is defined with respect to a binary treatment D in semi-parametric causal inference. Recently, [29] propose a semi-parametric approach to estimating ATT using black-box machine learning and implicit bias correction. We consider the more general learning problem in which treatment may be discrete, continuous, or text.

A generalization of ATT considers the quantity $\mathbb{P}(Y^{(1)}|D = 1) - \mathbb{P}(Y^{(0)}|D = 1)$, i.e. the distribution of potential outcomes with treatment $D = 1$ minus the distribution of potential outcomes with treatment $D = 0$ for the sub-population who received treatment. [27] refer to the latter term $\mathbb{P}(Y^{(0)}|D = 1)$ as the *counterfactual distribution* and present estimators that maximize a likelihood based on the empirical distributions for each sub-population. [104] study this learning problem as well, using mean embeddings of distributions for each sub-population. We consider the more general learning problem in which treatment may be discrete, continuous, or text. Our estimation procedure shares information across sub-populations by using conditional mean embeddings.

A.3.4 Conditional average treatment effect

Treatment effects may be heterogeneous: for different sub-populations, the treatment may differentially affect mean potential outcomes. Conditional average treatment effect (CATE) captures this phenomenon. The semi-parametric definition of CATE is $\mathbb{E}[Y^{(1)} - Y^{(0)}|V = v]$, i.e. the mean potential outcome with treatment $D = 1$ minus the mean potential outcome with treatment $D = 0$ for the sub-population with value $V = v$ where $V \subset X$ is an interpretable low-dimensional subset of the covariates such as age, race, or gender [6]. This quantity is defined with respect to a binary treatment D in existing work. [126] presents a series estimator around a de-biased pseudo-outcome and [28, 42, 150] present DML estimators that make use of Nadaraya-Watson smoothing. We consider

the more general learning problem in which treatment may be discrete, continuous, or text, and in which the analyst may be interested in means or distributions of potential outcomes.

A vast literature on heterogeneous treatment effects focuses on the case in which $V = X$, i.e. where CATE is defined with respect to the entire covariate vector: $\mathbb{E}[Y^{(1)} - Y^{(0)}|X = x]$. For this setting, [107] propose an RKHS approach and analyze mean square error, and [127] propose a regularized neural network approach and analyze mean square error. [145, 110] propose a random forest approaches and analyze pointwise consistency and pointwise asymptotic normality. The latter allows for continuous treatment in a partially linear setting. [94] present a meta-algorithm to use black-box machine learning to estimate $\mathbb{E}[Y^{(1)} - Y^{(0)}|X = x]$. Our estimators apply to this setting as well.

A.4 Mediation analysis

A recent literature in applied economic research estimates not only the treatment effect of an intervention but also the the economic *mechanism* at play. It is well documented that high quality early childhood interventions D substantially impact later life outcomes Y [93, 36]. In [64], the authors examine the extent to which the effects of these interventions are mediated by the mechanism of psychological skills M. Specifically, the authors examine the case study of the Perry Preschool program, a flagship early childhood intervention in the US that targeted disadvantaged children aged three to four, where data were collected for treatment and control groups through age 40. The authors provide important insights into the economics of personality [8]. While the empirical approach used by [64] is parametric, in the present work we propose an approach that is non-parametric.

Mediation analysis was initially studied in psychology, and has since gained popularity in epidemiology and public policy [88, 12]. For a comprehensive review, we recommend [143]. Existing methods often assume the treatment D and mediator M are binary, and hence are semi-parametric. [81] provides a Monte Carlo algorithm, while [140, 43] provide DML algorithms. [80] consider the non-parametric setting with continuous treatment and mediator, and analyze a generalized inverse propensity weighting estimator that relies on Nadaraya-Watson kernel density estimation. Interestingly, [69] bridge mediation analysis with the econometrics of estimating production technologies, e.g. as in [35]. We leave to future work the extension of our methods to the problem of estimating production technologies.

A.5 Off-policy planning

A rich literature that spans econometrics and epidemiology considers treatment effects in dynamic settings, where sequences of actions $D_{1:T}$ may be fixed or may depend on observed states $X_{1:T}$. [4] provide a thorough review. In epidemiology, this learning problem is studied under the name of *g-estimation*. The main assumption is a sequential generalization of selection on observables, which is satisfied in multi-stage clinical trials. [118] studies this setting with binary actions and states, [119] brings further clarity, [54] generalize analysis to continuous states and actions, and [97] generalizes analysis to continuous time. [95, 45] import this methodology to economic applications. [105] considers the related problem of estimating optimal treatment regimes. [37] provide a Monte Carlo algorithm for the setting with binary actions D_t. Building on previous work, we allow actions to be discrete, continuous, or text; we consider not only means but also distributions of potential outcomes; and we prove uniform consistency.

The starting point for dynamic treatment effect estimation in econometrics is [125]. In [125], the key identifying assumption is closely related to sequential selection on observables: conditional on observed state variables, unobserved state variables are temporally independent and independent of outcomes. [4] carefully compares this assumption with sequential selection on observables as well as the assumptions in [55, 47]. In subsequent work, [78, 79, 99, 98] invoke the same condition. In practice, the estimation strategy is typically parametric maximum likelihood. [109] presents a semi-parametric approach for binary actions D_t. Recently, [39] present a semi-parametric DML approach allowing for continuous actions D_t, where the value function takes a parametric form. [89] present an approach with Nadaraya-Watson smoothing around DML. To allow dynamic information accumulation, [5, 3] develop a continuous time event-history approach, while [68] develop a factor model approach. We leave to future work the extension of our methods to these important economic settings.

B Mediation analysis

B.1 Learning problem

Mediation analysis seeks to decompose the total effect of treatment D on outcome Y into the direct effect versus the indirect effect mediated via mechanism M. We denote the counterfactual mediator $M^{(d)}$ given a hypothetical intervention on treatment $D = d$. We denote the counterfactual outcome $Y^{(d,m)}$ given a hypothetical intervention on not only treatment $D = d$ but also mediator $M = m$. Our definitions of direct and indirect effect follow [120, 113].

Definition B.1 (Mediation analysis)**.** *We define the following mediated effects*

1. $\theta_0^{TE}(d, d') := \mathbb{E}[Y^{(d', M^{(d')})} - Y^{(d, M^{(d)})}]$ *is the total effect of new treatment value* d' *compared to old value* d
2. $\theta_0^{DE}(d, d') := \mathbb{E}[Y^{(d', M^{(d)})} - Y^{(d, M^{(d)})}]$ *is the direct effect of new treatment value* d' *compared to old value* d
3. $\theta_0^{IE}(d, d') := \mathbb{E}[Y^{(d', M^{(d')})} - Y^{(d', M^{(d)})}]$ *is the indirect effect of new treatment value* d' *compared to old value* d*, i.e. the component of total effect mediated by* M
4. $\theta_0^{ME}(d, d') := \mathbb{E}[Y^{(d', M^{(d)})}]$ *is the counterfactual mean outcome in the thought experiment that treatment is set at a new value* $D = d'$ *but the mediator* M *follows the distribution it would have followed if treatment were set at its old value* $D = d$

$\theta_0^{TE}(d, d')$ generalizes the notion of an average total effect. Average total effect of a binary treatment $D \in \{0, 1\}$ is $\mathbb{E}[Y^{(1, M^{(1)})} - Y^{(0, M^{(0)})}]$. The analyst is essentially estimating a 2-vector of counterfactual mean outcomes $(\mathbb{E}[Y^{(0, M^{(0)})}], \mathbb{E}[Y^{(1, M^{(1)})}])$, where the length of the vector is the cardinality of the support of treatment D. For treatment that is more generally discrete, continuous, or even text, the vector $\mathbb{E}[Y^{(d, M^{(d)})}]$ may be infinite-dimensional, which makes this problem fully non-parametric rather than semi-parametric.

$\theta_0^{TE}(d, d')$ captures the concept of total effect, but the total effect may be mostly mediated by some mechanism M. In that scenario, a policy-maker may prefer to intervene on M rather than D to achieve a social outcome. An analyst may therefore wish to measure how much of the total effect is direct: if the mediator were held at the original distribution corresponding to $D = d$, what would be the impact of treatment $D = d'$? In semi-parametrics, the direct effect is $\mathbb{E}[Y^{(1, M^{(0)})}] - \mathbb{E}[Y^{(0, M^{(0)})}]$.

The remaining component of total effect is the indirect effect, which answers: how much of the total effect would be achieved by simply intervening on the mediator? In semi-parametrics, the indirect effect is $\mathbb{E}[Y^{(1, M^{(1)})}] - \mathbb{E}[Y^{(1, M^{(0)})}]$. The comparison is between counterfactual mean outcomes where in the former term the mediator follows the counterfactual distribution under intervention $D = 1$ and in the latter it follows the counterfactual distribution under intervention $D = 0$.

The final target parameter is the highly conceptual $\theta_0^{ME}(d, d')$. It is only significant insofar as the interpretable target parameters used in practice–$\theta_0^{TE}(d, d')$, $\theta_0^{DE}(d, d')$, and $\theta_0^{IE}(d, d')$–can be expressed in terms of $\theta_0^{ME}(d, d')$. Conveniently, it is sufficient to estimate $\theta_0^{ME}(d, d')$. In semi-parametrics, this quantity would be a matrix in $\mathbb{R}^{2\times 2}$. In our non-parametric approach, it is a surface over $\mathcal{D} \times \mathcal{D}$.

Proposition B.1. *All mediated effects can be expressed in terms of* $\theta_0^{ME}(d, d')$*, since*

1. $\theta_0^{TE}(d, d') = \theta_0^{DE}(d, d') + \theta_0^{IE}(d, d') = \theta_0^{ME}(d', d') - \theta_0^{ME}(d, d)$
2. $\theta_0^{DE}(d, d') = \theta_0^{ME}(d, d') - \theta_0^{ME}(d, d)$
3. $\theta_0^{IE}(d, d') = \theta_0^{ME}(d', d') - \theta_0^{ME}(d, d')$

For comparison, note that under the assumption of no interference

4. $\theta_0^{TE}(d, d') = \theta_0^{ATE}(d') - \theta_0^{ATE}(d)$

B.2 Identification

In the seminal work [82], the authors state sufficient conditions under which mediated effects–philosophical quantities stated in terms of potential outcomes $\{Y^{(d,m)}\}$–can be measured from empirical quantities such as outcomes Y, treatments D, mediators M, and covariates X. We will refer to this collection of sufficient conditions as *selection on observables for mediation*.

Assumption B.1 (Selection on observables for mediation)**.** *Assume*

1. *No interference: if* $D = d$ *then* $M = M^{(d)}$*; if* $D = d$ *and* $M = m$ *then* $Y = Y^{(d,m)}$
2. *Conditional exchangeability:* $\{M^{(d)}, Y^{(d',m)}\} \perp\!\!\!\perp D|X$ *and* $Y^{(d',m)} \perp\!\!\!\perp M|D, X$
3. *Overlap: if* $f(d, x) > 0$ *then* $f(m|d, x) > 0$*; and if* $f(x) > 0$ *then* $f(d|x) > 0$

where $f(d, x) > 0$*,* $f(m|d, x) > 0$*,* $f(x)$*, and* $f(d|x)$ *are densities.*

No interference is also called the stable unit treatment value assumption (SUTVA). It rules out network effects, also called spillovers. Conditional exchangeability states that conditional on covariates X, treatment assignment is as good as random. Moreover, conditional on treatment D and covariates X, mediation assignment is as good as random. Overlap ensures that there is no covariate stratum $X = x$ such that treatment has a restricted support, and that there is no treatment-covariate stratum $(D, X) = (d, x)$ such that the mediator has a restricted support. [69] compares these assumptions with those required to estimate production technologies, e.g. as in [35].

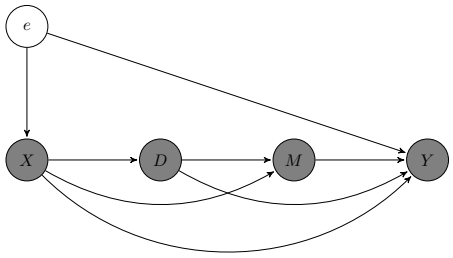


Figure 4: Selection on observables for mediation DAG

Formally, the theorem that uses these assumptions to express mediation effects in terms of data is known as an identification result. We quote a classic identification result below. Define the prediction

$$\gamma_0(d, m, x) := \mathbb{E}[Y|D = d, M = m, X = x]$$

Theorem B.1 (Identification of mediated effects [82])**.** *If Assumption B.1 holds then*

$$\theta_0^{ME}(d, d') = \int \gamma_0(d', m, x)\mathbb{P}(m|d, x)\mathbb{P}(x)$$

It is immediate by Proposition B.1 that all other quantities in Definition B.1 are also identified.

B.3 Algorithm

Theorem B.1 makes precise how each mediation effect is identified as a quantity of the form $\int \gamma_0(d', m, x)\mathbb{Q}$ for the distribution $\mathbb{Q} = \mathbb{P}(m|d, x)\mathbb{P}(x)$. We now assume that γ_0 is an element of a function space called a reproducing kernel Hilbert space, as in Section 2.

In our construction, we define scalar-valued RKHSs for treatment D, mediator M, and covariates X, then assume that the prediction is an element of a tensor product space. Let $k_{\mathcal{D}} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, $k_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$, and $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be measureable positive definite kernels corresponding to scalar-valued RKHSs $\mathcal{H}_{\mathcal{D}}$, $\mathcal{H}_{\mathcal{M}}$, and $\mathcal{H}_{\mathcal{X}}$. Denote the feature maps

$$\phi_{\mathcal{D}} : \mathcal{D} \to \mathcal{H}_{\mathcal{D}},\ d \mapsto k_{\mathcal{D}}(d, \cdot) \quad \phi_{\mathcal{M}} : \mathcal{M} \to \mathcal{H}_{\mathcal{M}},\ m \mapsto k_{\mathcal{M}}(m, \cdot) \quad \phi_{\mathcal{X}} : \mathcal{X} \to \mathcal{H}_{\mathcal{X}},\ x \mapsto k_{\mathcal{X}}(x, \cdot)$$

To lighten notation, we will suppress subscripts when arguments are provided, e.g. we will write $\phi(d) = \phi_{\mathcal{D}}(d)$.

For mediated effects, we assume the prediction γ_0 is an element of the RKHS with tensor product feature map $\phi(d) \otimes \phi(m) \otimes \phi(x)$, i.e. $\gamma_0 \in \mathcal{H} := \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{M}} \otimes \mathcal{H}_{\mathcal{X}}$. In this construction, we appeal to the fact that the product of positive definite kernels corresponding to $\mathcal{H}_{\mathcal{D}}$, $\mathcal{H}_{\mathcal{M}}$, and $\mathcal{H}_{\mathcal{X}}$ defines a new positive definite kernel corresponding to $\mathcal{H}$. We choose the product construction because it is provides a rich composite basis. Therefore by the reproducing property

$$\gamma_0(d, m, x) = \langle \gamma_0, \phi(d) \otimes \phi(m) \otimes \phi(x) \rangle_{\mathcal{H}}$$

With this RKHS construction, we obtain a representation of the target parameter as an inner product in the space $\mathcal{H}$.

Theorem B.2 (Representation of mediated effects)**.** *If $\gamma_0 \in \mathcal{H}$ then*

$$\theta_0^{ME}(d, d') = \langle \gamma_0, \phi(d') \otimes \int [\mu_m(d, x) \otimes \phi(x)] \mathbb{P}(x) \rangle_{\mathcal{H}}$$

where $\mu_m(d, x) := \int \phi(m) \mathbb{P}(m|d, x)$.

The quantity $\mu_m(d, x) := \int \phi(m) \mathbb{P}(m|d, x)$ is the conditional mean embedding of $\mathbb{P}(m|d, x)$. The quantity $\int [\mu_m(d, x) \otimes \phi(x)] \mathbb{P}(x)$ further encodes $\mathbb{P}(x)$. As far as we know, this sequential embedding is an innovation in the kernel methods literature. It is immediate by Proposition B.1 that all other quantities in Definition B.1 are also represented by inner products.

While this representation appears abstract, it is eminently useful for defining an estimator with a closed form solution that can be computed in one line of code (after computing kernel matrices). Our estimator will be $\hat{\theta}^{ME}(d, d') = \langle \hat{\gamma}, \phi(d') \otimes \frac{1}{n} \sum_{i=1}^{n} [\hat{\mu}_m(d, x_i) \otimes \phi(x_i)] \rangle_{\mathcal{H}}$. The estimator $\hat{\gamma}$ is a standard kernel ridge regression, with known closed form. The estimator $\hat{\mu}_m(d, x)$ is an appropriately defined kernel ridge regression.

Algorithm B.1 (Estimation of mediated effects)**.** *Denote the empirical kernel matrices*

$$K_{DD} \in \mathbb{R}^{n \times n}, \quad K_{MM} \in \mathbb{R}^{n \times n} \quad K_{XX} \in \mathbb{R}^{n \times n}$$

K_{DD}, K_{MM}, and K_{XX} are calculated from observations drawn from population $\mathbb{P}$. Denote by $\odot$ the element-wise product. Mediated effect estimators have closed form solutions based on

$$\hat{\theta}^{ME}(d, d') = \frac{1}{n} \sum_{i=1}^{n} Y^T (K_{DD} \odot K_{MM} \odot K_{XX} + n\lambda)^{-1} \\ (K_{Dd'} \odot \{K_{MM} (K_{DD} \odot K_{XX} + n\lambda_3)^{-1} (K_{Dd} \odot K_{Xx_i})\} \odot K_{Xx_i})$$

where (λ, λ_3) are ridge regression penalty parameters.[1]

It is immediate by Proposition B.1 that all other quantities in Definition B.1 are estimated as differences of $\hat{\theta}^{ME}(d, d')$. We give theoretical values for (λ, λ_3) in Appendix I that balance bias and variance. We give a practical tuning procedure in Appendix J based on leave-one-out cross validation.

B.4 Consistency

Towards a guarantee of uniform consistency, we place regularity conditions on the original spaces and scalar-valued RKHSs.

Assumption B.2 (Original space regularity conditions)**.** *Assume*

1. *$\mathcal{D}$, $\mathcal{M}$, and $\mathcal{X}$ are Polish spaces, i.e. separable and completely metrizable topological spaces*
2. *Y is bounded, i.e. $\exists C < \infty$ such that $|Y| \leq C$ almost surely*

A Polish space may be low-, high-, or infinite-dimensional. Random variables with support in a Polish space may be discrete, continuous, or even text. For simplicity of argument, we require that outcome $Y \in \mathbb{R}$ is bounded.

Assumption B.3 (RKHS regularity conditions)**.** *Assume*

1. *$k_{\mathcal{D}}$, $k_{\mathcal{M}}$, and $k_{\mathcal{X}}$ are continuous and bounded:*

$$\sup_{d \in \mathcal{D}} \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \leq \kappa_d, \quad \sup_{m \in \mathcal{M}} \|\phi(m)\|_{\mathcal{H}_{\mathcal{M}}} \leq \kappa_m, \quad \sup_{x \in \mathcal{X}} \|\phi(x)\|_{\mathcal{H}_{\mathcal{X}}} \leq \kappa_x$$

[1] Equivalently, $\hat{\theta}^{ME}(d, d') = Y^T (K_{DD} \odot K_{MM} \odot K_{XX} + n\lambda)^{-1} (K_{Dd'} \odot \{(K_{MM} (K_{DD} \odot K_{XX} + n\lambda_3)^{-1}) \odot \frac{1}{n} K_{XX}^2\} K_{Dd})$. This re-write can be shown by calculating elements of each matrix. Although the two expressions for $\hat{\theta}^{ME}$ are equivalent, the latter significantly speeds up computation.

2. *$\phi(d)$, $\phi(m)$, and $\phi(x)$ are measurable*

3. *$k_{\mathcal{M}}$ and $k_{\mathcal{X}}$ are characteristic*

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

Next, we assume the prediction γ_0 is smooth.

Assumption B.4 (Smoothness of prediction)**.** *Assume*

1. *the prediction is well-specified, i.e. $\gamma_0 \in \mathcal{H}$*

2. *the prediction is a particularly smooth element of $\mathcal{H}$. Formally, define the covariance operator T for $\mathcal{H}$. We assume $\exists g \in \mathcal{H}$ s.t. $\gamma_0 = T^{\frac{c-1}{2}} g$, $c \in (1, 2]$, and $\|g\|_{\mathcal{H}}^2 \leq \zeta$*

For θ_0^{ME}, $T := \mathbb{E}[\{\phi(D) \otimes \phi(M) \otimes \phi(X)\} \otimes \{\phi(D) \otimes \phi(M) \otimes \phi(X)\}]$.

For the conditional mean embedding, we place further smoothness conditions on the corresponding conditional expectation operator.

Assumption B.5 (Smoothness for θ_0^{ME})**.** *Assume*

1. *the conditional expectation operator E_3 is well-specified as a Hilbert-Schmidt operator between RKHSs, i.e. $E_3 \in \mathcal{L}_2(\mathcal{H}_{\mathcal{M}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})$, where*

$$E_3 : \mathcal{H}_{\mathcal{M}} \to \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}}, \quad f(\cdot) \mapsto \mathbb{E}[f(M)|D = \cdot, X = \cdot]$$

2. *the conditional expectation operator is a particularly smooth element of $\mathcal{L}_2(\mathcal{H}_{\mathcal{M}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})$. Formally, define the covariance operator $T_3 := \mathbb{E}[\{\phi(D) \otimes \phi(X)\} \otimes \{\phi(D) \otimes \phi(X)\}]$ for $\mathcal{L}_2(\mathcal{H}_{\mathcal{M}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})$. We assume $\exists G_3 \in \mathcal{L}_2(\mathcal{H}_{\mathcal{M}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})$ s.t. $E_3 = (T_3)^{\frac{c_3-1}{2}} \circ G_3$, $c_3 \in (1, 2]$, and $\|G_3\|^2_{\mathcal{L}_2(\mathcal{H}_{\mathcal{M}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})} \leq \zeta_3$*

With these assumptions, we arrive at our second main result.

Theorem B.3 (Consistency)**.** *Suppose Assumptions B.1, B.2, B.3, B.4, and B.5 hold. Then*

$$\|\hat{\theta}^{ME} - \theta_0^{ME}\|_\infty = O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + n^{-\frac{1}{2}\frac{c_3-1}{c_3+1}}\right)$$

It is immediate by Proposition B.1 that all other quantities in Definition B.1 are uniformly consistent with the same rate. The exact finite sample rate is given in Appendix I. This rate is at best $n^{-\frac{1}{6}}$, by setting $(c, c_3) = 2$. The slow rate reflects the challenge of a sup-norm guarantee, which is much stronger than a prediction guarantee, and which encodes caution about worst-case scenarios when informing policy decisions. Our analysis is agnostic about the spectral decay; further assumptions on spectral decay (interpretable as effective dimension) will lead to a faster rate of prediction.

C Off-policy planning

C.1 Learning problem

So far we have considered the effect of a one-off treatment D. A rich literature instead considers the effect of a sequence of treatments $D_{1:T} = d_{1:T}$ on counterfactual outcome $Y^{(d_{1:T})}$. If the sequence of treatment values $d_{1:T}$ was actually observed in the data, this problem is called on-policy planning; if the sequence of treatment values was not actually observed, this problem is called off-policy planning. In reinforcement learning and structural estimation, the tuple $(Y, D_{1:T}, X_{1:T})$ may instead be denoted $(R, A_{1:T}, S_{1:T})$ for reward, action, and state.

Definition C.1 (Off-policy planning)**.** *We define the following off-policy effects*

1. *$\theta_0^{SATE}(d_{1:T}) := \mathbb{E}[Y^{(d_{1:T})}]$ is the counterfactual mean outcome given interventions $D_{1:T} = d_{1:T}$ for the entire population*

2. $\theta_0^{SDS}(d_{1:T}, \tilde{\mathbb{P}}) := \mathbb{E}_{\tilde{\mathbb{P}}}[Y^{(d_{1:T})}]$ *is the counterfactual mean outcome given interventions* $D_{1:T} = d_{1:T}$ *for an alternative population with data distribution* $\tilde{\mathbb{P}}$

$\theta_0^{SATE}(d_{1:T})$ is a sequential generalization of $\theta_0^{ATE}(d)$. Whereas the semi-parametric literature restricts $d_t \in \{0, 1\}$, we allow d_t to be discrete, continuous, or even text. We consider a fully non-parametric approach to off-policy planning. Likewise, $\theta_0^{SDS}(d_{1:T})$ is a sequential generalization of $\theta_0^{DS}(d_{1:T})$. In the spirit of off-policy planning, we consider the additional difficulty of a shift in the distribution from $\mathbb{P}$ to $\tilde{\mathbb{P}}$. As in Section 2, additional nuanced off-policy effects may be defined.

In the present work, we consider only the deterministic, fixed counterfactual policy $d_{1:T}$. It is deterministic in the sense that it is non-random. It is fixed in the sense that it does not depend on the observed sequence of covariates $X_{1:T}$. Impressively, the causal inference literature on off-policy planning extends to policies that may be randomized and that may be dynamic. Denote by $\Delta(\mathcal{D}_t)$ the space of distributions over $\mathcal{D}_t$. Let $\Delta(\mathcal{D}_t)$ be the space of distributions over $\mathcal{D}_t$. A randomized, dynamic policy may be recursively defined by the strategy $g_t : \mathcal{D}_{1:t-1} \times \mathcal{X}_{1:t} \to \Delta(\mathcal{D}_t)$ which in turn admits a representation as the mappings $g_t' : \mathcal{X}_{1:t} \to \Delta(\mathcal{D}_t)$. This richer learning problem is important for any setting with a strategic agent. We leave analysis of the richer learning problem for future work.

C.2 Identification

In the seminal work [118], the author states sufficient conditions under which the off-policy effect– a philosophical quantity stated in terms of potential outcomes $\{Y^{(d_{1:T})}\}$–can be measured from empirical quantities such as outcomes Y, treatments $D_{1:T}$, and covariates $X_{1:T}$. Colloquially, this collection of sufficient conditions is known as *sequential selection on observables*.

Assumption C.1 (Sequential selection on observables)**.** *Assume*

1. *No interference: if* $D_{1:T} = d_{1:T}$ *then* $Y = Y^{(d_{1:T})}$
2. *Conditional exchangeability:* $\{Y^{(d_{1:T})}\} \perp\!\!\!\perp D_t | D_{1:t-1}, X_{1:t}$ *for all* $t = 1 : T$
3. *Overlap: if* $f(d_{1:t-1}, x_{1:t}) > 0$ *then* $f(d_t | d_{1:t-1}, x_{1:t}) > 0$

where $f(d_{1:t-1}, x_{1:t})$ *and* $f(d_t | d_{1:t-1}, x_{1:t})$ *are densities.*

Observe that Assumption C.1 is a sequential generalization of Assumption 2.1. As before, no interference rules out network effects, also called spillovers. Conditional exchangeability states that conditional on the history of treatments $D_{1:t-1}$ and covariates $X_{1:t}$, treatment assignment D_t is as good as random. Overlap ensures that there is no treatment-covariate history $(d_{1:t-1}, x_{1:t})$ such that treatment has a restricted support. In [125], a cornerstone of dynamic discrete choice models, the key identifying assumption is closely related: conditional on observed state variables, unobserved state variables are temporally independent and independent of outcomes. [4] carefully relates these assumptions.

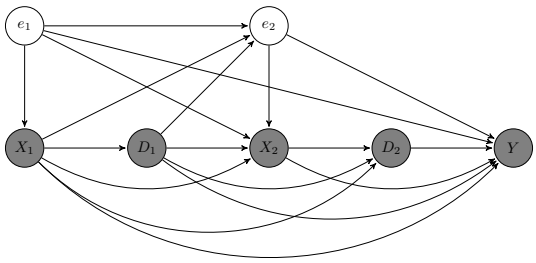


Figure 5: Sequential selection on observables DAG

To handle θ_0^{SDS}, we also make a standard assumption in transfer learning.

Assumption C.2 (Distribution shift)**.** *The difference in population distributions* $\mathbb{P}$ *and* $\tilde{\mathbb{P}}$ *is only in the distribution of treatments and covariates.*

$$\mathbb{P}(Y, D_{1:T}, X_{1:T}) = \mathbb{P}(Y | D_{1:T}, X_{1:T}) \mathbb{P}(D_{1:T}, X_{1:T})$$
$$\tilde{\mathbb{P}}(Y, D_{1:T}, X_{1:T}) = \mathbb{P}(Y | D_{1:T}, X_{1:T}) \tilde{\mathbb{P}}(D_{1:T}, X_{1:T})$$

Clearly, Assumption C.2 is a sequential generalization of Assumption 2.2. An immediate consequence is that the prediction function $\gamma_0(d_{1:T}, x_{1:T}) := \mathbb{E}[Y | D_{1:T} = d_{1:T}, X_{1:T} = x_{1:T}]$ remains the same

across the different populations $\mathbb{P}$ and $\tilde{\mathbb{P}}$. Formally, the theorem that uses these assumptions to express treatment effects in terms of data is known as an identification result. We quote a classic identification result below.

Theorem C.1 (Identification of off-policy effects [118])**.** *If Assumption C.1 holds then*

1. $\theta_0^{SATE}(d_{1:T}) = \int \gamma_0(d_{1:T}, x_{1:T})\mathbb{P}(x_1)\prod_{t=2}^{T}\mathbb{P}(x_t|d_{1:t-1}, x_{1:t-1})$
2. *If in addition Assumption C.2 holds then* $\theta_0^{SDS}(d_{1:T}, \tilde{\mathbb{P}}) = \int \gamma_0(d_{1:T}, x_{1:T})\tilde{\mathbb{P}}(x_1)\prod_{t=2}^{T}\tilde{\mathbb{P}}(x_t|d_{1:t-1}, x_{1:t-1})$

The integral is the famous *g-formula* from epidemiology stated for sequences of treatments that may be discrete, continuous, or even text. We consider a fully non-parametric g-formula that allows for distribution shift.

C.3 Algorithm

Theorem C.1 makes precise how each off-policy effect is identified as a quantity of the form $\int \gamma_0(d_{1:T}, x_{1:T})\mathbb{Q}$ for the distribution $\mathbb{Q} = \mathbb{P}$ or $\mathbb{Q} = \tilde{\mathbb{P}}$. We now assume that γ_0 is an element of a function space called a reproducing kernel Hilbert space, as in Sections 2 and B. For clarity, we present the algorithm with $T = 2$.

In our construction, we define scalar-valued RKHSs for each treatment D_t and each covariate X_t, then assume that the prediction is an element of a tensor product space. Let $k_{\mathcal{D}} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ and $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be measureable positive definite kernels corresponding to scalar-valued RKHSs $\mathcal{H}_{\mathcal{D}}$ and $\mathcal{H}_{\mathcal{X}}$. Denote the feature maps

$$\phi_{\mathcal{D}} : \mathcal{D} \to \mathcal{H}_{\mathcal{D}}, \quad d \mapsto k_{\mathcal{D}}(d, \cdot) \quad \phi_{\mathcal{X}} : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}, \quad x \mapsto k_{\mathcal{X}}(x, \cdot)$$

To lighten notation, we will suppress subscripts when arguments are provided, e.g. we will write $\phi(d) = \phi_{\mathcal{D}}(d)$.

For off-policy effects with $T = 2$, we assume the prediction γ_0 is an element of the RKHS with tensor product feature map $\phi(d_1) \otimes \phi(d_2) \otimes \phi(x_1) \otimes \phi(x_2)$, i.e. $\gamma_0 \in \mathcal{H} := \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}$. In this construction, we appeal to the fact that the product of positive definite kernels corresponding to $\mathcal{H}_{\mathcal{D}}$ and $\mathcal{H}_{\mathcal{X}}$ defines a new positive definite kernel corresponding to $\mathcal{H}$. We choose the product construction because it is provides a rich composite basis. Therefore by the reproducing property

$$\gamma_0(d_1, d_2, x_1, x_2) = \langle \gamma_0, \phi(d_1) \otimes \phi(d_2) \otimes \phi(x_1) \otimes \phi(x_2) \rangle_{\mathcal{H}}$$

With this RKHS construction, we obtain a representation of the target parameter as an inner product in the space $\mathcal{H}$.

Theorem C.2 (Representation of off-policy effects)**.** *If $\gamma_0 \in \mathcal{H}$ then*

1. $\theta_0^{SATE}(d_1, d_2) = \langle \gamma_0, \phi(d_1) \otimes \phi(d_2) \otimes \int \phi(x_1) \otimes \mu_{x_2}(d_1, x_1)\mathbb{P}(x_1) \rangle_{\mathcal{H}}$ *where* $\mu_{x_2}(d_1, x_1) := \int \phi(x_2)\mathbb{P}(x_2|d_1, x_1)$
2. $\theta_0^{SDS}(d_1, d_2, \tilde{\mathbb{P}}) = \langle \gamma_0, \phi(d_1) \otimes \phi(d_2) \otimes \int \phi(x_1) \otimes \nu_{x_2}(d_1, x_1)\tilde{\mathbb{P}}(x_1) \rangle_{\mathcal{H}}$ *where* $\nu_{x_2}(d_1, x_1) := \int \phi(x_2)\tilde{\mathbb{P}}(x_2|d_1, x_1)$

In θ_0^{SATE}, the quantity $\mu_{x_2}(d_1, x_1) := \int \phi(x_2)\mathbb{P}(x_2|d_1, x_1)$ is the conditional mean embedding of $\mathbb{P}(x_2|d_1, x_1)$. The quantity $\int [\phi(x_1) \otimes \mu_{x_2}(d_1, x_1)]\mathbb{P}(x_1)$ further encodes $\mathbb{P}(x)$. Likewise for θ_0^{SDS}. As far as we know, this sequential embedding is an innovation in the kernel methods literature.

While these representations appear abstract, they are eminently useful for defining estimators with closed form solutions that can be computed in one line of code (after computing kernel matrices). For example for $\theta_0^{SATE}(d_1, d_2)$, our estimator will be $\hat{\theta}^{SATE}(d_1, d_2) = \langle \hat{\gamma}, \phi(d_1) \otimes \phi(d_2) \otimes \frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i}) \otimes \hat{\mu}_{x_2}(d_1, x_{1i})] \rangle_{\mathcal{H}}$. The estimator $\hat{\gamma}$ is a standard kernel ridge regression, with known closed form. The estimator $\hat{\mu}_{x_2}(d_1, x_1)$ is an appropriately defined kernel ridge regression.

Algorithm C.1 (Estimation of off-policy effects)**.** *Denote the empirical kernel matrices*

$$K_{D_1D_1} \in \mathbb{R}^{n \times n}, \quad K_{D_2D_2} \in \mathbb{R}^{n \times n}, \quad K_{X_1X_1} \in \mathbb{R}^{n \times n}, \quad K_{X_2X_2} \in \mathbb{R}^{n \times n}$$

calculated from observations drawn from population $\mathbb{P}$. Likewise denote the empirical kernel matrices

$$K_{\tilde{D}_1\tilde{D}_1} \in \mathbb{R}^{n\times n}, \quad K_{\tilde{D}_2\tilde{D}_2} \in \mathbb{R}^{n\times n}, \quad K_{\tilde{X}_1\tilde{X}_1} \in \mathbb{R}^{n\times n}, \quad K_{\tilde{X}_2\tilde{X}_2} \in \mathbb{R}^{n\times n}$$

calculated from observations drawn from population $\tilde{\mathbb{P}}$.

Denote by $\odot$ the element-wise product. Off-policy effect estimators have the closed form solutions

1. $\hat{\theta}^{SATE} = \frac{1}{n}\sum_{i=1}^{n} Y^T(K_{D_1D_1} \odot K_{D_2D_2} \odot K_{X_1X_1} \odot K_{X_2X_2} + n\lambda)^{-1}$
   $(K_{D_1d_1} \odot K_{D_2d_2} \odot K_{X_1x_{1i}} \odot \{K_{X_2X_2}(K_{D_1D_1} \odot K_{X_1X_1} + n\lambda_4)^{-1}(K_{D_1d_1} \odot K_{X_1x_{1i}})\})$
2. $\hat{\theta}^{SDS} = \frac{1}{n}\sum_{i=1}^{n} Y^T(K_{D_1D_1} \odot K_{D_2D_2} \odot K_{X_1X_1} \odot K_{X_2X_2} + n\lambda)^{-1}$
   $(K_{D_1d_1} \odot K_{D_2d_2} \odot K_{X_1\tilde{x}_{1i}} \odot \{K_{X_2\tilde{X}_2}(K_{\tilde{D}_1\tilde{D}_1} \odot K_{\tilde{X}_1\tilde{X}_1} + n\lambda_4)^{-1}(K_{\tilde{D}_1d_1} \odot K_{\tilde{X}_1\tilde{x}_{1i}})\})$

where (λ, λ_4) are ridge regression penalty parameters.

We give theoretical values for (λ, λ_4) in Appendix I that balance bias and variance. We give a practical tuning procedure in Appendix J based on leave-one-out cross validation.

C.4 Consistency

Towards a guarantee of uniform consistency, we place regularity conditions on the original spaces and scalar-valued RKHSs.

Assumption C.3 (Original space regularity conditions)**.** *Assume*

1. *$\mathcal{D}$ and $\mathcal{X}$ are Polish spaces, i.e. separable and completely metrizable topological spaces*
2. *Y is bounded, i.e. $\exists C < \infty$ such that $|Y| \leq C$ almost surely*

A Polish space may be low-, high-, or infinite-dimensional. Random variables with support in a Polish space may be discrete, continuous, or even text. For simplicity of argument, we require that outcome $Y \in \mathbb{R}$ is bounded.

Assumption C.4 (RKHS regularity conditions)**.** *Assume*

1. *$k_{\mathcal{D}}$ and $k_{\mathcal{X}}$ are continuous and bounded:*
$$\sup_{d\in\mathcal{D}} \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \leq \kappa_d, \quad \sup_{x\in\mathcal{X}} \|\phi(x)\|_{\mathcal{H}_{\mathcal{X}}} \leq \kappa_x$$
2. *$\phi(d)$ and $\phi(x)$ are measurable*
3. *$k_{\mathcal{X}}$ is characteristic*

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

Next, we assume the prediction γ_0 is smooth.

Assumption C.5 (Smoothness of prediction)**.** *Assume*

1. *the prediction is well-specified, i.e. $\gamma_0 \in \mathcal{H}$*
2. *the prediction is a particularly smooth element of $\mathcal{H}$. Formally, define the covariance operator T for $\mathcal{H}$. We assume $\exists g \in \mathcal{H}$ s.t. $\gamma_0 = T^{\frac{c-1}{2}} g$, $c \in (1, 2]$, and $\|g\|_{\mathcal{H}}^2 \leq \zeta$*

For θ_0^{SATE} and θ_0^{SDS}, $T := \mathbb{E}[\{\phi(D_1) \otimes \phi(D_2) \otimes \phi(X_1) \otimes \phi(X_2)\} \otimes \{\phi(D_1) \otimes \phi(D_2) \otimes \phi(X_1) \otimes \phi(X_2)\}]$.

For learning problems with conditional mean embeddings, we place further smoothness conditions on the corresponding conditional expectation operators.

Assumption C.6 (Smoothness for θ_0^{SATE})**.** *Assume*

1. *the conditional expectation operator E_4 is well-specified as a Hilbert-Schmidt operator between RKHSs, i.e. $E_4 \in \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X})$, where*

$$E_4 : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X}, \quad f(\cdot) \mapsto \mathbb{E}[f(X_2)|D_1 = \cdot, X_1 = \cdot]$$

2. *the conditional expectation operator is a particularly smooth element of $\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X})$. Formally, define the covariance operator $T_4 := \mathbb{E}[\{\phi(D_1) \otimes \phi(X_1)\} \otimes \{\phi(D_1) \otimes \phi(X_1)\}]$ for $\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X})$. We assume $\exists G_4 \in \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X})$ s.t. $E_4 = (T_4)^{\frac{c_4-1}{2}} \circ G_4$, $c_4 \in (1, 2]$, and $\|G_4\|^2_{\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X})} \leq \zeta_4$*

Assumption C.7 (Smoothness for θ_0^{SDS})**.** *Assume*

1. *the conditional expectation operator E_5 is well-specified as a Hilbert-Schmidt operator between RKHSs, i.e. $E_5 \in \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X})$, where*

$$E_5 : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X}, \quad f(\cdot) \mapsto \mathbb{E}_{\tilde{\mathbb{P}}}[f(X_2)|D_1 = \cdot, X_1 = \cdot]$$

2. *the conditional expectation operator is a particularly smooth element of $\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X})$. Formally, define the covariance operator $T_5 := \mathbb{E}[\{\phi(D_1) \otimes \phi(X_1)\} \otimes \{\phi(D_1) \otimes \phi(X_1)\}]$ for $\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X})$. We assume $\exists G_5 \in \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X})$ s.t. $E_5 = (T_5)^{\frac{c_5-1}{2}} \circ G_5$, $c_5 \in (1, 2]$, and $\|G_5\|^2_{\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X})} \leq \zeta_5$*

With these assumptions, we arrive at our third main result.

Theorem C.3 (Consistency)**.** *Suppose Assumptions C.1, C.3, C.4, and C.5 hold.*

1. *If in addition Assumption C.6 holds then*

$$\|\hat{\theta}^{SATE} - \theta_0^{SATE}\|_\infty = O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + n^{-\frac{1}{2}\frac{c_4-1}{c_4+1}}\right)$$

2. *If in addition Assumptions C.2 and C.7 hold then*

$$\|\hat{\theta}^{SDS} - \theta_0^{SDS}\|_\infty = O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + \tilde{n}^{-\frac{1}{2}\frac{c_5-1}{c_5+1}}\right)$$

Exact finite sample rates are given in Appendix I. These rates are at best $n^{-\frac{1}{6}}$, by setting $(c, c_4, c_5) = 2$. The slow rates reflect the challenge of a sup-norm guarantee, which is much stronger than a prediction guarantee, and which encodes caution about worst-case scenarios when informing policy decisions. Our analysis is agnostic about the spectral decay; further assumptions on spectral decay (interpretable as effective dimension) will lead to faster rates of prediction.

D Graphical effect

D.1 An alternative language for causal inference

In computer science, causal directed acyclic graphs (DAGs) rather than potential outcomes provide a language for causal inference [114]. Rather than reasoning about $\mathbb{P}(Y^{(d)})$ one reasons about $\mathbb{P}(Y|do(D = d))$. [72] draws the connection between the do operator and structural equations more familiar in econometrics [62]. Both expressions $\mathbb{P}(Y^{(d)})$ and $\mathbb{P}(Y|do(D = d))$ are concerned with the distribution of outcome Y given intervention $D = d$. In our view, the two languages are complementary: for a specific setting, graphical criteria in terms of the DAG can help verify conditional independence statements in terms of potential outcomes.

In this section, we provide identification results in terms of causal DAGs, analogous to the identification results in terms of potential outcomes given in the main text. In doing so, we emphasize that the algorithms provided in the present work also apply to a variety of settings that appear in computer science research. In particular, we focus on back-door and front-door criteria, which are the fundamental building blocks of DAG-based causal inference.

Assume the analyst has access to a causal DAG G with vertex set W, partitioned into four disjoint sets $W = \{Y, D, X, U\}$. Y is the outcome, D is the set of treatments, X is the set of covariates, and

U is the set of unobserved variables. In the language of a *control problem*, the sets $\{Y, D, X\}$ may be denoted $\{R, A, S\}$ for reward, action, and state.

Since counterfactual inquiries involve intervention on the graph G, we require notation for graph modification. Denote by $G_{\bar{D}}$ the graph obtained by deleting from G all arrows pointing into nodes in D. Denote by $G_{\underline{D}}$ the graph obtained by deleting from G all arrows emerging from nodes in D. We denote d-separation by $\perp\!\!\!\perp_d$. Note that d-separation implies statistical independence. Throughout this section, we make the standard faithfulness assumption: d-connection implies statistical dependence.

D.2 Treatment effect

We define treatment effects in terms of the do-operator on the DAG. For clarity of exposition, we focus on the case where (D, Y) are nodes rather than sets of nodes.

Definition D.1 (Treatment effects: DAG)**.** $\check{\theta}_0^{ATE}(d) := \mathbb{E}[Y|do(D = d)]$ *is the counterfactual mean outcome given intervention* $D = d$ *for the entire population*

In the seminal works [111, 112], the author states sufficient conditions under which this treatment effect–a philosophical quantity defined in terms of interventions on the graph–can be measured from empirical quantities such as outcomes Y, treatments D, and covariates X. We present two sets of sufficient conditions, known as the *back-door criterion* and *front-door criterion*.

Assumption D.1 (Back-door criterion)**.** *Assume*

1. *no node in* X *is a descendent of* D
2. X *blocks every path between* D *and* Y *that contains an arrow into* D*. In other words*

$$(Y \perp\!\!\!\perp_d D|X)_{G_{\underline{D}}}$$

Intuitively, the analyst requires sufficiently many and sufficiently well placed covariates X in the context of the graph G. Assumption D.1 is satisfied if there is no unobserved confounding U, or if any unobserved confounder U with a back-door path into treatment D is blocked by X.

Assumption D.2 (Front door criterion)**.** *Assume*

1. X *intercepts all directed paths from* D *to* Y
2. *there is no unblocked back-door path from* D *to* X
3. *all back-door paths from* X *to* Y *are blocked by* D
4. $\mathbb{P}(d, x) > 0$

Intuitively, these conditions ensure that X serves to block all spurious paths from D to Y; to leave all directed paths unperturbed; and to create no new spurious paths.

As before, define the prediction

$$\gamma_0(d, x) := \mathbb{E}[Y|D = d, X = x]$$

Theorem D.1 (Identification of treatment effects: DAG [111, 112])**.** *Depending on which criterion holds, the causal parameter* $\check{\theta}_0^{ATE}(d)$ *has different expressions*

1. *If Assumption D.1 holds then* $\check{\theta}_0^{ATE}(d) = \int \gamma_0(d, x)\mathbb{P}(x)$
2. *If Assumption D.2 holds then* $\check{\theta}_0^{ATE}(d) = \int \gamma_0(d', x)\mathbb{P}(d')\mathbb{P}(x|d)$

Comparing Theorem D.1 with Theorem 2.1, we see that if Assumption D.1 holds then our estimator $\hat{\theta}^{ATE}$ for θ_0^{ATE} is also a uniformly consistent estimator of $\check{\theta}_0^{ATE}$. In the remainder of this section, we therefore focus on what happens if Assumption D.2 holds instead. We maintain notation from Section 2.

Theorem D.2 (Representation of treatment effects: DAG)**.** *If* $\gamma_0 \in \mathcal{H}$ *and Assumption D.2 holds then*

$$\check{\theta}_0^{ATE} = \langle \gamma_0, \mu_d \otimes \mu_x(d) \rangle_{\mathcal{H}}$$

where $\mu_d := \int \phi(d)\mathbb{P}(d)$ *and* $\mu_x(d) := \int \phi(x)\mathbb{P}(x|d)$.

The quantity $\mu_d := \int \phi(d)\mathbb{P}(d)$ is the mean embedding of $\mathbb{P}(d)$. The quantity $\mu_x(d) := \int \phi(x)\mathbb{P}(x|d)$ is the conditional mean embedding of $\mathbb{P}(x|d)$.

While this representation appears abstract, it is in fact eminently useful for defining an estimator with a closed form solution that can be computed in one line of code (after computing kernel matrices). For $\check{\theta}_0^{ATE}(d)$, our estimator will be $\hat{\theta}^{FD}(d) = \langle \hat{\gamma}, \hat{\mu}_d \otimes \hat{\mu}_x(d) \rangle_{\mathcal{H}}$. The estimator $\hat{\gamma}$ is a standard kernel ridge regression, with known closed form. The estimator $\hat{\mu}_d$ is an empirical mean. The estimator $\hat{\mu}_x(d)$ is an appropriately defined kernel ridge regression.

Algorithm D.1 (Estimation of treatment effects: DAG)**.** *Denote the empirical kernel matrices*

$$K_{DD} \in \mathbb{R}^{n \times n}, \quad K_{XX} \in \mathbb{R}^{n \times n}$$

K_{DD} and K_{XX} are calculated from observations drawn from population $\mathbb{P}$. Denote by $\odot$ the element-wise product. The front-door criterion estimator has the closed form solution

$$\hat{\theta}^{FD}(d) = \frac{1}{n} \sum_{i=1}^{n} Y^T (K_{DD} \odot K_{XX} + n\lambda)^{-1} (K_{Dd_i} \odot \{K_{XX}(K_{DD} + n\lambda_1)^{-1} K_{Dd}\})$$

where (λ, λ_1) are ridge regression penalty hyper-parameters.

We give theoretical values for (λ, λ_1) in Appendix I that balance bias and variance. We give a practical tuning procedure in Appendix J based on leave-one-out cross validation.

Towards a guarantee of uniform consistency, we place the same assumptions as in Section 2.

Theorem D.3 (Consistency: DAG)**.** *Suppose Assumptions 2.3, 2.4, 2.5, 2.6, and D.2 hold. Then*

$$\|\hat{\theta}^{FD} - \check{\theta}_0^{ATE}\|_\infty = O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + n^{-\frac{1}{2}\frac{c_1-1}{c_1+1}}\right)$$

Exact finite sample rates are given in Appendix I. These rates are at best $n^{-\frac{1}{6}}$, by setting $(c, c_1) = 2$. The slow rates reflect the challenge of a *sup*-norm guarantee, which is much stronger than a prediction guarantee, and which encodes caution about worst-case scenarios when informing policy decisions. Our analysis is agnostic about the spectral decay; further assumptions on spectral decay (interpretable as effective dimension) will lead to faster rates of prediction.

D.3 Off-policy planning

For sequential settings, we now allow $D = \{D_t\}_{t=1:T}$ to refer to the set of treatments. The ordering $t = 1 : T$ is such that every D_t is a non-descendent of $D_{t'}$ with $t' > t$ in G. Y is a descendent of D_T. We define off-policy effects in terms of the do-operator on the the DAG.

Definition D.2 (Off-policy planning: DAG)**.** *$\check{\theta}_0^{SATE}(d_{1:T}) := \mathbb{E}[Y|do(D_{1:T} = d_{1:T})]$ is the counterfactual mean outcome given interventions $D_{1:T} = d_{1:T}$ for the entire population*

Clearly, $\check{\theta}_0^{SATE}(d_{1:T})$ is a sequential generalization of $\check{\theta}_0^{ATE}(d)$. In the seminal work [115], the authors state sufficient conditions under which this off-policy effect–a philosophical quantity defined in terms of interventions on the graph–can be measured from empirical quantities such as outcomes Y, treatments $D_{1:T}$, and covariates $X_{1:T}$. This collection of sufficient conditions is known as the *sequential back-door criterion*.

Assumption D.3 (Sequential back-door criterion)**.** *Assume that for all $t = 1 : T$, there exists a set X_t of covariates such that*

1. *X_t consists of non-descendants of $\{D_{t:T}\}$*
2. *$(Y \perp\!\!\!\perp_d D_t | D_{1:t-1} X_{1:t})_{G_{\underline{D}_t, \bar{D}_{t+1:T}}}$*

As expected, Assumption D.3 is a sequential generalization of Assumption D.1. Finally, we see a sequential generalization of Theorem D.1. Define the prediction

$$\gamma_0(d_{1:T}, x_{1:T}) := \mathbb{E}[Y|D_{1:T} = d_{1:T}, X_{1:T} = x_{1:T}]$$

Theorem D.4 (Identification of off-policy effects: DAG [115])**.** *If Assumption D.3 holds then*

$$\check{\theta}_0^{SATE}(d_{1:T}) = \int \gamma_0(d_{1:T}, x_{1:T}) \mathbb{P}(x_1) \prod_{t=2}^{T} \mathbb{P}(x_t | d_{1:t-1}, x_{1:t-1})$$

Comparing Theorem D.4 with Theorem C.1, we see that our estimator $\hat{\theta}^{SATE}$ for θ_0^{SATE} is also a uniformly consistent estimator of $\check{\theta}_0^{SATE}$.

E Distribution effect

E.1 Learning problem

In the main text, we study target parameters defined as *means* of potential outcomes. In fact, the framework of causal adjustment extends to target parameters defined as *distributions* of potential outcomes. In this section, we focus on how to extend the algorithms and analyses presented in the main text to distribution target parameters.

For clarity, we focus on the distributional generalization of two simple parameters, $\theta_0^{ATE}(d)$ and $\theta_0^{ATT}(d, d')$. As a distribution, a target parameter can be encoded by a kernel mean embedding using a new feature map $\phi(y)$ for a new scalar-valued RKHS $\mathcal{H}_\mathcal{Y}$ corresponding to the outcome space $\mathcal{Y}$. In this formulation, we will allow the outcomes Y to be discrete, continuous, or text.

Definition E.1 (Distribution effects and embeddings)**.** *We define the distribution effects and their mean embeddings by*

1. $\theta_0^{DATE}(d) := \mathbb{P}(Y^{(d)})$ *is the counterfactual distribution of outcomes given intervention* $D = d$ *for the entire population.*
2. $\check{\theta}_0^{DATE}(d) := \mathbb{E}[\phi(Y^{(d)}]$ *is the embedding of the counterfactual distribution of outcomes given intervention* $D = d$ *for the entire population.*
3. $\theta_0^{DATT}(d, d') := \mathbb{P}(Y^{(d')}|D = d)$ *is the counterfactual distribution of outcomes given intervention* $D = d'$ *for the sub-population who actually received treatment* $D = d$*.*
4. $\check{\theta}_0^{DATT}(d, d') := \mathbb{E}[\phi(Y^{(d')}|D = d]$ *is the embedding of the counterfactual distribution of outcomes given intervention* $D = d'$ *for the sub-population who actually received treatment* $D = d$*.*

Likewise, distribution target parameters can be defined generalizing all quantities in Table 1.

E.2 Identification

The same identification results for the mean target parameters in the main text apply to distribution target parameters.

Theorem E.1 (Identification of distribution effects and embeddings)**.** *If Assumption 2.1 holds then*

1. $\theta_0^{DATE}(d) = \int \mathbb{P}(Y|D = d, X = x)\mathbb{P}(x)$
2. $\check{\theta}_0^{DATE}(d) = \int \mathbb{E}[\phi(Y)|D = d, X = x]\mathbb{P}(x)$
3. $\theta_0^{DATT}(d, d') = \int \mathbb{P}(Y|D = d', X = x)\mathbb{P}(x|d)$
4. $\check{\theta}_0^{DATT}(d, d') = \int \mathbb{E}[\phi(Y)|D = d', X = x]\mathbb{P}(x|d)$

Likewise, Theorems 2.1, B.1, and C.1 generalize for the remaining distribution target parameters. Clearly, the identification result for $\check{\theta}_0^{DATE}(d)$ and $\check{\theta}_0^{DATT}(d, d')$ resembles those presented in the main text. Define the prediction

$$\gamma_0(d, x) := \mathbb{E}[\phi(Y)|D = d, X = x]$$

Then we can write the latter result in the familiar form

$$\check{\theta}_0^{DATE}(d) = \int \gamma_0(d, x)\mathbb{P}(x)$$

$$\check{\theta}_0^{DATT}(d, d') = \int \gamma_0(d', x)\mathbb{P}(x|d)$$

E.3 Algorithm

To estimate θ_0^{DATE} and θ_0^{DATT}, we extend the RKHS construction used for θ_0^{ATE} and θ_0^{ATT}. As before, define scalar-valued RKHSs for treatment D and covariates X. As previewed above, define an additional scalar-valued RKHS for outcome Y. Let $k_{\mathcal{D}} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and $k_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be measureable positive definite kernels corresponding to scalar-valued RKHSs $\mathcal{H}_{\mathcal{D}}$, $\mathcal{H}_{\mathcal{X}}$, and $\mathcal{H}_{\mathcal{Y}}$. Denote the feature maps

$$\phi_{\mathcal{D}} : \mathcal{D} \to \mathcal{H}_{\mathcal{D}}, \quad d \mapsto k_{\mathcal{D}}(d, \cdot) \quad \phi_{\mathcal{X}} : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}, \quad x \mapsto k_{\mathcal{X}}(x, \cdot) \quad \phi_{\mathcal{Y}} : \mathcal{Y} \to \mathcal{H}_{\mathcal{Y}}, \quad y \mapsto k_{\mathcal{Y}}(y, \cdot)$$

To lighten notation, we will suppress subscripts when arguments are provided, e.g. we will write $\phi(d) = \phi_{\mathcal{D}}(d)$.

Because the prediction γ_0 is now a conditional mean embedding, we will present a construction involving a conditional expectation operator. Define the conditional expectation operator

$$E_6 : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}}, \quad f(\cdot) \mapsto \mathbb{E}[f(Y)|D = \cdot, X = \cdot]$$

Importantly, by construction

$$\gamma_0(d, x) = E_6^*[\phi(d) \otimes \phi(x)]$$

Note that $E_6 \in \mathcal{L}_2(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})$ is analogous to E_ρ in [129].

With this RKHS construction, we obtain a representation of the intermediate target parameters $\check{\theta}_0^{DATE}$ and $\check{\theta}_0^{DATT}$ as evaluations of E_6^*.

Theorem E.2 (Representation of distribution embeddings)**.** *If $E_6 \in \mathcal{L}_2(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})$ then*

1. $\check{\theta}_0^{DATE}(d) = E_6^*[\phi(d) \otimes \mu_x]$ *where* $\mu_x := \int \phi(x)\mathbb{P}(x)$
2. $\check{\theta}_0^{DATE}(d) = E_6^*[\phi(d') \otimes \mu_x(d)]$ *where* $\mu_x(d) := \int \phi(x)\mathbb{P}(x|d)$

The quantity $\mu_x := \int \phi(x)\mathbb{P}(x)$ is called the mean embedding of $\mathbb{P}(x)$. It encodes the distribution $\mathbb{P}(x)$ as a vector $\mu_x \in \mathcal{H}_{\mathcal{X}}$ such that the target parameter $\check{\theta}_0^{DATE}(d)$ can be expressed as an evaluation of E_6^*. Likewise, the quantity $\mu_x(d) := \int \phi(x)\mathbb{P}(x|d)$ is the mean embedding of $\mathbb{P}(x|d)$.

While these representations appear abstract, they are in fact eminently useful for defining estimators with closed form solutions that can be computed in one line of code (after computing kernel matrices). For example, for $\check{\theta}_0^{DATE}(d)$, our estimator will be $\hat{\theta}^{DATE}(d) = \hat{E}_6^*[\phi(d) \otimes \hat{\mu}_x]$. The estimators $\hat{E}_6$ and $\hat{\mu}_x(d)$ are generalized kernel ridge regressions, with known closed form. The estimator $\hat{\mu}_x$ is an empirical mean.

Algorithm E.1 (Estimation of distribution embeddings)**.** *Denote the empirical kernel matrices*

$$K_{DD} \in \mathbb{R}^{n \times n}, \quad K_{XX} \in \mathbb{R}^{n \times n} \quad K_{YY} \in \mathbb{R}^{n \times n}$$

K_{DD}, K_{XX}, and K_{YY} are calculated from observations drawn from population $\mathbb{P}$. Denote by $\odot$ the element-wise product. The distribution embedding estimators have the closed form solutions

1. $\hat{\theta}_0^{DATE}(d) = \frac{1}{n}\sum_{i=1}^n K_{\cdot Y}(K_{DD} \odot K_{XX} + n\lambda_6)^{-1}(K_{Dd} \odot K_{Xx_i})$
2. $\hat{\theta}_0^{DATT}(d, d') = K_{\cdot Y}(K_{DD} \odot K_{XX} + n\lambda_6)^{-1}(K_{Dd'} \odot [K_{XX}(K_{DD} + n\lambda_1)^{-1}K_{Dd}])$

where $(\lambda, \lambda_1, \lambda_6)$ are ridge regression penalty hyper-parameters.

We give theoretical values for $(\lambda, \lambda_1, \lambda_6)$ in Appendix I that balance bias and variance. We give a practical tuning procedure in Appendix J based on leave-one-out cross validation.

Importantly, Algorithm E.1 estimates counterfactual distribution *embeddings*. The ultimate parameters of interest, however, are counterfactual distributions $\theta_0^{DATE}(d)$ and $\theta_0^{DATT}(d, d')$. To recover the distribution from the distribution embedding, we present a deterministic procedure that uses the distribution embedding to provide samples $\{\tilde{y_j}\}$ from the distribution. The procedure is a variant of kernel herding, adapted from [147, 104].

Algorithm E.2 (Estimation of distribution effects)**.** *Recall that $\hat{\theta}_0^{DATE}(d)$ and $\hat{\theta}_0^{DATT}(d, d')$ are mappings from $\mathcal{Y}$ to $\mathbb{R}$.*

1. *Given* $\hat{\theta}_0^{DATE}(d)$*, calculate*

$$\tilde{y}_1 = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \left\{ [\hat{\theta}_0^{DATE}(d)](y) \right\}$$

$$\tilde{y}_j = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \left\{ [\hat{\theta}_0^{DATE}(d)](y) - \frac{1}{j+1} \sum_{\ell=1}^{j-1} k_{\mathcal{Y}}(\tilde{y}_\ell, y) \right\}, \quad j > 1$$

2. *Given* $\hat{\theta}_0^{DATT}(d, d')$*, calculate*

$$\tilde{y}_1 = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \left\{ [\hat{\theta}_0^{DATT}(d, d')](y) \right\}$$

$$\tilde{y}_j = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \left\{ [\hat{\theta}_0^{DATE}(d)](y) - \frac{1}{j+1} \sum_{\ell=1}^{j-1} k_{\mathcal{Y}}(\tilde{y}_\ell, y) \right\}, \quad j > 1$$

By this procedure, samples from $\theta_0^{DATE}(d)$ and $\theta_0^{DATT}(d, d')$ are easy to compute for any given values (d, d'). With such samples, one may visualize a histogram as an estimator of the counterfactual density of potential outcomes given values (d, d'). Alternatively, one may test statistical hypotheses using samples $\{\tilde{y_j}\}$.

E.4 Consistency

Towards a guarantee of uniform consistency, we place regularity conditions on the original spaces and scalar-valued RKHSs.

Assumption E.1 (Original space regularity conditions)**.** *Assume* $\mathcal{D}$*,* $\mathcal{X}$*, and* $\mathcal{Y}$ *are Polish spaces, i.e. separable and completely metrizable topological spaces*

A Polish space may be low-, high-, or infinite-dimensional. Random variables with support in a Polish space may be discrete, continuous, or even text.

Assumption E.2 (RKHS regularity conditions)**.** *Assume*

1. $k_{\mathcal{D}}$*,* $k_{\mathcal{X}}$*, and* $k_{\mathcal{Y}}$ *are continuous and bounded:*

$$\sup_{d \in \mathcal{D}} \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \leq \kappa_d, \quad \sup_{x \in \mathcal{X}} \|\phi(x)\|_{\mathcal{H}_{\mathcal{X}}} \leq \kappa_x, \quad \sup_{y \in \mathcal{Y}} \|\phi(y)\|_{\mathcal{H}_{\mathcal{Y}}} \leq \kappa_y$$

2. $\phi(d)$*,* $\phi(x)$*, and* $\phi(y)$ *are measurable*

3. $k_{\mathcal{X}}$ *and* $k_{\mathcal{Y}}$ *are characteristic*

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

Next, we assume the prediction γ_0 is smooth, this time parameterized by smoothness of the conditional expectation operator E_6.

Assumption E.3 (Smoothness of prediction)**.** *Assume*

1. *the conditional expectation operator* E_6 *is well-specified as a Hilbert-Schmidt operator between RKHSs, i.e.* $E_6 \in \mathcal{L}_2(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})$*, where*

$$E_6 : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{D}}, \quad f(\cdot) \mapsto \mathbb{E}[f(X)|D = \cdot]$$

2. *the conditional expectation operator is a particularly smooth element of* $\mathcal{L}_2(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})$*. Formally, define the covariance operator* $T_6 := \mathbb{E}[\{\phi(D) \otimes \phi(X)\} \otimes \{\phi(D) \otimes \phi(X)\}]$ *for* $\mathcal{L}_2(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})$*. We assume* $\exists G_6 \in \mathcal{L}_2(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})$ *s.t.* $E_6 = (T_6)^{\frac{c_6 - 1}{2}} \circ G_6$*,* $c_6 \in (1, 2]$*, and* $\|G_6\|^2_{\mathcal{L}_2(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}})} \leq \zeta_6$

With these assumptions, we arrive at our penultimate result.

Theorem E.3 (Consistency)**.** *Suppose Assumptions 2.1, E.1, E.2, and E.3 hold.*

1. *Then*
$$\sup_{d\in\mathcal{D}} \|\hat{\theta}^{DATE}(d) - \check{\theta}_0^{DATE}(d)\|_{\mathcal{H}_\mathcal{Y}} = O_p\left(n^{-\frac{1}{2}\frac{c_6-1}{c_6+1}}\right)$$
2. *If in addition Assumption 2.6 holds, then*
$$\sup_{d\in\mathcal{D}} \|\hat{\theta}^{DATE}(d) - \check{\theta}_0^{DATE}(d)\|_{\mathcal{H}_\mathcal{Y}} = O_p\left(n^{-\frac{1}{2}\frac{c_6-1}{c_6+1}} + n^{-\frac{1}{2}\frac{c_1-1}{c_1+1}}\right)$$

Exact finite sample rates are given in Appendix I. These rates are at best $n^{-\frac{1}{6}}$, by setting $(c_1, c_6) = 2$. The slow rates reflect the challenge of a uniform guarantee, which encodes caution about worst-case scenarios when informing policy decisions. Our analysis is agnostic about the spectral decay; further assumptions on spectral decay (interpretable as effective dimension) will lead to faster rates of prediction.

Finally, we state an additional regularity condition under which we can prove that the samples $\{\tilde{y}_j\}$ calculated from the distribution embeddings weakly converge to the desired distribution.

Assumption E.4 (Regularity)**.** *Assume*

1. *$\mathcal{Y}$ is locally compact*
2. *$\mathcal{H}_\mathcal{Y} \subset \mathcal{C}_0$, where $\mathcal{C}_0$ is the space of bounded, continuous, real-valued functions that vanish at infinity*

As discussed by [128], the combined assumptions that $\mathcal{Y}$ is Polish and locally compact can become restrictive. In particular, if $\mathcal{Y}$ is a Banach space, then to satisfy both conditions it must be finite-dimensional. Trivially, $\mathcal{Y} = \mathbb{R}$ satisfies both conditions. We arrive at our final result.

Theorem E.4 (Convergence in distribution)**.** *Suppose Assumptions 2.1, E.1, E.2, E.3, and E.4 hold. Suppose samples $\{\tilde{y}_j\}$ are calculated for $\theta_0^{DATE}(d)$ and $\theta_0^{DATT}(d, d')$ as described in Algorithm E.2.*

1. *Then $\{\tilde{y}_j\} \xrightarrow{d} \theta_0^{DATE}(d)$*
2. *If in addition Assumption 2.6 holds, then $\{\tilde{y}_j\} \xrightarrow{d} \theta_0^{DATT}(d, d')$*

Note that samples are drawn for given values (d, d'). Though our consistency results throughout the paper are uniform, this convergence in distribution result is pointwise.

F A gentle introduction to kernel methods

F.1 Kernel and feature map

Motivated by the close fit between the aims of causal inference and the capabilities of RKHSs, we provide an introduction to RKHS methods for a general economics audience. Appendix G formalizes this discussion. We present (1) the feature map and kernel as basic building blocks; (2) the covariance operator, which generalizes the covariance matrix; (3) the mean embedding, which is a way to encode a probability distribution; and finally, (4) the composite RKHS, which is composed from simpler RKHSs.

Polynomials, splines, and Sobolev spaces are intuitive examples of RKHSs. More generally, an abstract RKHS $\mathcal{H}_\mathcal{W}$ is a function space with elements that are functions $f : \mathcal{W} \to \mathbb{R}$. The original space $\mathcal{W}$ can be any Polish space: it can be low-, high-, or even infinite-dimensional. As such, random variable W with support $\mathcal{W}$ can be discrete, continuous, or even text. For example, one may take $\mathcal{W} = [0, 1]$.

An RKHS is fully characterized by its feature map. The feature map takes a point w in the original space $\mathcal{W}$, and maps it to a feature $\phi(w)$ in the RKHS $\mathcal{H}_\mathcal{W}$. For example, the feature map may be $\phi(w) = (1, \sqrt{2}w, w^2)^T$. The closure of $span\{\phi(w)\}_{w\in\mathcal{W}}$ is the RKHS $\mathcal{H}_\mathcal{W}$. In other words, $\{\phi(w)\}_{w\in\mathcal{W}}$ can be viewed as the dictionary of basis functions for the RKHS $\mathcal{H}_\mathcal{W}$. In the example

with $\mathcal{W} = [0,1]$ and $\phi(w) = (1, \sqrt{2}w, w^2)^T$, the RKHS $\mathcal{H}_\mathcal{W}$ consists of quadratic functions over the unit interval.

The kernel $k : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ is the inner product of features $\phi(w)$ and $\phi(w')$ rather than original values w and w'. Formally, $k(w, w') = \langle \phi(w), \phi(w') \rangle_{\mathcal{H}_\mathcal{W}}$. In the running example, $k(w, w') = (1, \sqrt{2}w, w^2)(1, \sqrt{2}w', \{w'\}^2)^T = (w \cdot w' + 1)^2$. The kernel k encodes notions of angles and distances in the RKHS. Though we have constructed the kernel from the feature map, it turns out that any positive definite kernel k induces a corresponding feature map $\phi : w \mapsto k(w, \cdot)$. For commonly used kernels, the RKHS $\mathcal{H}_\mathcal{W}$ is dense in $L^2(\mathcal{W})$, the space of square integrable functions classically used in non-parametric statistics [134, 135]. See [133, Example 1] for kernels corresponding to splines and Sobolev spaces.

We have already seen that if $f \in \mathcal{H}_\mathcal{W}$, then $f : \mathcal{W} \to \mathbb{R}$. With the additional notation of the feature map, we write $f : w \mapsto \langle f, \phi(w) \rangle_{\mathcal{H}_\mathcal{W}}$. In our running example, the function $f(w) = w^2 + 2w + 3$ can be represented as $f(w) = (3, \sqrt{2}, 1)(1, \sqrt{2}w, w^2)^T$, so $f = (3, \sqrt{2}, 1)^T$ in this RKHS. Importantly, $\phi(w)$ and hence f may be infinite-dimensional quantities. By Cauchy-Schwarz, we see the relation between RKHS norm and sup-norm frequently used in this work: $\|f\|_\infty \leq \|f\|_{\mathcal{H}_\mathcal{W}} \cdot \sup_{w \in \mathcal{W}} \|\phi(w)\|_{\mathcal{H}_\mathcal{W}}$.

For computation, the kernel is much more convenient than the feature map. Therefore, the so-called *kernel trick* in machine learning is to write a linear algorithm in terms of inner products so that a nonlinear generalization of the linear algorithm is equally easy to compute. In practice, kernel algorithms are written in terms of the kernel matrix $K_{WW} \in \mathbb{R}^{n \times n}$ with (i,j)-th entry $k(w_i, w_j)$. Ridge regression in the RKHS can be implemented in one line: $\hat{f}(w) = Y^T (K_{WW} + n\lambda)^{-1} K_{Ww}$, where $K_{Ww} \in \mathbb{R}^n$ has i-th entry $k(w_i, w)$. In the running example, we can implement ridge regression in the space of quadratic functions by constructing K_{WW} with (i,j)-th entry $(w_i \cdot w_j + 1)^2$ and K_{Ww} with i-th entry $(w_i \cdot w + 1)^2$.

F.2 Covariance operator

Next, we introduce a key statistical quantity for RKHS analysis. The covariance operator is $T = \mathbb{E}[\phi(W) \otimes \phi(W)]$ where $\otimes$ means tensor product. Recall tensor product notation: $[a \otimes b]c = a\langle b, c \rangle$. In the running example, $T = \mathbb{E} \begin{bmatrix} 1 & \sqrt{2}W & W^2 \\ \sqrt{2}W & 2W^2 & \sqrt{2}W^3 \\ W^2 & \sqrt{2}W^3 & W^4 \end{bmatrix}$. This quantity generalizes the uncentered covariance matrix $\mathbb{E}[WW^T]$. The relation between RKHS norm and L^2 norm (i.e. prediction norm) is $\|T^{\frac{1}{2}} f\|_\mathcal{H} = \|f\|_{L^2}$ [16]. In this sense, RKHS norm reflects the spectrum of the covariance T, which depends on both the choice of kernel and the population distribution. For comparison, L^2 norm depends on only the population distribution. In estimation, rates are parametrized by properties of T.

The *range-space* assumption requires that the target parameter $f_0 \in \mathcal{H}_\mathcal{W}$ is a particularly smooth element of the RKHS $\mathcal{H}_\mathcal{W}$ [130, 16]. Formally, one assumes that f_0 can be expressed as $f_0 = T^{\frac{c-1}{2}} g,\ c \in (1, 2],\ g \in \mathcal{H}_\mathcal{W}$. f_0 is some positive power of the covariance operator T applied to another element of the RKHS g. In other words, f_0 is well-approximated by the top of the spectrum of T, which facilitates analysis of approximation error. This condition appears under the name of the *source* condition in the econometrics literature on inverse problems [19, 18, 21, 20]. It is similar to the approximation error assumptions used in analyzing series-based estimators. Specifically, it generalizes to an arbitrary RKHS the degree of smoothness of f_0 usually defined for regression in Sobolev spaces [40, 142, 61, 34, 144]. A larger value of c corresponds to a smoother target f_0. [137, Theorem 4.6] formalizes how the range space of $T^{\frac{c-1}{2}}$ is equal to an interpolation space between L^2 and the RKHS. We will make such assumptions.

Researchers also place direct assumptions on the decay of the spectrum of T, interpretable as an assumption on the effective dimension of the statistical problem. The spectrum may have a finite number of nonzero eigenvalues; it may have eigenvalues that decay exponentially; or it may have eigenvalues that decay polynomially. Note that this behavior is a joint assumption over the kernel k and the data distribution $\mathbb{P}$. In our analysis, we do not make such assumptions; we are agnostic about the spectrum of T. For faster rates of prediction, one can place such spectral assumptions [149, 101, 16].

F.3 Mean embedding

We have seen that the feature map takes a value in the original space $w \in \mathcal{W}$ and maps it to a feature in the RKHS $\phi(w) \in \mathcal{H}_{\mathcal{W}}$. Now we wish to generalize this idea from values w in the original space to distributions $\mathbb{Q}$ over the original space.

Denote by δ_w the Dirac distribution at w. One interpretation of the role of the feature map is to provide a sufficiently rich representation of the distribution δ_w. Indeed, we may conceptualize the embedding $\delta_w \mapsto \mathbb{E}_{\delta_w}[\phi(W)] = \phi(w)$. Now, let us consider a general distribution $\mathbb{Q}$ over the original space. We may similarly conceptualize the embedding $\mathbb{Q} \mapsto \mathbb{E}_{\mathbb{Q}}[\phi(W)] := \mu$. Just as a value w in the original space is embedded as a vector $\phi(w)$ in the RKHS , so too the distribution $\mathbb{Q}$ over the original space is embedded as a vector μ in the RKHS. Mean embeddings faciliate the evaluation of expecations of RKHS functions: for $f \in \mathcal{H}_{\mathcal{W}}$, $\mathbb{E}_{\mathbb{Q}}[f(W)] = \langle f, \mu \rangle_{\mathcal{H}_{\mathcal{W}}}$. In the running example, $\mu = (1, \sqrt{2}\mathbb{E}[W], \mathbb{E}[W]^2)^T$, and we can confirm $\mathbb{E}[f(W)] = \mathbb{E}[W^2 + 2W + 3] = (3, \sqrt{2}, 1)(1, \sqrt{2}\mathbb{E}[W], \mathbb{E}[W]^2)^T$.

A natural question is whether the embedding $\mathbb{Q} \mapsto \mathbb{E}_{\mathbb{Q}}[\phi(W)]$ is injective, i.e. whether the RKHS vector representation is without loss. This property is called the *characteristic* property of the kernel k, and we will assume that it holds in the present work. The characteristic property holds for kernels commonly used by machine learning practitioners [134, 135, 133]. Though it does not hold for polynomials, it does hold for splines and Sobolev spaces over $\mathbb{R}^d$ [133].

F.4 Composite RKHS

Suppose we define the RKHSs $\mathcal{H}_{\mathcal{W}}$ and $\mathcal{H}_{\mathcal{Z}}$ with positive definite kernels $k_{\mathcal{W}} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ and $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, respectively. Recall that this implies an element $f \in \mathcal{H}_{\mathcal{W}}$ is a function $f : \mathcal{W} \to \mathbb{R}$ and an element $g \in \mathcal{H}_{\mathcal{Z}}$ is a function $g : \mathcal{Z} \to \mathbb{R}$. In the present work, we will make use of composite RKHSs that build on $\mathcal{H}_{\mathcal{W}}$ and $\mathcal{H}_{\mathcal{Z}}$. Specifically, we will make use of the tensor product RKHS and vector-valued RKHS.

The tensor product RKHS $\mathcal{H} := \mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{Z}}$ is the RKHS with the kernel $k : (\mathcal{W} \times \mathcal{Z}) \times (\mathcal{W} \times \mathcal{Z}) \to \mathbb{R},\ (w, z), (w', z') \mapsto k_{\mathcal{W}}(w, w') \cdot k_{\mathcal{Z}}(z, z')$. Equivalently, the tensor product RKHS $\mathcal{H}$ has feature map $\phi(w) \otimes \phi(z)$. An element of the tensor product RKHS $h \in \mathcal{H}$ is a function $h : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$. Taking $\mathcal{H}_{\mathcal{W}}$ and $\mathcal{H}_{\mathcal{Z}}$ to both be the RKHS of quadratic functions over the unit interval, the tensor product RKHS has kernel $k((w, z), (w', z')) = (w \cdot w' + 1)^2 \cdot (z \cdot z' + 1)^2$. In the present work, we will assume the prediction function $\gamma_0(w, z) := \mathbb{E}[Y|W = w, Z = z]$ is an element of a tensor product RKHS, i.e. $\gamma_0 \in \mathcal{H}$. We will estimate γ_0 by a kernel ridge regression in $\mathcal{H}$.

The vector-valued RKHS $\mathcal{L}_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{W}})$ is unlike the RKHSs discussed so far. Rather than being a space of real-valued functions, it is a space of Hilbert-Schmidt operators from one RKHS to another. If the operator $E \in \mathcal{L}_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{W}})$, then $E : \mathcal{H}_{\mathcal{Z}} \to \mathcal{H}_{\mathcal{W}}$. In the running example, E would be an operator from the RKHS of quadratic functions over the unit interval to the RKHS of quadratic functions over the unit interval. Formally, it can be shown that $\mathcal{L}_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{W}})$ is an RKHS in its own right with an appropriately defined kernel and feature map. In the present work, we will assume the conditional expectation operator $E : g(\cdot) \mapsto \mathbb{E}[g(Z)|W = \cdot]$ is an element of a vector-valued RKHS, i.e. $E \in \mathcal{L}_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{W}})$. We will estimate E by a kernel ridge regression in $\mathcal{L}_2(\mathcal{H}_{\mathcal{Z}}, \mathcal{H}_{\mathcal{W}})$, which in turn implies an estimator of the corresponding conditional mean embedding.

G A formal introduction to kernel methods

G.1 Scalar-valued RKHS

We briefly review the theory of the scalar-valued RKHS as it relates to causal adjustment. The primary reference is [136]. For clarity, we continue the abstract example in Section F, where $f \in \mathcal{H}_{\mathcal{W}}$.

G.1.1 Kernels

Definition G.1 (Definition 4.1 of [136]). *Let $\mathcal{W}$ be a non-empty set. A function $k : \mathcal{W} \times \mathcal{W} \to \mathbb{K}$ is a kernel on $\mathcal{X}$ if $\exists$ a $\mathbb{K}$-Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{W} \to \mathcal{H}$ s.t. $\forall w, w' \in \mathcal{W}$*

$$k(w, w') = \langle \phi(w), \phi(w') \rangle_{\mathcal{H}}$$

ϕ is the feature map and $\mathcal{H}$ is the feature space of k

Definition G.2 (Definition 4.15 of [136]). *A function $k : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ is positive definite if $\forall n \in \mathbb{N}$, $\alpha_1, ..., \alpha_n \in \mathbb{R}$, $w_1, ..., w_n \in \mathcal{W}$*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(w_i, w_j) \geq 0$$

k is symmetric if $k(w, w') = k(w', w)$ for all $w, w' \in \mathcal{W}$.

Proposition G.1 (Theorem 4.16 of [136]). *A function $k : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ is a kernel iff it is symmetric and positive definite.*

G.1.2 Reproducing kernels

Definition G.3 (Definition 4.18 of [136]). *Let $\mathcal{W} \neq \emptyset$ and $\mathcal{H}$ be a $\mathbb{K}$-Hilbert function space over $\mathcal{W}$, i.e. a $\mathbb{K}$-Hilbert space that consists of functions mapping from $\mathcal{W}$ into $\mathbb{K}$.*

1. *A function $k : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if $k(w, \cdot) \in \mathcal{H}$ for all $w \in \mathcal{W}$ and the reproducing property holds:*

$$f(w) = \langle f, k(w, \cdot) \rangle_{\mathcal{W}}, \quad \forall f \in \mathcal{H}, \quad \forall w \in \mathcal{W}$$

2. *The space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) over $\mathcal{W}$ if $\forall w \in \mathcal{W}$, the Dirac functional*

$$\delta_w : \mathcal{W} \to \mathbb{K}, \quad f \mapsto f(w)$$

is continuous

Proposition G.2 (Lemma 4.19 of [136]). *Let $\mathcal{H}$ be a Hilbert function space over $\mathcal{X}$ that has a reproducing kernel k. Then $\mathcal{H}$ is an RKHS and $\mathcal{H}$ is also a feature space of k, where the feature map is given by*

$$\phi : \mathcal{W} \to \mathcal{H}, \quad w \mapsto k(w, \cdot)$$

ϕ is the canonical feature map.

Proposition G.3 (Theorem 4.20 of [136]). *Let $\mathcal{H}$ be an RKHS over $\mathcal{W}$. Then*

$$k : \mathcal{W} \times \mathcal{W} \to \mathbb{K}, \quad w, w' \mapsto \langle \delta_w, \delta'_w \rangle_{\mathcal{H}}$$

is the only reproducing kernel of $\mathcal{H}$.

Proposition G.4 (Theorem 4.21 of [136]). *Let $\mathcal{W} \neq \emptyset$ and k be a kernel of $\mathcal{W}$ with feature space $\mathcal{H}_0$ and feature map $\phi_0 : \mathcal{W} \to \mathcal{H}_0$. Then*

$$\mathcal{H} := \{f : \mathcal{W} \to \mathbb{K} : \exists v \in \mathcal{H}_0 \text{ s.t. } \forall w \in \mathcal{W},\ f(w) = \langle v, \phi_0(w) \rangle_{\mathcal{H}_0}\}$$

equipped with the norm

$$\|f\|_{\mathcal{H}} := \inf\{\|v\|_{\mathcal{H}_0} : v \in \mathcal{H}_0 \text{ s.t. } f = \langle v, \phi_0(\cdot) \rangle_{\mathcal{H}_0}\}$$

is the only RKHS for which k is a reproducing kernel. Moreover, the set

$$\mathcal{H}_{pre} := \left\{ \sum_{i=1}^{n} \alpha_i k(w_i, \cdot) : n \in \mathbb{N}, \alpha_1, ..., \alpha_n \in \mathbb{K}, w_1, ..., w_n \in \mathcal{W} \right\}$$

is dense in $\mathcal{H}$ and for $f := \sum_{i=1}^{n} \alpha_i k(w_i, \cdot) \in \mathcal{H}_{pre}$,

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \bar{\alpha}_j k(x_i, x_j)$$

G.1.3 Properties

Proposition G.5 (eq. 4.15 of [136]). *A kernel k with feature map $\phi : \mathcal{W} \to \mathcal{H}$ is bounded iff*

$$\kappa := \sup_{w \in \mathcal{W}} \sqrt{k(w, w)} = \sup_{w \in \mathcal{W}} \|\phi(w)\|_{\mathcal{H}} < \infty$$

Proposition G.6 (Lemma 4.23 of [136]). *Let $\mathcal{W}$ be a set and k be a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$. Then k is bounded iff every $f \in \mathcal{H}$ is bounded.*

Proposition G.7 (Lemma 4.24 of [136]). *Let $\mathcal{W}$ be a measurable space and k be a kernel on $\mathcal{W}$ with RKHS $\mathcal{H}$. Then all $f \in \mathcal{H}$ are measurable iff $k(w, \cdot) : \mathcal{W} \to \mathbb{R}$ is measurable for all $w \in \mathcal{W}$*

Proposition G.8 (Lemma 4.25 of [136]). *Let $\mathcal{W}$ be a measurable space and k be a kernel on $\mathcal{W}$ s.t. $k(w, \cdot) : \mathcal{W} \to \mathbb{R}$ is measurable for all $w \in \mathcal{W}$. If the RKHS $\mathcal{H}$ of k is spearable, then both the canonical feature map $\phi : \mathcal{W} \to \mathcal{H}$ and $k : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ are measurable.*

Proposition G.9 (Lemma 4.33 of [136]). *If $\mathcal{W}$ is a separable topological space and k is a continous kernel on $\mathcal{W}$ then the RKHS of k is separable.*

G.2 Tensor-product RKHS

We briefly review the theory of the tensor-product RKHS as it relates to causal adjustment. The primary references are [136, 56, 129]. For clarity, we focus on the example of θ_0^{ATT}, where $\gamma_0 \in \mathcal{H} := \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X}$ and $T_1 := \mathbb{E}[\phi(d) \otimes \phi(d)]$.

G.2.1 Tensor product

Recall tensor product notation: if $a, b \in \mathcal{H}_1$ and $c \in \mathcal{H}_2$ then $[c \otimes a]b = c\langle a, b\rangle_{\mathcal{H}_1}$. Denote by $\mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)$ the space of Hilbert-Schmidt operators from $\mathcal{H}_1$ to $\mathcal{H}_2$.

Proposition G.10 (eq. 3.6 of [56]). *If $\mathcal{H}_1$ and $\mathcal{H}_2$ are separable RKHSs, then*

$$\|c \otimes a\|_{\mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)} = \|a\|_{\mathcal{H}_1} \|c\|_{\mathcal{H}_2}$$

and $c \otimes a \in \mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)$.

Proposition G.11 (eq. 3.7 of [56]). *Assume $\mathcal{H}_1$ and $\mathcal{H}_2$ are separable RKHSs. If $C \in \mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)$ then*

$$\langle C, c \otimes a\rangle_{\mathcal{L}_2(\mathcal{H}_1, \mathcal{H}_2)} = \langle c, Ca\rangle_{\mathcal{H}_2}$$

Proposition G.12. *Let $k_\mathcal{D}$ be a kernel on $\mathcal{D}$ and $k_\mathcal{X}$ be a kernel on $\mathcal{X}$. Then $k := k_\mathcal{D} \cdot k_\mathcal{X}$ is a kernel on $\mathcal{D} \times \mathcal{X}$. In particular,*

$$k : (\mathcal{D} \times \mathcal{X}) \times (\mathcal{D} \times \mathcal{X}) \to \mathbb{R}$$

and

$$\begin{aligned} k((d, x), (d', x')) &= k_\mathcal{D}(d, d') \cdot k_\mathcal{X}(x, x') \\ &= \langle \phi_\mathcal{D}(d), \phi_\mathcal{D}(d')\rangle_{\mathcal{H}_\mathcal{D}} \cdot \langle \phi_\mathcal{X}(x), \phi_\mathcal{X}(x')\rangle_{\mathcal{H}_\mathcal{X}} \\ &= \langle \phi_\mathcal{D}(d) \otimes \phi_\mathcal{X}(x), \phi_\mathcal{D}(d') \otimes \phi_\mathcal{X}(x')\rangle_{\mathcal{H}_\mathcal{D} \hat{\otimes} \mathcal{H}_\mathcal{X}} \\ &= \langle \phi_\mathcal{D}(d) \otimes \phi_\mathcal{X}(x), \phi_\mathcal{D}(d') \otimes \phi_\mathcal{X}(x')\rangle_{\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D})} \end{aligned}$$

where $\mathcal{H}_\mathcal{D} \hat{\otimes} \mathcal{H}_\mathcal{X}$ is the completion of $\mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X}$

Proof. Lemma 4.6 and Appendix A.5.2 of [136] □

G.2.2 Covariance operator

We prove existence of T_1 and derive its properties as a covariance operator. Identical arguments hold for all other covariance operators in the paper.

In Assumption 2.3, we assume that treatment space $\mathcal{D}$ is separable. In Assumption 2.4, we assume RKHS $\mathcal{H}_\mathcal{D}$ has continuous, bounded kernel $k_\mathcal{D}$ with feature map $\phi_\mathcal{D}$. By Proposition G.9, it follows that $\mathcal{H}_\mathcal{D}$ is separable, i.e. it has a countable orthonormal basis that we now denote $\{e_i^\mathcal{D}\}_{i=1}^\infty$.

Denote by $\mathcal{L}_2(\mathcal{H}_\mathcal{D}, \mathcal{H}_\mathcal{D})$ the space of Hilbert-Schmidt operators $A : \mathcal{H}_\mathcal{D} \to \mathcal{H}_\mathcal{D}$ with inner product $\langle A, B\rangle_{\mathcal{L}_2(\mathcal{H}_\mathcal{D}, \mathcal{H}_\mathcal{D})} = \sum_{i=1}^\infty \langle Ae_i^\mathcal{D}, Be_i^\mathcal{D}\rangle_{\mathcal{H}_\mathcal{D}}$. When it is contextually clear, we abbreviate the space as $\mathcal{L}_2$.

Proposition G.13 (Proposition 17 of [129]). *Suppose Assumptions 2.3 and 2.4 hold. Then $\exists T_1 \in \mathcal{L}_2(\mathcal{H}_\mathcal{D}, \mathcal{H}_\mathcal{D})$ s.t.*

$$\langle T_1, A\rangle_{\mathcal{L}_2} = \mathbb{E}\langle \phi(D) \otimes \phi(D), A\rangle_{\mathcal{L}_2}$$

Proposition G.14 (Proposition 18 of [129]). *Suppose Assumptions 2.3 and 2.4 hold.*

$$\langle \ell, T_1 \ell' \rangle_{\mathcal{H}_\mathcal{D}} = \mathbb{E}[\ell(D)\ell'(D)], \quad \forall \ell, \ell' \in \mathcal{H}_\mathcal{D}$$

Proposition G.15 (Proposition 19 of [129]). *Suppose Assumptions 2.3 and 2.4 hold.*

$$tr(T_1) \leq \kappa^2$$

Since covariance operator T_1 has finite trace, its eigendecomposition is well-defined. Recall that the prediction covariance operator T consists of functions from $\mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X}$ to $\mathcal{Y} = \mathbb{R}$. Since these functions have finite-dimensional output, it is immediate that T has finite trace and its eigendecomposition is well-defined [16, Remark 1].

Definition G.4. *The powers of operators T_1 and T are defined as*

$$T_1^a = \sum_{k=1}^{\infty} \nu_k^a e_k^\mathcal{D} \langle \cdot, e_k^\mathcal{D} \rangle_{\mathcal{H}_\mathcal{D}}$$
$$T^a = \sum_{k=1}^{\infty} \iota_k^a e_k \langle \cdot, e_k \rangle_{\mathcal{H}_\Omega}$$

where $(\{\nu_k\}, \{e_k^\mathcal{D}\})$ is the spectrum of T_1 and $(\{\iota_k\}, \{e_k\})$ is the spectrum of T.

G.3 Vector-valued RKHS

We briefly review the theory of the vector-valued RKHS as it relates to causal adjustment. The primary references are the appendices of [59, 129]. For clarity, we focus on the example of θ_0^{ATT}, where $E_1 \in \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D})$. Identical arguments hold for all other conditional expectation operators in the paper.

G.3.1 Vector-valued RKHS as tensor-product RKHS

Proposition G.16 (Theorem A.2 of [59]). *Let $I_{\mathcal{H}_\mathcal{D}} : \mathcal{H}_\mathcal{D} \to \mathcal{H}_\mathcal{D}$ be the identity operator. $\Gamma(h, h') = \langle h, h' \rangle_{\mathcal{H}_\mathcal{X}} I_{\mathcal{H}_\mathcal{D}}$ is a kernel of positive type.*

Proposition G.17 (Proposition 2.3 of [17]). *Consider a kernel of positive type $\Gamma : \mathcal{H}_\mathcal{X} \times \mathcal{H}_\mathcal{X} \to \mathcal{L}(\mathcal{H}_\mathcal{D})$, where $\mathcal{L}(\mathcal{H}_\mathcal{D})$ is the space of bounded linear operators from $\mathcal{H}_\mathcal{D}$ to $\mathcal{H}_\mathcal{D}$. It corresponds to a unique RKHS $\mathcal{H}_\Gamma$ with reproducing kernel Γ.*

Proposition G.18 (Theorem B.1 of [59]). *Each $E \in \mathcal{H}_\Gamma$ is a bounded linear operator $E : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{D}$.*

Proposition G.19. *$\mathcal{H}_\Gamma = \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{D})$ and the inner products are equal.*

Proof. [13, Theorem 13] and [60, eq. 12]. □

Proposition G.20 (Theorem B.2 of [59].). *If $\exists E, G \in \mathcal{H}_\Gamma$ s.t. $\forall x \in \mathcal{X},\ E\phi(x) = G\phi(x)$ then $E = G$. Furthermore, if $\phi(x)$ is continuous in x then it is sufficient that $E\phi(x) = G\phi(x)$ on a dense subset of $\mathcal{X}$.*

Proposition G.21 (Theorem B.3 of [59]). *$\forall E \in \mathcal{H}_\Gamma$, $\exists E^* \in \mathcal{H}_{\Gamma^*}$ where $\mathcal{H}_{\Gamma^*}$ is the vector-valued RKHS with reproducing kernel $\Gamma^*(\ell, \ell') = \langle \ell, \ell' \rangle_{\mathcal{H}_\mathcal{D}} I_{\mathcal{H}_\mathcal{X}}$. $\forall h \in \mathcal{H}_\mathcal{X}$ and $\forall \ell \in \mathcal{H}_\mathcal{D}$,*

$$\langle Eh, \ell \rangle_{\mathcal{H}_\mathcal{D}} = \langle h, E^* \ell \rangle_{\mathcal{H}_\mathcal{X}}$$

The operator $A \circ E = E^*$ is an isometric isomorphism from $\mathcal{H}_\Gamma$ to $\mathcal{H}_{\Gamma^*}$; $\mathcal{H}_\Gamma \cong \mathcal{H}_{\Gamma^*}$ and $\|E\|_{\mathcal{H}_\Gamma} = \|E^*\|_{\mathcal{H}_{\Gamma^*}}$.

Proposition G.22 (Theorem B.4 of [59].). *The set of self-adjoint operators in $\mathcal{H}_\Gamma$ is a closed linear subspace.*

G.3.2 Conditional expectation operator to conditional mean embedding

Proposition G.23 (Lemma 15 of [31]). *$\mathcal{H}_{\Gamma^*}$ is isometrically isomorphic to $\mathcal{H}_\Xi$, the vector-valued RKHS with reproducing kernel $\Xi(d, d') = k_{\mathcal{D}}(d, d')I_{\mathcal{H}_\mathcal{X}}$. $\forall \mu_x \in \mathcal{H}_\Xi$, $\exists! E^* \in \mathcal{H}_{\Gamma^*}$ s.t.*

$$\mu_x(d) = E^*\phi(d), \quad \forall d \in \mathcal{D}$$

Proposition G.24. *Under Assumption 2.4,*

$$\mathbb{E}_{X|D=d}[\gamma_0(d', X)] = \langle \gamma_0, \phi(d') \otimes [E_1\phi(X)](d)\rangle_\mathcal{H} = \langle \gamma_0, \phi(d') \otimes \mu_x(d)\rangle_\mathcal{H}$$

Proof. Assumption 2.4 implies that the feature map is Bochner integrable [136, Definition A.5.20] for the conditional distributions considered: $\forall d \in \mathcal{D}$, $\mathbb{E}_{X|D=d}\|\phi(d') \otimes \phi(X)\| < \infty$.

The first equality holds by Bochner integrability of the feature map, since it allows us to exchange the order of expectation and dot product.

$$\begin{aligned}
\mathbb{E}_{X|D=d}[\gamma_0(d', X)] &= \mathbb{E}_{X|D=d}\langle \gamma_0, \phi(d') \otimes \phi(X)\rangle_\mathcal{H} \\
&= \langle \gamma_0, \phi(d') \otimes \mathbb{E}_{X|D=d}[\phi(X)]\rangle_\mathcal{H} \\
&= \langle \gamma_0, \phi(d') \otimes E_1[\phi(X)](d)\rangle_\mathcal{H} \\
&= \langle \gamma_0, \phi(d') \otimes \mu_x(d)\rangle_\mathcal{H}
\end{aligned}$$

□

Proposition G.25. *Our RKHS construction implies that*

$$[E_1 h](\cdot) = \mathbb{E}_{X|D=(\cdot)}[h(X)] \in \mathcal{H}_\mathcal{D}, \quad \forall h \in \mathcal{H}_\mathcal{X}$$

Proof. After defining $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{D}$, we define the conditional expectation operator $E : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{D}$ such that $[Eh](d) = \mathbb{E}_{X|D=d}h(X)$. By construction, $\mathbb{E}_{X|D=(\cdot)}h(X) \in \mathcal{H}_\mathcal{D}$, $\forall h \in \mathcal{H}_\mathcal{X}$. This is precisely the condition required for the surrogate risk to coincide with the natural risk for the conditional expectation operator [60, 59]. □

H Algorithm derivation

H.1 Treatment effect

Proof of Theorem 2.2. By Theorem 2.1 and linearity of expectation

1. $\theta_0^{ATE}(d) = \langle \gamma_0, \phi(d) \otimes \mu_x\rangle_\mathcal{H}$ where $\mu_x := \int \phi(x)\mathbb{P}(x)$

$$\begin{aligned}
\theta_0^{ATE}(d) &= \int \gamma_0(d, x)\mathbb{P}(x) \\
&= \int \langle \gamma_0, \phi(d) \otimes \phi(x)\rangle_\mathcal{H}\mathbb{P}(x) \\
&= \langle \gamma_0, \phi(d) \otimes \int \phi(x)\mathbb{P}(x)\rangle_\mathcal{H} \\
&= \langle \gamma_0, \phi(d) \otimes \mu_x\rangle_\mathcal{H}
\end{aligned}$$

2. $\theta_0^{DS}(d, \mathbb{Q}) = \langle \gamma_0, \phi(d) \otimes \nu_x\rangle_\mathcal{H}$ where $\nu_x := \int \phi(x)\mathbb{Q}(x)$

$$\begin{aligned}
\theta_0^{DS}(d) &= \int \gamma_0(d, x)\mathbb{Q}(x) \\
&= \int \langle \gamma_0, \phi(d) \otimes \phi(x)\rangle_\mathcal{H}\mathbb{Q}(x) \\
&= \langle \gamma_0, \phi(d) \otimes \int \phi(x)\mathbb{Q}(x)\rangle_\mathcal{H} \\
&= \langle \gamma_0, \phi(d) \otimes \nu_x\rangle_\mathcal{H}
\end{aligned}$$

3. $\theta_0^{ATT}(d, d') = \langle \gamma_0, \phi(d') \otimes \mu_x(d) \rangle_{\mathcal{H}}$ where $\mu_x(d) := \int \phi(x)\mathbb{P}(x|d)$

$$\begin{aligned}
\theta_0^{ATT}(d, d') &= \int \gamma_0(d', x)\mathbb{P}(x|d) \\
&= \int \langle \gamma_0, \phi(d') \otimes \phi(x) \rangle_{\mathcal{H}} \mathbb{P}(x|d) \\
&= \langle \gamma_0, \phi(d') \otimes \int \phi(x)\mathbb{P}(x|d) \rangle_{\mathcal{H}} \\
&= \langle \gamma_0, \phi(d') \otimes \mu_x(d) \rangle_{\mathcal{H}}
\end{aligned}$$

4. $\theta_0^{CATE}(d, v) = \langle \gamma_0, \phi(d) \otimes \phi(v) \otimes \mu_x(v) \rangle_{\mathcal{H}}$ where $\mu_x(v) = \int \phi(x)\mathbb{P}(x|v)$

$$\begin{aligned}
\theta_0^{CATE}(d, v) &= \int \gamma_0(d, x)\mathbb{P}(x|v) \\
&= \int \langle \gamma_0, \phi(d) \otimes \phi(v) \otimes \phi(x) \rangle_{\mathcal{H}} \mathbb{P}(x|v) \\
&= \langle \gamma_0, \phi(d) \otimes \phi(v) \otimes \int \phi(x)\mathbb{P}(x|v) \rangle_{\mathcal{H}} \\
&= \langle \gamma_0, \phi(d) \otimes \phi(v) \otimes \mu_x(v) \rangle_{\mathcal{H}}
\end{aligned}$$

□

Derivation of Algorithm 2.1. By standard arguments

$$\hat{\gamma}(d, x) = \langle \hat{\gamma}, \phi(d) \otimes \phi(x) \rangle_{\mathcal{H}} = Y^T (K_{DD} \odot K_{XX} + n\lambda)^{-1} (K_{Dd} \odot K_{Xx})$$

1. $\hat{\theta}^{ATE}(d) = \frac{1}{n} \sum_{i=1}^n Y^T (K_{DD} \odot K_{XX} + n\lambda)^{-1} (K_{Dd} \odot K_{Xx_i})$
   Write the mean embedding
   $$\hat{\mu}_x = \frac{1}{n} \sum_{i=1}^n \phi(x_i)$$
   Finally appeal to
   $$\hat{\theta}^{ATE}(d) = \langle \hat{\gamma}, \phi(d) \otimes \hat{\mu}_x \rangle_{\mathcal{H}}$$

2. $\hat{\theta}^{DS}(d, \mathbb{Q}) = \frac{1}{n} \sum_{i=1}^n Y^T (K_{DD} \odot K_{XX} + n\lambda)^{-1} (K_{Dd} \odot K_{X\tilde{x}_i})$
   Write the mean embedding
   $$\hat{\nu}_x = \frac{1}{n} \sum_{i=1}^n \phi(\tilde{x}_i)$$
   Finally appeal to
   $$\hat{\theta}^{DS}(d, \mathbb{Q}) = \langle \hat{\gamma}, \phi(d) \otimes \hat{\nu}_x \rangle_{\mathcal{H}}$$

3. $\hat{\theta}^{ATT}(d, d') = Y^T (K_{DD} \odot K_{XX} + n\lambda)^{-1} (K_{Dd'} \odot [K_{XX}(K_{DD} + n\lambda_1)^{-1} K_{Dd}])$
   By [129, Algorithm 1], write the conditional mean embedding
   $$\hat{\mu}_x(d) = K_{\cdot X}(K_{DD} + n\lambda_1)^{-1} K_{Dd}$$
   Finally appeal to
   $$\hat{\theta}^{ATT}(d, d') = \langle \hat{\gamma}, \phi(d') \otimes \hat{\mu}_x(d) \rangle_{\mathcal{H}}$$

4. $\hat{\theta}^{CATE}(d,v) = Y^T(K_{DD} \odot K_{VV} \odot K_{XX} + n\lambda)^{-1}(K_{Dd} \odot K_{Vv} \odot K_{XX}(K_{VV} + n\lambda_2)^{-1}K_{Vv})$

   By [129, Algorithm 1], write the conditional mean embedding

$$\hat{\mu}_x(v) = K_{\cdot X}(K_{VV} + n\lambda_2)^{-1}K_{Vv}$$

   Finally appeal to

$$\hat{\theta}^{CATE}(d,v) = \langle \hat{\gamma}, \phi(d) \otimes \phi(v) \otimes \hat{\mu}_x(v) \rangle_{\mathcal{H}}$$

□

H.2 Mediation analysis

Proof of Theorem B.2.

$$\begin{aligned}
\theta_0^{ME}(d,d') &= \int \gamma_0(d',m,x)\mathbb{P}(m|d,x)\mathbb{P}(x) \\
&= \int \langle \gamma_0, \phi(d') \otimes \phi(m) \otimes \phi(x) \rangle_{\mathcal{H}} \mathbb{P}(m|d,x)\mathbb{P}(x) \\
&= \int \langle \gamma_0, \phi(d') \otimes \int \phi(m)\mathbb{P}(m|d,x) \otimes \phi(x) \rangle_{\mathcal{H}} \mathbb{P}(x) \\
&= \int \langle \gamma_0, \phi(d') \otimes \mu_m(d,x) \otimes \phi(x) \rangle_{\mathcal{H}} \mathbb{P}(x) \\
&= \langle \gamma_0, \phi(d') \otimes \int [\mu_m(d,x) \otimes \phi(x)]\mathbb{P}(x) \rangle_{\mathcal{H}}
\end{aligned}$$

□

Derivation of Algorithm B.1. By standard arguments

$$\begin{aligned}
\hat{\gamma}(d,m,x) &= \langle \hat{\gamma}, \phi(d) \otimes \phi(m) \otimes \phi(x) \rangle_{\mathcal{H}} \\
&= Y^T(K_{DD} \odot K_{MM} \odot K_{XX} + n\lambda)^{-1}(K_{Dd} \odot K_{Mm} \odot K_{Xx})
\end{aligned}$$

By [129, Algorithm 1], write the conditional mean

$$\hat{\mu}_m(d,x) = K_{\cdot M}(K_{DD} \odot K_{XX} + n\lambda_3)^{-1}(K_{Dd} \odot K_{Xx})$$

Therefore

$$\frac{1}{n}\sum_{i=1}^n [\hat{\mu}_m(d,x_i) \otimes \phi(x_i)] = \frac{1}{n}\sum_{i=1}^n [\{K_{\cdot M}(K_{DD} \odot K_{XX} + n\lambda_3)^{-1}(K_{Dd} \odot K_{Xx_i})\} \otimes \phi(x_i)]$$

and

$$\begin{aligned}
&\hat{\theta}^{ME}(d,d') \\
&= \langle \hat{\gamma}, \phi(d') \otimes \frac{1}{n}\sum_{i=1}^n [\hat{\mu}_m(d,x_i) \otimes \phi(x_i)] \rangle_{\mathcal{H}} \\
&= \frac{1}{n}\sum_{i=1}^n Y^T(K_{DD} \odot K_{MM} \odot K_{XX} + n\lambda)^{-1} \\
&\quad (K_{Dd'} \odot \{K_{MM}(K_{DD} \odot K_{XX} + n\lambda_3)^{-1}(K_{Dd} \odot K_{Xx_i})\} \odot K_{Xx_i})
\end{aligned}$$

□

H.3 Off-policy planning

Proof of Theorem C.2.

$$\begin{aligned}
\theta_0^{SATE}(d_1,d_2) &= \int \gamma_0(d_1,d_2,x_1,x_2)\mathbb{P}(x_2|d_1,x_1)\mathbb{P}(x_1) \\
&= \int \langle \gamma_0, \phi(d_1)\otimes\phi(d_2)\otimes\phi(x_1)\otimes\phi(x_2)\rangle_{\mathcal{H}}\mathbb{P}(x_2|d_1,x_1)\mathbb{P}(x_1) \\
&= \int \langle \gamma_0, \phi(d_1)\otimes\phi(d_2)\otimes\phi(x_1)\otimes \int \phi(x_2)\mathbb{P}(x_2|d_1,x_1)\rangle_{\mathcal{H}}\mathbb{P}(x_1) \\
&= \int \langle \gamma_0, \phi(d_1)\otimes\phi(d_2)\otimes\phi(x_1)\otimes \mu_{x_2}(d_1,x_1)\rangle_{\mathcal{H}}\mathbb{P}(x_1) \\
&= \langle \gamma_0, \phi(d_1)\otimes\phi(d_2)\otimes \int \phi(x_1)\otimes \mu_{x_2}(d_1,x_1)\mathbb{P}(x_1)\rangle_{\mathcal{H}}
\end{aligned}$$

The argument for θ_0^{SDS} is identical □

Derivation of Algorithm C.1. By standard arguments

$$\begin{aligned}
&\hat{\gamma}(d_1,d_2,x_1,x_2) \\
&= \langle \hat{\gamma}, \phi(d_1)\otimes\phi(d_2)\otimes\phi(x_1)\otimes\phi(x_2)\rangle_{\mathcal{H}} \\
&= Y^T(K_{D_1D_1}\odot K_{D_2D_2}\odot K_{X_1X_1}\odot K_{X_2X_2}+n\lambda)^{-1}(K_{D_1d_1}\odot K_{D_2d_2}\odot K_{X_1x_1}\odot K_{X_2x_2})
\end{aligned}$$

1. θ_0^{SATE}

   By [129, Algorithm 1], write the conditional mean

   $$\hat{\mu}_{x_2}(d_1,x_1) = K_{\cdot X_2}(K_{D_1D_1}\odot K_{X_1X_1}+n\lambda_4)^{-1}(K_{D_1d_1}\odot K_{X_1x_1})$$

   Therefore

   $$\begin{aligned}
   &\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i})\otimes\hat{\mu}_{x_2}(d_1,x_{1i})] \\
   &= \frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i})\otimes\{K_{\cdot X_2}(K_{D_1D_1}\odot K_{X_1X_1}+n\lambda_4)^{-1}(K_{D_1d_1}\odot K_{X_1x_{1i}})\}]
   \end{aligned}$$

   and

   $$\begin{aligned}
   &\hat{\theta}^{SATE}(d_1,d_2) \\
   &= \langle \hat{\gamma}, \phi(d_1)\otimes\phi(d_2)\otimes\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i})\otimes\hat{\mu}_{x_2}(d_1,x_{1i})]\rangle_{\mathcal{H}} \\
   &= \frac{1}{n}\sum_{i=1}^{n} Y^T(K_{D_1D_1}\odot K_{D_2D_2}\odot K_{X_1X_1}\odot K_{X_2X_2}+n\lambda)^{-1} \\
   &\quad (K_{D_1d_1}\odot K_{D_2d_2}\odot K_{X_1x_{1i}}\odot\{K_{X_2X_2}(K_{D_1D_1}\odot K_{X_1X_1}+n\lambda_4)^{-1}(K_{D_1d_1}\odot K_{X_1x_{1i}})\})
   \end{aligned}$$

2. θ_0^{SDS} By [129, Algorithm 1], write the conditional mean

   $$\hat{\nu}_{x_2}(d_1,x_1) = K_{\cdot\tilde{X}_2}(K_{\tilde{D}_1\tilde{D}_1}\odot K_{\tilde{X}_1\tilde{X}_1}+n\lambda_4)^{-1}(K_{\tilde{D}_1d_1}\odot K_{\tilde{X}_1x_1})$$

   Therefore

   $$\begin{aligned}
   &\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}[\phi(\tilde{x}_{1i})\otimes\hat{\mu}_{x_2}(d_1,\tilde{x}_{1i})] \\
   &= \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}[\phi(\tilde{x}_{1i})\otimes\{K_{\cdot\tilde{X}_2}(K_{\tilde{D}_1\tilde{D}_1}\odot K_{\tilde{X}_1\tilde{X}_1}+n\lambda_4)^{-1}(K_{\tilde{D}_1d_1}\odot K_{\tilde{X}_1\tilde{x}_{1i}})\}]
   \end{aligned}$$

and

$$\hat{\theta}^{SDS}(d_1, d_2)$$

$$= \langle \hat{\gamma}, \phi(d_1) \otimes \phi(d_2) \otimes \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} [\phi(\tilde{x}_{1i}) \otimes \hat{\mu}_{x_2}(d_1, \tilde{x}_{1i})] \rangle_{\mathcal{H}}$$

$$= \frac{1}{n} \sum_{i=1}^{n} Y^T (K_{D_1 D_1} \odot K_{D_2 D_2} \odot K_{X_1 X_1} \odot K_{X_2 X_2} + n\lambda)^{-1}$$
$$(K_{D_1 d_1} \odot K_{D_2 d_2} \odot K_{X_1 \tilde{x}_{1i}} \odot \{K_{X_2 \tilde{X}_2} (K_{\tilde{D}_1 \tilde{D}_1} \odot K_{\tilde{X}_1 \tilde{X}_1} + n\lambda_4)^{-1} (K_{\tilde{D}_1 d_1} \odot K_{\tilde{X}_1 \tilde{x}_{1i}})\})$$

□

H.4 Graphical effect

Proof of Theorem D.2.

$$\begin{aligned}
\check{\theta}_0^{ATE}(d) &= \int \gamma_0(d', x) \mathbb{P}(d') \mathbb{P}(x|d) \\
&= \int \langle \gamma_0, \phi(d') \otimes \phi(x) \rangle_{\mathcal{H}} \mathbb{P}(d') \mathbb{P}(x|d) \\
&= \langle \gamma_0, \int \phi(d') \mathbb{P}(d') \otimes \int \phi(x) \mathbb{P}(x|d) \rangle_{\mathcal{H}} \\
&= \langle \gamma_0, \mu_d \otimes \mu_x(d) \rangle_{\mathcal{H}}
\end{aligned}$$

□

Derivation of Algorithm D.1. By standard arguments

$$\hat{\gamma}(d, x) = \langle \hat{\gamma}, \phi(d) \otimes \phi(x) \rangle_{\mathcal{H}} = Y^T (K_{DD} \odot K_{XX} + n\lambda)^{-1} (K_{Dd} \odot K_{Xx})$$

Write the mean embedding

$$\hat{\mu}_x = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$$

By [129, Algorithm 1], write the conditional mean embedding

$$\hat{\mu}_x(d) = K_{\cdot X} (K_{DD} + n\lambda_1)^{-1} K_{Dd}$$

Finally appeal to

$$\hat{\theta}^{FD}(d) = \langle \hat{\gamma}, \hat{\mu}_d \otimes \hat{\mu}_x(d) \rangle_{\mathcal{H}}$$

Therefore

$$\hat{\theta}^{FD}(d) = \frac{1}{n} \sum_{i=1}^{n} Y^T (K_{DD} \odot K_{XX} + n\lambda)^{-1} (K_{Dd_i} \odot \{K_{XX} (K_{DD} + n\lambda_1)^{-1} K_{Dd}\})$$

□

H.5 Distribution effect

Proof of Theorem E.2. By Theorem 2.1 and linearity of expectation

1. $\check{\theta}_0^{DATE}(d)$

$$\begin{aligned}
\check{\theta}_0^{DATE}(d) &= \int \gamma_0(d, x) \mathbb{P}(x) \\
&= \int E_6^*[\phi(d) \otimes \phi(x)] \mathbb{P}(x) \\
&= E_6^*[\phi(d) \otimes \int \phi(x) \mathbb{P}(x)] \\
&= E_6^*[\phi(d) \otimes \mu_x]
\end{aligned}$$

2. $\check{\theta}_0^{DATT}(d, d')$

$$\begin{aligned}\check{\theta}_0^{DATE}(d) &= \int \gamma_0(d', x)\mathbb{P}(x|d) \\ &= \int E_6^*[\phi(d') \otimes \phi(x)]\mathbb{P}(x|d) \\ &= E_6^*[\phi(d') \otimes \int \phi(x)\mathbb{P}(x|d)] \\ &= E_6^*[\phi(d') \otimes \mu_x(d)]\end{aligned}$$

□

Derivation of Algorithm E.1. By [129, Algorithm 1]

$$\hat{\gamma}(d, x) = \hat{E}_6^*[\phi(d) \otimes \phi(x)] = K_{\cdot Y}(K_{DD} \odot K_{XX} + n\lambda_6)^{-1}(K_{Dd} \odot K_{Xx})$$

1. $\hat{\theta}^{DATE}(d)$

   Write the mean embedding as

   $$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$$

   Finally appeal to

   $$\hat{\theta}^{DATE}(d) = \hat{E}_6^*[\phi(d) \otimes \hat{\mu}_x]$$

2. $\hat{\theta}^{DATT}(d)$

   By [129, Algorithm 1], write the conditional mean embedding

   $$\hat{\mu}_x(d) = K_{\cdot X}(K_{DD} + n\lambda_1)^{-1}K_{Dd}$$

   Finally appeal to

   $$\hat{\theta}^{DATE}(d) = \hat{E}_6^*[\phi(d) \otimes \hat{\mu}_x(d)]$$

□

I Consistency proof

I.1 Probability

Proposition I.1 (Lemma 2 of [130])**.** *Let ξ be a random variable taking values in a real separable Hilbert space $\mathcal{K}$. Suppose $\exists \tilde{M}$ s.t.*

$$\begin{aligned}&\|\xi\|_{\mathcal{K}} \leq \tilde{M} < \infty \quad \textit{a.s.} \\ &\sigma^2(\xi) := \mathbb{E}\|\xi\|_{\mathcal{K}}^2\end{aligned}$$

Then $\forall n \in \mathbb{N}, \forall \eta \in (0, 1)$,

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbb{E}\xi\right\|_{\mathcal{K}} \leq \frac{2\tilde{M}\ln(2/\eta)}{n} + \sqrt{\frac{2\sigma^2(\xi)\ln(2/\eta)}{n}}\right] \geq 1 - \eta$$

I.2 Treatment effect

I.2.1 Regression

For expositional purposes, we present classic results for the kernel ridge regression estimator $\hat{\gamma}$ for $\gamma_0(d, x) := \mathbb{E}[Y|D = d, X = x]$. The same results hold for $\gamma_0(d, v, x) := \mathbb{E}[Y|D = d, V =$

$v, X = x]$.

$$\gamma_0 = \operatorname*{argmin}_{\gamma \in \mathcal{H}} \mathcal{E}(\gamma), \quad \mathcal{E}(\gamma) = \mathbb{E}[Y - \gamma(D, X)]^2$$
$$\gamma_\lambda = \operatorname*{argmin}_{\gamma \in \mathcal{H}} \mathcal{E}_\lambda(\gamma), \quad \mathcal{E}_\lambda(\gamma) = \mathcal{E}(\gamma) + \lambda \|\gamma\|_{\mathcal{H}}^2$$
$$\hat{\gamma} = \operatorname*{argmin}_{\gamma \in \mathcal{H}} \hat{\mathcal{E}}(\gamma), \quad \hat{\mathcal{E}}(\gamma) = \frac{1}{n} \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \|\gamma\|_{\mathcal{H}}^2$$

Proposition I.2.

$$\|T^{\frac{1}{2}}(\gamma - \gamma_0)\|_{\mathcal{H}}^2 = \|\gamma - \gamma_0\|_{L^2}^2 = \mathbb{E}[f(X) - \gamma_0(X)]^2 = \mathcal{E}(f) - \mathcal{E}(\gamma_0)$$

Proof.

$$\mathcal{E}(\gamma) = \mathbb{E}[Y - \gamma(D, X)]^2 = \mathbb{E}[Y - \gamma_0(D, X) + \gamma_0(D, X) - \gamma(D, X)]^2$$

Expanding the square we see that the cross terms are 0 by law of iterated expectation. For the first equality, see [16, eq. 24] □

Proposition I.3.

$$\gamma_\lambda = \operatorname*{argmin}_{\gamma \in \mathcal{H}} \mathbb{E}[\gamma(D, X) - \gamma_0(D, X)]^2 + \lambda \|\gamma\|_{\mathcal{H}}^2$$

Proof. Corollary of Proposition I.2. □

Proposition I.4.

$$\gamma_\lambda = [(T + \lambda)^{-1} \circ T]\gamma_0$$

Proof. [16, eq. 25] □

Proposition I.5 (Sampling error)**.** *Suppose Assumptions 2.3 and 2.4 hold. Then*

1. *for* $\gamma_0(d, x)$*, w.p.* $1 - \delta$

$$\|\hat{\gamma} - \gamma_\lambda\|_{\mathcal{H}} \leq \frac{6C\kappa_d\kappa_x \ln(2/\delta)}{\sqrt{n}\lambda}$$

2. *for* $\gamma_0(d, v, x)$*, w.p.* $1 - \delta$

$$\|\hat{\gamma} - \gamma_\lambda\|_{\mathcal{H}} \leq \frac{6C\kappa_d\kappa_v\kappa_x \ln(2/\delta)}{\sqrt{n}\lambda}$$

Proof. [130, Theorem 1], and using $\|\phi(d) \otimes \phi(x)\|_{\mathcal{H}} \leq \kappa_d\kappa_x$ and $\|\phi(d) \otimes \phi(v) \otimes \phi(x)\|_{\mathcal{H}} \leq \kappa_d\kappa_v\kappa_x$. □

Proposition I.6 (Approximation error)**.** *Suppose Assumptions 2.3 and 2.5 hold. Then*

$$\|\gamma_\lambda - \gamma_0\|_{\mathcal{H}} \leq \lambda^{\frac{c-1}{2}} \sqrt{\zeta}$$

Proof. By Hypothesis 2.5, $\exists\ g$ s.t.

$$g = T^{\frac{1-c}{2}}\gamma_0 = \sum_k \nu_k^{\frac{1-c}{2}} e_k \langle e_k, \gamma_0 \rangle_{\mathcal{H}} = \sum_k \nu_k^{\frac{1-c}{2}} [e_k \otimes e_k]\gamma_0$$

Hence

$$\zeta = \|g\|_{\mathcal{H}}^2 = \sum_k \nu_k^{1-c} \|\gamma_0\|_{\mathcal{H}}^2$$

By Proposition I.4, write

$$\begin{aligned}\gamma_\lambda - \gamma_0 &= [(T+\lambda I)^{-1} \circ T - I]\gamma_0 \\ &= \sum_k \left(\frac{\nu_k}{\nu_k+\lambda} - 1\right) e_k \langle e_k, \gamma_0\rangle_{\mathcal{H}_\mathcal{Z}} \\ &= \sum_k \left(\frac{\nu_k}{\nu_k+\lambda} - 1\right) [e_k \otimes e_k]\gamma_0\end{aligned}$$

Hence

$$\begin{aligned}\|\gamma_\lambda - \gamma_0\|_{\mathcal{H}}^2 &= \sum_k \left(\frac{\nu_k}{\nu_k+\lambda} - 1\right)^2 \|\gamma_0\|_{\mathcal{H}}^2 \\ &= \sum_k \left(\frac{\lambda}{\nu_k+\lambda}\right)^2 \|\gamma_0\|_{\mathcal{H}}^2 \\ &= \sum_k \left(\frac{\lambda}{\nu_k+\lambda}\right)^2 \|\gamma_0\|_{\mathcal{H}}^2 \left(\frac{\lambda}{\lambda}\cdot\frac{\nu_k}{\nu_k}\cdot\frac{\nu_k+\lambda}{\nu_k+\lambda}\right)^{c-1} \\ &= \lambda^{c-1}\sum_k \nu_k^{1-c}\|\gamma_0\|_{\mathcal{H}}^2 \left(\frac{\lambda}{\nu_k+\lambda}\right)^{3-c}\left(\frac{\nu_k}{\nu_k+\lambda}\right)^{c-1} \\ &\le \lambda^{c-1}\sum_k \nu_k^{1-c}\|\gamma_0\|_{\mathcal{H}}^2 \\ &= \lambda^{c-1}\|g\|_{\mathcal{H}}^2 \\ &\le \lambda^{c-1}\zeta\end{aligned}$$

□

Theorem I.1 (Regression rate)**.** *Suppose Assumptions 2.3, 2.4, and 2.5 hold. Then*

1. *for $\gamma_0(d,x)$, w.p. $1-\delta$*

$$\|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \le r_\gamma(n,\delta,c) := \frac{\sqrt{\zeta}(c+1)}{4^{\frac{1}{c+1}}}\left(\frac{6C\kappa_d\kappa_x \ln(2/\delta)}{\sqrt{n\zeta}(c-1)}\right)^{\frac{c-1}{c+1}}$$

2. *for $\gamma_0(d,v,x)$, w.p. $1-\delta$*

$$\|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \le r_\gamma(n,\delta,c) := \frac{\sqrt{\zeta}(c+1)}{4^{\frac{1}{c+1}}}\left(\frac{6C\kappa_d\kappa_v\kappa_x \ln(2/\delta)}{\sqrt{n\zeta}(c-1)}\right)^{\frac{c-1}{c+1}}$$

Proof. We focus on case of $\gamma_0(d,x)$. By triangle inequality,

$$\|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \le \|\hat{\gamma} - \gamma_\lambda\|_{\mathcal{H}} + \|\gamma_\lambda - \gamma_0\|_{\mathcal{H}} \le \frac{6C\kappa_d\kappa_x \ln(2/\delta)}{\sqrt{n}\lambda} + \lambda^{\frac{c-1}{2}}\sqrt{\zeta}$$

Minimize the RHS w.r.t. λ. Rewrite the objective as

$$A\lambda^{-1} + B\lambda^{\frac{c-1}{2}}$$

then the FOC yields

$$\lambda = \left(\frac{2A}{B(c-1)}\right)^{\frac{2}{c+1}} = \left(\frac{12C\kappa_d\kappa_x \ln(2/\delta)}{\sqrt{n\zeta}(c-1)}\right)^{\frac{2}{c+1}} = O(n^{\frac{-1}{c+1}})$$

Substituting this value of λ, the RHS becomes

$$\begin{aligned}
&A\left(\frac{2A}{B(c-1)}\right)^{-\frac{2}{c+1}} + B\left(\frac{2A}{B(c-1)}\right)^{\frac{c-1}{c+1}} \\
&= \frac{B(c+1)}{4^{\frac{1}{c+1}}}\left(\frac{A}{B(c-1)}\right)^{\frac{c-1}{c+1}} \\
&= \frac{\sqrt{\zeta}(c+1)}{4^{\frac{1}{c+1}}}\left(\frac{6C\kappa_d\kappa_x \ln(2/\delta)}{\sqrt{n\zeta}(c-1)}\right)^{\frac{c-1}{c+1}}
\end{aligned}$$

□

I.2.2 Unconditional mean embedding

Theorem I.2 (Mean embedding rate)**.** *Suppose Assumptions 2.3 and 2.4 hold. Then w.p.* $1-\delta$*,*

$$\|\hat{\mu}_x - \mu_x\|_{\mathcal{H}_\mathcal{X}} \leq r_\mu(n,\delta) := \frac{4\kappa_x \ln(2/\delta)}{\sqrt{n}}$$

Likewise, w.p. $1-\delta$

$$\|\hat{\nu}_x - \nu_x\|_{\mathcal{H}_\mathcal{X}} \leq r_\nu(\tilde{n},\delta) := \frac{4\kappa_x \ln(2/\delta)}{\sqrt{\tilde{n}}}$$

Proof. By Lemma I.1 with $\xi_i = \phi(x_i)$, since

$$\left\|\frac{1}{n}\sum_{i=1}^n \phi(x_i) - \mathbb{E}[\phi(X)]\right\|_{\mathcal{H}_\mathcal{X}} \leq \frac{2\kappa_x \ln(2/\delta)}{n} + \sqrt{\frac{2\kappa_x^2 \ln(2/\delta)}{n}} \leq \frac{4\kappa_x \ln(2/\delta)}{\sqrt{n}}$$

The argument for ν_x is identical, using $\xi_i = \phi(\tilde{x}_i)$ □

I.2.3 Conditional mean embeddings

Theorem I.3 (Conditional mean embedding rate)**.** *Suppose Assumptions 2.3 and 2.4 hold.*

1. *If in addition Assumption 2.6 holds then w.p.* $1-\delta$*,* $\forall d \in \mathcal{D}$

$$\|\hat{\mu}_x(d) - \mu_x(d)\|_{\mathcal{H}_\mathcal{X}} \leq r_\mu^{ATT}(\delta, n, c_1)$$

where

$$r_\mu^{ATT}(\delta,n,c_1) := \kappa_d \cdot \frac{\sqrt{\zeta_1}(c_1+1)}{4^{\frac{1}{c_1+1}}}\left(\frac{4\kappa_d(\kappa_x + \kappa_d\|E_1\|_{\mathcal{L}_2})\ln(2/\delta)}{\sqrt{n\zeta_1}(c_1-1)}\right)^{\frac{c_1-1}{c_1+1}}$$

2. *If in addition Assumption 2.7 holds then w.p.* $1-\delta$*,* $\forall v \in \mathcal{V}$

$$\|\hat{\mu}_x(v) - \mu_x(v)\|_{\mathcal{H}_\mathcal{X}} \leq r_\mu^{CATE}(\delta, n, c_2)$$

where

$$r_\mu^{CATE}(\delta,n,c_2) := \kappa_v \cdot \frac{\sqrt{\zeta_2}(c_2+1)}{4^{\frac{1}{c_2+1}}}\left(\frac{4\kappa_v(\kappa_x + \kappa_v\|E_2\|_{\mathcal{L}_2})\ln(2/\delta)}{\sqrt{n\zeta_2}(c_2-1)}\right)^{\frac{c_2-1}{c_2+1}}$$

Proof. Immediate from [129, Corollary 1], observing that

$$E_1 : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{D}, \quad \|\phi(x)\|_{\mathcal{H}_\mathcal{X}} \leq \kappa_x, \quad \|\phi(d)\|_{\mathcal{H}_\mathcal{D}} \leq \kappa_d$$

and

$$E_2 : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{V}, \quad \|\phi(x)\|_{\mathcal{H}_\mathcal{X}} \leq \kappa_x, \quad \|\phi(v)\|_{\mathcal{H}_\mathcal{V}} \leq \kappa_v$$

□

I.2.4 Target parameters

Proposition I.7. *In summary*

$$\begin{aligned}
\|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} &= O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}}\right) \\
\|\hat{\mu}_x - \mu_x\|_{\mathcal{H}_{\mathcal{X}}} &= O_p\left(n^{-\frac{1}{2}}\right) \\
\|\hat{\nu}_x - \nu_x\|_{\mathcal{H}_{\mathcal{X}}} &= O_p\left(\tilde{n}^{-\frac{1}{2}}\right) \\
\|\hat{\mu}_x(d) - \mu_x(d)\|_{\mathcal{H}_{\mathcal{X}}} &= O_p\left(n^{-\frac{1}{2}\frac{c_1-1}{c_1+1}}\right) \\
\|\hat{\mu}_x(v) - \mu_x(v)\|_{\mathcal{H}_{\mathcal{X}}} &= O_p\left(n^{-\frac{1}{2}\frac{c_2-1}{c_2+1}}\right)
\end{aligned}$$

Proof of Theorem 2.3. We consider each treatment effect parameter

1. θ_0^{ATE}

$$\begin{aligned}
&\hat{\theta}^{ATE}(d) - \theta_0^{ATE}(d) \\
&= \langle \hat{\gamma}, \phi(d) \otimes \hat{\mu}_x \rangle_{\mathcal{H}} - \langle \gamma_0, \phi(d) \otimes \mu_x \rangle_{\mathcal{H}} \\
&= \langle \hat{\gamma}, \phi(d) \otimes [\hat{\mu}_x - \mu_x] \rangle_{\mathcal{H}} + \langle [\hat{\gamma} - \gamma_0], \phi(d) \otimes \mu_x \rangle_{\mathcal{H}} \\
&= \langle [\hat{\gamma} - \gamma_0], \phi(d) \otimes [\hat{\mu}_x - \mu_x] \rangle_{\mathcal{H}} + \langle \gamma_0, \phi(d) \otimes [\hat{\mu}_x - \mu_x] \rangle_{\mathcal{H}} + \langle [\hat{\gamma} - \gamma_0], \phi(d) \otimes \mu_x \rangle_{\mathcal{H}}
\end{aligned}$$

Therefore w.p. $1 - 2\delta$

$$\begin{aligned}
&|\hat{\theta}^{ATE}(d) - \theta_0^{ATE}(d)| \\
&\leq \|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \|\hat{\mu}_x - \mu_x\|_{\mathcal{H}_{\mathcal{X}}} + \|\gamma_0\|_{\mathcal{H}} \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \|\hat{\mu}_x - \mu_x\|_{\mathcal{H}_{\mathcal{X}}} \\
&\quad + \|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \|\mu_x\|_{\mathcal{H}_{\mathcal{X}}} \\
&\leq \kappa_d \cdot r_\gamma(n, \delta, c) \cdot r_\mu(n, \delta) + \kappa_d \cdot \|\gamma_0\|_{\mathcal{H}} \cdot r_\mu(n, \delta) + \kappa_d \kappa_x \cdot r_\gamma(n, \delta, c) \\
&= O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}}\right)
\end{aligned}$$

2. θ_0^{DS}

By the same argument as for θ_0^{ATE}, w.p. $1 - 2\delta$

$$\begin{aligned}
&|\hat{\theta}^{DS}(d, \tilde{P}) - \theta_0^{DS}(d, \tilde{P})| \\
&\leq \kappa_d \cdot r_\gamma(n, \delta, c) \cdot r_\nu(\tilde{n}, \delta) + \kappa_d \cdot \|\gamma_0\|_{\mathcal{H}} \cdot r_\nu(\tilde{n}, \delta) + \kappa_d \kappa_x \cdot r_\gamma(n, \delta, c) \\
&= O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + \tilde{n}^{-\frac{1}{2}}\right)
\end{aligned}$$

3. θ_0^{ATT}

$$\begin{aligned}
&\hat{\theta}^{ATT}(d, d') - \theta_0^{ATT}(d, d') \\
&= \langle \hat{\gamma}, \phi(d') \otimes \hat{\mu}_x(d) \rangle_{\mathcal{H}} - \langle \gamma_0, \phi(d') \otimes \mu_x(d) \rangle_{\mathcal{H}} \\
&= \langle \hat{\gamma}, \phi(d') \otimes [\hat{\mu}_x(d) - \mu_x(d)] \rangle_{\mathcal{H}} + \langle [\hat{\gamma} - \gamma_0], \phi(d') \otimes \mu_x(d) \rangle_{\mathcal{H}} \\
&= \langle [\hat{\gamma} - \gamma_0], \phi(d') \otimes [\hat{\mu}_x(d) - \mu_x(d)] \rangle_{\mathcal{H}} + \langle \gamma_0, \phi(d') \otimes [\hat{\mu}_x(d) - \mu_x(d)] \rangle_{\mathcal{H}} \\
&\quad + \langle [\hat{\gamma} - \gamma_0], \phi(d') \otimes \mu_x(d) \rangle_{\mathcal{H}}
\end{aligned}$$

Therefore w.p. $1 - 2\delta$

$$\begin{aligned}
&|\hat{\theta}^{ATT}(d, d') - \theta_0^{ATT}(d, d')| \\
&\leq \|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \|\phi(d')\|_{\mathcal{H}_{\mathcal{D}}} \|\hat{\mu}_x(d) - \mu_x(d)\|_{\mathcal{H}_{\mathcal{X}}} + \|\gamma_0\|_{\mathcal{H}} \|\phi(d')\|_{\mathcal{H}_{\mathcal{D}}} \|\hat{\mu}_x(d) - \mu_x(d)\|_{\mathcal{H}_{\mathcal{X}}} \\
&\quad + \|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \|\phi(d')\|_{\mathcal{H}_{\mathcal{D}}} \|\mu_x(d)\|_{\mathcal{H}_{\mathcal{X}}} \\
&\leq \kappa_d \cdot r_\gamma(n, \delta, c) \cdot r_\mu^{ATT}(n, \delta, c_1) + \kappa_d \cdot \|\gamma_0\|_{\mathcal{H}} \cdot r_\mu^{ATT}(n, \delta, c_1) + \kappa_d \kappa_x \cdot r_\gamma(n, \delta, c) \\
&= O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + n^{-\frac{1}{2}\frac{c_1-1}{c_1+1}}\right)
\end{aligned}$$

4. θ_0^{CATE}

$$
\begin{aligned}
&\hat{\theta}^{CATE}(d,v) - \theta_0^{CATE}(d,v) \\
&= \langle \hat{\gamma}, \phi(d) \otimes \phi(v) \otimes \hat{\mu}_x(v) \rangle_{\mathcal{H}} - \langle \gamma_0, \phi(d) \otimes \phi(v) \otimes \mu_x(v) \rangle_{\mathcal{H}} \\
&= \langle \hat{\gamma}, \phi(d) \otimes \phi(v) \otimes [\hat{\mu}_x(v) - \mu_x(v)] \rangle_{\mathcal{H}} + \langle [\hat{\gamma} - \gamma_0], \phi(d) \otimes \phi(v) \otimes \mu_x(v) \rangle_{\mathcal{H}} \\
&= \langle [\hat{\gamma} - \gamma_0], \phi(d) \otimes \phi(v) \otimes [\hat{\mu}_x(v) - \mu_x(v)] \rangle_{\mathcal{H}} \\
&\quad + \langle \gamma_0, \phi(d) \otimes \phi(v) \otimes [\hat{\mu}_x(v) - \mu_x(v)] \rangle_{\mathcal{H}} \\
&\quad + \langle [\hat{\gamma} - \gamma_0], \phi(d) \otimes \phi(v) \otimes \mu_x(v) \rangle_{\mathcal{H}}
\end{aligned}
$$

Therefore w.p. $1 - 2\delta$

$$
\begin{aligned}
&|\hat{\theta}^{CATE}(d,v) - \theta_0^{CATE}(d,v)| \\
&\leq \|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \|\phi(v)\|_{\mathcal{H}_{\mathcal{V}}} \|\hat{\mu}_x(v) - \mu_x(v)\|_{\mathcal{H}_{\mathcal{X}}} \\
&\quad + \|\gamma_0\|_{\mathcal{H}} \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \|\phi(v)\|_{\mathcal{H}_{\mathcal{V}}} \|\hat{\mu}_x(v) - \mu_x(v)\|_{\mathcal{H}_{\mathcal{X}}} \\
&\quad + \|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \|\phi(v)\|_{\mathcal{H}_{\mathcal{V}}} \|\mu_x(v)\|_{\mathcal{H}_{\mathcal{X}}} \\
&\leq \kappa_d \kappa_v \cdot r_\gamma(n, \delta, c) \cdot r_\mu^{CATE}(n, \delta, c_2) + \kappa_d \kappa_v \cdot \|\gamma_0\|_{\mathcal{H}} \cdot r_\mu^{CATE}(n, \delta, c_2) \\
&\quad + \kappa_d \kappa_v \kappa_x \cdot r_\gamma(n, \delta, c) \\
&= O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + n^{-\frac{1}{2}\frac{c_2-1}{c_2+1}} \right)
\end{aligned}
$$

□

I.3 Mediation analysis

I.3.1 Regression

We state a result analogous to Theorem I.1 for $\gamma_0(d, m, x)$.

Theorem I.4 (Regression rate)**.** *Suppose Assumptions B.2, B.3, and B.4 hold. Then w.p.* $1 - \delta$

$$
\|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \leq r_\gamma(n, \delta, c) := \frac{\sqrt{\zeta}(c+1)}{4^{\frac{1}{c+1}}} \left(\frac{6C\kappa_d \kappa_m \kappa_x \ln(2/\delta)}{\sqrt{n\zeta}(c-1)} \right)^{\frac{c-1}{c+1}}
$$

Proof. Identical to Theorem I.1 □

I.3.2 Unconditional mean embedding

Theorem I.5 (Mean embedding rate)**.** *Suppose Assumptions B.2 and B.3 hold. Then w.p.* $1 - \delta$*,* $\forall d \in \mathcal{D}$

$$
\left\| \frac{1}{n} \sum_{i=1}^{n} [\mu_m(d, x_i) \otimes \phi(x_i)] - \int [\mu_m(d, X) \otimes \phi(X)] \mathbb{P}(x) \right\|_{\mathcal{H}_{\mathcal{M}} \otimes \mathcal{H}_{\mathcal{X}}} \leq r_\mu^{ME}(n, \delta)
$$

where

$$
r_\mu^{ME}(n, \delta) := \frac{4\kappa_m \kappa_x \ln(2/\delta)}{\sqrt{n}}
$$

Proof. By Lemma I.1 with $\xi_i = [\mu_m(d, x_i) \otimes \phi(x_i)]$, since

$$
\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^{n} [\mu_m(d, x_i) \otimes \phi(x_i)] - \mathbb{E}[\mu_m(d, X) \otimes \phi(X)] \right\|_{\mathcal{H}_{\mathcal{M}} \otimes \mathcal{H}_{\mathcal{X}}} \\
&\leq \frac{2\kappa_m \kappa_x \ln(2/\delta)}{n} + \sqrt{\frac{2\kappa_m^2 \kappa_x^2 \ln(2/\delta)}{n}} \\
&\leq \frac{4\kappa_m \kappa_x \ln(2/\delta)}{\sqrt{n}}
\end{aligned}
$$

□

I.3.3 Conditional mean embedding

Theorem I.6 (Conditional mean embedding rate)**.** *Suppose Assumptions B.2, B.3, and B.5 hold. Then w.p.* $1-\delta$*,* $\forall d \in \mathcal{D}$ *and* $\forall x \in \mathcal{X}$

$$\|\hat{\mu}_m(d,x) - \mu_m(d,x)\|_{\mathcal{H}_\mathcal{M}} \leq r_\mu^{ME}(\delta, n, c_3)$$

where

$$r_\mu^{ME}(\delta, n, c_3) := \kappa_d \kappa_x \cdot \frac{\sqrt{\zeta_3}(c_3+1)}{4^{\frac{1}{c_3+1}}} \left(\frac{4\kappa_d\kappa_x(\kappa_m + \kappa_d\kappa_x\|E_3\|_{\mathcal{L}_2})\ln(2/\delta)}{\sqrt{n\zeta_3}(c_3-1)} \right)^{\frac{c_3-1}{c_3+1}}$$

Proof. Immediate from [129, Corollary 1], observing that

$$E_3 : \mathcal{H}_\mathcal{M} \to \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X}, \quad \|\phi(m)\|_{\mathcal{H}_\mathcal{M}} \leq \kappa_m, \quad \|\phi(d)\|_{\mathcal{H}_\mathcal{D}} \leq \kappa_d, \quad \|\phi(x)\|_{\mathcal{H}_\mathcal{X}} \leq \kappa_x$$

□

I.3.4 Target parameter

Proposition I.8. *In summary*

$$\begin{aligned}
&\|\hat{\gamma} - \gamma_0\|_\mathcal{H} = O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}}\right) \\
&\left\| \frac{1}{n}\sum_{i=1}^n [\mu_m(d,x_i)\otimes\phi(x_i)] - \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x) \right\|_{\mathcal{H}_\mathcal{M}\otimes\mathcal{H}_\mathcal{X}} = O_p\left(n^{-\frac{1}{2}}\right) \\
&\|\hat{\mu}_m(d,x) - \mu_m(d,x)\|_{\mathcal{H}_\mathcal{M}} = O_p\left(n^{-\frac{1}{2}\frac{c_3-1}{c_3+1}}\right)
\end{aligned}$$

Proposition I.9. *Suppose Assumptions B.2, B.3, and B.5 hold. Then w.p.* $1-2\delta$

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^n [\hat{\mu}_m(d,x_i)\otimes\phi(x_i)] - \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x) \right\|_{\mathcal{H}_\mathcal{M}\otimes\mathcal{H}_\mathcal{X}} \\
&\leq \kappa_x \cdot r_\mu^{ME}(n,\delta,c_3) + r_\mu^{ME}(n,\delta)
\end{aligned}$$

Proof. By triangle inequality, it is sufficient to control

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^n [\hat{\mu}_m(d,x_i)\otimes\phi(x_i)] - [\mu_m(d,x_i)\otimes\phi(x_i)] \right\|_{\mathcal{H}_\mathcal{M}\otimes\mathcal{H}_\mathcal{X}} \\
&= \left\| \frac{1}{n}\sum_{i=1}^n [\hat{\mu}_m(d,x_i) - \mu_m(d,x_i)]\otimes\phi(x_i) \right\|_{\mathcal{H}_\mathcal{M}\otimes\mathcal{H}_\mathcal{X}} \\
&\leq \kappa_x \max_i \|\hat{\mu}_m(d,x_i) - \mu_m(d,x_i)\|_{\mathcal{H}_\mathcal{M}} \\
&\leq \kappa_x \cdot r_\mu^{ME}(n,\delta,c_3)
\end{aligned}$$

and

$$\left\| \frac{1}{n}\sum_{i=1}^n [\mu_m(d,x_i)\otimes\phi(x_i)] - \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x) \right\|_{\mathcal{H}_\mathcal{M}\otimes\mathcal{H}_\mathcal{X}} \leq r_\mu^{ME}(n,\delta)$$

□

Proof of Theorem B.3.

$$
\begin{aligned}
&\hat{\theta}^{ME}(d,d') - \theta_0^{ME}(d,d') \\
&= \langle \hat{\gamma}, \phi(d') \otimes \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_m(d,x_i)\otimes\phi(x_i)]\rangle_{\mathcal{H}} - \langle \gamma_0, \phi(d') \otimes \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x)\rangle_{\mathcal{H}} \\
&= \langle \hat{\gamma}, \phi(d') \otimes \left\{ \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_m(d,x_i)\otimes\phi(x_i)] - \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x) \right\} \rangle_{\mathcal{H}} \\
&\quad + \langle [\hat{\gamma}-\gamma_0], \phi(d') \otimes \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x)\rangle_{\mathcal{H}} \\
&= \langle [\hat{\gamma}-\gamma_0], \phi(d') \otimes \left\{ \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_m(d,x_i)\otimes\phi(x_i)] - \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x) \right\} \rangle_{\mathcal{H}} \\
&\quad + \langle \gamma_0, \phi(d') \otimes \left\{ \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_m(d,x_i)\otimes\phi(x_i)] - \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x) \right\} \rangle_{\mathcal{H}} \\
&\quad + \langle [\hat{\gamma}-\gamma_0], \phi(d') \otimes \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x)\rangle_{\mathcal{H}}
\end{aligned}
$$

Therefore w.p. $1-3\delta$

$$
\begin{aligned}
&|\hat{\theta}^{ME}(d,d') - \theta_0^{ME}(d,d')| \\
&\leq \|\hat{\gamma}-\gamma_0\|_{\mathcal{H}}\|\phi(d')\|_{\mathcal{H}_{\mathcal{D}}} \left\| \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_m(d,x_i)\otimes\phi(x_i)] - \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x) \right\|_{\mathcal{H}_{\mathcal{M}}\otimes\mathcal{H}_{\mathcal{X}}} \\
&\quad + \|\gamma_0\|_{\mathcal{H}}\|\phi(d')\|_{\mathcal{H}_{\mathcal{D}}} \left\| \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_m(d,x_i)\otimes\phi(x_i)] - \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x) \right\|_{\mathcal{H}_{\mathcal{M}}\otimes\mathcal{H}_{\mathcal{X}}} \\
&\quad + \|\hat{\gamma}-\gamma_0\|_{\mathcal{H}}\|\phi(d')\|_{\mathcal{H}_{\mathcal{D}}} \left\| \int [\mu_m(d,X)\otimes\phi(X)]\mathbb{P}(x) \right\|_{\mathcal{H}_{\mathcal{M}}\otimes\mathcal{H}_{\mathcal{X}}} \\
&\leq \kappa_d \cdot r_\gamma(n,\delta,c) \cdot \{\kappa_x \cdot r_\mu^{ME}(n,\delta,c_3) + r_\mu^{ME}(n,\delta)\} \\
&\quad + \kappa_d \cdot \|\gamma_0\|_{\mathcal{H}} \cdot \{\kappa_x \cdot r_\mu^{ME}(n,\delta,c_3) + r_\mu^{ME}(n,\delta)\} \\
&\quad + \kappa_m\kappa_d\kappa_x \cdot r_\gamma(n,\delta,c) \\
&= O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + n^{-\frac{1}{2}\frac{c_3-1}{c_3+1}} \right)
\end{aligned}
$$

□

I.4 Off-policy planning

I.4.1 Regression

We state a result analogous to Theorem I.1 for $\gamma_0(d_1,d_2,x_1,x_2)$.

Theorem I.7 (Regression rate)**.** *Suppose Assumptions C.3, C.4, and C.5 hold. Then w.p.* $1-\delta$

$$
\|\hat{\gamma}-\gamma_0\|_{\mathcal{H}} \leq r_\gamma(n,\delta,c) := \frac{\sqrt{\zeta}(c+1)}{4^{\frac{1}{c+1}}} \left(\frac{6C\kappa_d^2\kappa_x^2\ln(2/\delta)}{\sqrt{n\zeta}(c-1)} \right)^{\frac{c-1}{c+1}}
$$

Proof. Identical to Theorem I.1 □

I.4.2 Unconditional mean embedding

Theorem I.8 (Mean embedding rate)**.** *Suppose Assumptions C.3 and C.4 hold.*

1. *Then w.p.* $1-\delta$, $\forall d_1 \in \mathcal{D}$

$$\left\| \frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i}) \otimes \mu_{x_2}(d_1, x_{1i})] - \int [\phi(x_1) \otimes \mu_{x_2}(d_1, x_1)]\mathbb{P}(x_1) \right\|_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{X}} \leq r_\mu^{SATE}(n, \delta)$$

where

$$r_\mu^{SATE}(n, \delta) := \frac{4\kappa_x^2 \ln(2/\delta)}{\sqrt{n}}$$

2. *And w.p.* $1-\delta$, $\forall d_1 \in \mathcal{D}$

$$\left\| \frac{1}{n}\sum_{i=1}^{n}[\phi(\tilde{x}_{1i}) \otimes \mu_{x_2}(d_1, \tilde{x}_{1i})] - \int [\phi(x_1) \otimes \mu_{x_2}(d_1, x_1)]\tilde{\mathbb{P}}(x_1) \right\|_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{X}} \leq r_\nu^{SDS}(\tilde{n}, \delta)$$

where

$$r_\nu^{SDS}(\tilde{n}, \delta) := \frac{4\kappa_x^2 \ln(2/\delta)}{\sqrt{\tilde{n}}}$$

Proof. We show the result for θ_0^{SATE}. The result for θ_0^{SDS} is identical.
By Lemma I.1 with $\xi_i = [\phi(x_{1i}) \otimes \mu_{x_2}(d_1, x_{1i}) \otimes]$, we have

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i}) \otimes \mu_{x_2}(d_1, x_{1i})] - \int [\phi(x_1) \otimes \mu_{x_2}(d_1, x_1)]\mathbb{P}(x_1) \right\|_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{X}} \\
&\leq \frac{2\kappa_x^2 \ln(2/\delta)}{n} + \sqrt{\frac{2\kappa_x^4 \ln(2/\delta)}{n}} \\
&\leq \frac{4\kappa_x^2 \ln(2/\delta)}{\sqrt{n}}
\end{aligned}$$

□

I.4.3 Conditional mean embedding

Theorem I.9 (Conditional mean embedding rate)**.** *Suppose Assumptions C.3 and C.4.*

1. *If in addition Assumption C.6 holds then w.p.* $1-\delta$, $\forall d_1 \in \mathcal{D}$ *and* $\forall x_1 \in \mathcal{X}$

$$\|\hat{\mu}_{x_2}(d_1, x_1) - \mu_{x_2}(d_1, x_1)\|_{\mathcal{H}_\mathcal{X}} \leq r_\mu^{SATE}(\delta, n, c_4)$$

where

$$r_\mu^{SATE}(\delta, n, c_4) := \kappa_d\kappa_x \cdot \frac{\sqrt{\zeta_4}(c_4+1)}{4^{\frac{1}{c_4+1}}} \left(\frac{4\kappa_d\kappa_x(\kappa_x + \kappa_d\kappa_x\|E_4\|_{\mathcal{L}_2})\ln(2/\delta)}{\sqrt{n\zeta_4}(c_4-1)} \right)^{\frac{c_4-1}{c_4+1}}$$

2. *If in addition Assumption C.7 holds then w.p.* $1-\delta$, $\forall d_1 \in \mathcal{D}$ *and* $\forall x_1 \in \mathcal{X}$

$$\|\hat{\nu}_{x_2}(d_1, x_1) - \nu_{x_2}(d_1, x_1)\|_{\mathcal{H}_\mathcal{X}} \leq r_\nu^{SDS}(\delta, \tilde{n}, c_5)$$

where

$$r_\nu^{SDS}(\delta, \tilde{n}, c_5) := \kappa_d\kappa_x \cdot \frac{\sqrt{\zeta_5}(c_5+1)}{4^{\frac{1}{c_5+1}}} \left(\frac{4\kappa_d\kappa_x(\kappa_x + \kappa_d\kappa_x\|E_5\|_{\mathcal{L}_2})\ln(2/\delta)}{\sqrt{\tilde{n}\zeta_5}(c_5-1)} \right)^{\frac{c_5-1}{c_5+1}}$$

Proof. Immediate from [129, Corollary 1], observing that

$$E_4, E_5 : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{D} \otimes \mathcal{H}_\mathcal{X}, \quad \|\phi(d)\|_{\mathcal{H}_\mathcal{D}} \leq \kappa_d, \quad \|\phi(x)\|_{\mathcal{H}_\mathcal{X}} \leq \kappa_x$$

□

I.4.4 Target parameter

Proposition I.10. *In summary*

$$\|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} = O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}}\right)$$
$$\left\|\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i}) \otimes \mu_{x_2}(d_1, x_{1i})] - \int [\phi(x_1) \otimes \mu_{x_2}(d_1, x_1)]\mathbb{P}(x_1)\right\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}} = O_p\left(n^{-\frac{1}{2}}\right)$$
$$\left\|\frac{1}{n}\sum_{i=1}^{n}[\phi(\tilde{x}_{1i}) \otimes \mu_{x_2}(d_1, \tilde{x}_{1i})] - \int [\phi(x_1) \otimes \mu_{x_2}(d_1, x_1)]\tilde{\mathbb{P}}(x_1)\right\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}} = O_p\left(\tilde{n}^{-\frac{1}{2}}\right)$$
$$\|\hat{\mu}_{x_2}(d_1, x_1) - \mu_{x_2}(d_1, x_1)\|_{\mathcal{H}_{\mathcal{X}}} = O_p\left(n^{-\frac{1}{2}\frac{c_4-1}{c_4+1}}\right)$$
$$\|\hat{\nu}_{x_2}(d_1, x_1) - \nu_{x_2}(d_1, x_1)\|_{\mathcal{H}_{\mathcal{X}}} = O_p\left(\tilde{n}^{-\frac{1}{2}\frac{c_5-1}{c_5+1}}\right)$$

Proposition I.11. *Suppose Assumptions C.3 and C.4 hold.*

1. *If in addition Assumption C.6 holds then w.p.* $1 - 2\delta$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i}) \otimes \hat{\mu}_{x_2}(d_1, x_{1i})] - \int [\phi(x_1) \otimes \mu_{x_2}(d_1, x_1)]\mathbb{P}(x_1)\right\|_{\mathcal{H}_{\mathcal{M}} \otimes \mathcal{H}_{\mathcal{X}}}$$
$$\leq \kappa_x \cdot r_{\mu}^{SATE}(n, \delta, c_4) + r_{\mu}^{SATE}(n, \delta)$$

2. *If in addition Assumption C.7 holds then w.p.* $1 - 2\delta$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i}) \otimes \hat{\nu}_{x_2}(d_1, x_{1i})] - \int [\phi(x_1) \otimes \nu_{x_2}(d_1, x_1)]\tilde{\mathbb{P}}(x_1)\right\|_{\mathcal{H}_{\mathcal{M}} \otimes \mathcal{H}_{\mathcal{X}}}$$
$$\leq \kappa_x \cdot r_{\nu}^{SDS}(\tilde{n}, \delta, c_5) + r_{\nu}^{SDS}(\tilde{n}, \delta)$$

Proof. We prove the result for θ_0^{SATE}. The argument for θ_0^{SDS} is identical.
By triangle inequality, it is sufficient to control

$$\left\|\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i}) \otimes \hat{\mu}_{x_2}(d_1, x_{1i})] - [\phi(x_{1i}) \otimes \mu_{x_2}(d_1, x_{1i})]\right\|_{\mathcal{H}_{\mathcal{M}} \otimes \mathcal{H}_{\mathcal{X}}}$$
$$= \left\|\frac{1}{n}\sum_{i=1}^{n}\phi(x_{1i}) \otimes [\hat{\mu}_{x_2}(d_1, x_{1i}) - \mu_{x_2}(d_1, x_{1i})]\right\|_{\mathcal{H}_{\mathcal{M}} \otimes \mathcal{H}_{\mathcal{X}}}$$
$$\leq \kappa_x \max_i \|\hat{\mu}_{x_2}(d_1, x_{1i}) - \mu_{x_2}(d_1, x_{1i})\|_{\mathcal{H}_{\mathcal{M}}}$$
$$\leq \kappa_x \cdot r_{\mu}^{SATE}(n, \delta, c_4)$$

and

$$\left\|\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i}) \otimes \mu_{x_2}(d_1, x_{1i})] - \int [\phi(x_1) \otimes \mu_{x_2}(d_1, x_1)]\mathbb{P}(x_1)\right\|_{\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{X}}} \leq r_{\mu}^{SATE}(n, \delta)$$

□

Proof of Theorem C.3. We consider each off-policy effect parameter

1. θ_0^{SATE}

$$
\begin{aligned}
&\hat{\theta}^{SATE}(d_1,d_2) - \theta_0^{SATE}(d_1,d_2) \\
&= \langle \hat{\gamma}, \phi(d_1)\otimes\phi(d_2)\otimes \frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i})\otimes\hat{\mu}_{x_2}(d_1,x_{1i})]\rangle_{\mathcal{H}} \\
&\quad - \langle \gamma_0, \phi(d_1)\otimes\phi(d_2)\otimes \int \phi(x_1)\otimes\mu_{x_2}(d_1,x_1)\mathbb{P}(x_1)\rangle_{\mathcal{H}} \\
&= \langle \hat{\gamma}, \phi(d_1)\otimes\phi(d_2)\otimes \\
&\qquad \left\{\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i})\otimes\hat{\mu}_{x_2}(d_1,x_{1i})] - \int \phi(x_1)\otimes\mu_{x_2}(d_1,x_1)\mathbb{P}(x_1)\right\}\rangle_{\mathcal{H}} \\
&\quad + \langle [\hat{\gamma}-\gamma_0], \phi(d_1)\otimes\phi(d_2)\otimes \int \phi(x_1)\otimes\mu_{x_2}(d_1,x_1)\mathbb{P}(x_1)\rangle_{\mathcal{H}} \\
&= \langle [\hat{\gamma}-\gamma_0], \phi(d_1)\otimes\phi(d_2)\otimes \\
&\qquad \left\{\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i})\otimes\hat{\mu}_{x_2}(d_1,x_{1i})] - \int \phi(x_1)\otimes\mu_{x_2}(d_1,x_1)\mathbb{P}(x_1)\right\}\rangle_{\mathcal{H}} \\
&\quad + \langle \gamma_0, \phi(d_1)\otimes\phi(d_2)\otimes \\
&\qquad \left\{\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i})\otimes\hat{\mu}_{x_2}(d_1,x_{1i})] - \int \phi(x_1)\otimes\mu_{x_2}(d_1,x_1)\mathbb{P}(x_1)\right\}\rangle_{\mathcal{H}} \\
&\quad + \langle [\hat{\gamma}-\gamma_0], \phi(d_1)\otimes\phi(d_2)\otimes \int \phi(x_1)\otimes\mu_{x_2}(d_1,x_1)\mathbb{P}(x_1)\rangle_{\mathcal{H}}
\end{aligned}
$$

Therefore w.p. $1-3\delta$

$$
\begin{aligned}
&|\hat{\theta}^{SATE}(d_1,d_2) - \theta_0^{SATE}(d_1,d_2)| \\
&\le \|\hat{\gamma}-\gamma_0\|_{\mathcal{H}}\|\phi(d_1)\|_{\mathcal{H}_{\mathcal{D}}}\|\phi(d_2)\|_{\mathcal{H}_{\mathcal{D}}} \\
&\qquad \left\|\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i})\otimes\hat{\mu}_{x_2}(d_1,x_{1i})] - \int[\phi(x_1)\otimes\mu_{x_2}(d_1,x_1)]\mathbb{P}(x_1)\right\|_{\mathcal{H}_{\mathcal{M}}\otimes\mathcal{H}_{\mathcal{X}}} \\
&\quad + \|\gamma_0\|_{\mathcal{H}}\|\phi(d_1)\|_{\mathcal{H}_{\mathcal{D}}}\|\phi(d_2)\|_{\mathcal{H}_{\mathcal{D}}}\|_{\mathcal{H}_{\mathcal{D}}} \\
&\qquad \left\|\frac{1}{n}\sum_{i=1}^{n}[\phi(x_{1i})\otimes\hat{\mu}_{x_2}(d_1,x_{1i})] - \int[\phi(x_1)\otimes\mu_{x_2}(d_1,x_1)]\mathbb{P}(x_1)\right\|_{\mathcal{H}_{\mathcal{M}}\otimes\mathcal{H}_{\mathcal{X}}} \\
&\quad + \|\hat{\gamma}-\gamma_0\|_{\mathcal{H}}\|\phi(d_1)\|_{\mathcal{H}_{\mathcal{D}}}\|\phi(d_2)\|_{\mathcal{H}_{\mathcal{D}}} \left\|\int[\phi(x_1)\otimes\mu_{x_2}(d_1,x_1)]\mathbb{P}(x_1)\right\|_{\mathcal{H}_{\mathcal{M}}\otimes\mathcal{H}_{\mathcal{X}}} \\
&\le \kappa_d^2\cdot r_\gamma(n,\delta,c)\cdot\{\kappa_x\cdot r_\mu^{SATE}(n,\delta,c_4) + r_\mu^{SATE}(n,\delta)\} \\
&\quad + \kappa_d^2\cdot\|\gamma_0\|_{\mathcal{H}}\cdot\{\kappa_x\cdot r_\mu^{SATE}(n,\delta,c_4) + r_\mu^{SATE}(n,\delta)\} \\
&\quad + \kappa_d^2\kappa_x^2\cdot r_\gamma(n,\delta,c) \\
&= O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + n^{-\frac{1}{2}\frac{c_4-1}{c_4+1}}\right)
\end{aligned}
$$

2. θ_0^{SDS}

By the same argument

$$
\begin{aligned}
&|\hat{\theta}^{SDS}(d_1,d_2) - \theta_0^{SDS}(d_1,d_2)| \\
&\leq \kappa_d^2 \cdot r_\gamma(n,\delta,c) \cdot \{\kappa_x \cdot r_\nu^{SDS}(\tilde{n},\delta,c_5) + r_\nu^{SDS}(\tilde{n},\delta)\} \\
&\quad + \kappa_d^2 \cdot \|\gamma_0\|_{\mathcal{H}} \cdot \{\kappa_x \cdot r_\nu^{SDS}(\tilde{n},\delta,c_5) + r_\nu^{SDS}(\tilde{n},\delta)\} \\
&\quad + \kappa_d^2 \kappa_x^2 \cdot r_\gamma(n,\delta,c) \\
&= O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + \tilde{n}^{-\frac{1}{2}\frac{c_5-1}{c_5+1}}\right)
\end{aligned}
$$

□

I.5 Graphical effect

I.5.1 Regression

See Theorem I.1.

I.5.2 Unconditional mean embedding

See Theorem I.2.

I.5.3 Conditional mean embedding

See Theorem I.3.

I.5.4 Target parameter

Proposition I.12. *In summary*

$$
\begin{aligned}
\|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} &= O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}}\right) \\
\|\hat{\mu}_d - \mu_d\|_{\mathcal{H}_{\mathcal{D}}} &= O_p\left(n^{-\frac{1}{2}}\right) \\
\|\hat{\mu}_x(d) - \mu_x(d)\|_{\mathcal{H}_{\mathcal{X}}} &= O_p\left(n^{-\frac{1}{2}\frac{c_1-1}{c_1+1}}\right)
\end{aligned}
$$

Proposition I.13. *If Assumptions 2.3, 2.4, and 2.6 hold, then w.p.* $1-2\delta$

$$
\|\hat{\mu}_d \otimes \hat{\mu}_x(d) - \mu_d \otimes \mu_x(d)\|_{\mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}}} \leq \kappa_x \cdot r_\mu(n,\delta) + \kappa_d \cdot r_\mu^{ATT}(n,\delta,c_1)
$$

where r_μ *is as defined in Theorem I.2 (replacing* κ_x *with* κ_d*) and* r_μ^{ATT} *is as defined in Theorem I.3.*

Proposition I.14. *By triangle inequality, it is sufficient to control*

$$
\|\hat{\mu}_d \otimes \hat{\mu}_x(d) - \hat{\mu}_d \otimes \mu_x(d)\|_{\mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}}} \leq \|\hat{\mu}_d\|_{\mathcal{H}_{\mathcal{D}}} \cdot \|\hat{\mu}_x(d) - \mu_x(d)\|_{\mathcal{H}_{\mathcal{X}}} \leq \kappa_d \cdot r_\mu^{ATT}(n,\delta,c_1)
$$

and

$$
\|\hat{\mu}_d \otimes \mu_x(d) - \mu_d \otimes \mu_x(d)\|_{\mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}}} \leq \|\hat{\mu}_d - \mu_d\|_{\mathcal{H}_{\mathcal{D}}} \cdot \|\mu_x(d)\|_{\mathcal{H}_{\mathcal{X}}} \leq \kappa_x \cdot r_\mu(n,\delta)
$$

Proof of Theorem D.3.

$$
\begin{aligned}
&\hat{\theta}^{FD}(d) - \check{\theta}_0^{ATE}(d) \\
&= \langle \hat{\gamma}, \hat{\mu}_d \otimes \hat{\mu}_x(d) \rangle_{\mathcal{H}} - \langle \gamma_0, \mu_d \otimes \mu_x(d) \rangle_{\mathcal{H}} \\
&= \langle \hat{\gamma}, [\hat{\mu}_d \otimes \hat{\mu}_x(d) - \mu_d \otimes \mu_x(d)] \rangle_{\mathcal{H}} + \langle [\hat{\gamma} - \gamma_0], \mu_d \otimes \mu_x(d) \rangle_{\mathcal{H}} \\
&= \langle [\hat{\gamma} - \gamma_0], [\hat{\mu}_d \otimes \hat{\mu}_x(d) - \mu_d \otimes \mu_x(d)] \rangle_{\mathcal{H}} + \langle \gamma_0, [\hat{\mu}_d \otimes \hat{\mu}_x(d) - \mu_d \otimes \mu_x(d)] \rangle_{\mathcal{H}} \\
&\quad + \langle [\hat{\gamma} - \gamma_0], \mu_d \otimes \mu_x(d) \rangle_{\mathcal{H}}
\end{aligned}
$$

Therefore w.p. $1-3\delta$

$$
\begin{aligned}
&|\hat{\theta}^{FD}(d) - \check{\theta}_0^{ATE}(d)| \\
&\leq \|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \|\hat{\mu}_d \otimes \hat{\mu}_x(d) - \mu_d \otimes \mu_x(d)\|_{\mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}}} \\
&\quad + \|\gamma_0\|_{\mathcal{H}} \|\hat{\mu}_d \otimes \hat{\mu}_x(d) - \mu_d \otimes \mu_x(d)\|_{\mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}}} \\
&\quad + \|\hat{\gamma} - \gamma_0\|_{\mathcal{H}} \|\mu_d\|_{\mathcal{H}_{\mathcal{D}}} \|\mu_x(d)\|_{\mathcal{H}_{\mathcal{X}}} \\
&\leq r_\gamma(n, \delta, c)\{\kappa_x \cdot r_\mu(n, \delta) + \kappa_d \cdot r_\mu^{ATT}(n, \delta, c_1)\} \\
&\quad + \|\gamma_0\|_{\mathcal{H}} \|\{\kappa_x \cdot r_\mu(n, \delta) + \kappa_d \cdot r_\mu^{ATT}(n, \delta, c_1)\} \\
&\quad + \kappa_d \kappa_x \cdot r_\gamma(n, \delta, c) \\
&= O_p\left(n^{-\frac{1}{2}\frac{c-1}{c+1}} + n^{-\frac{1}{2}\frac{c_1-1}{c_1+1}}\right)
\end{aligned}
$$

□

I.6 Distribution effect

I.6.1 Conditional expectation operator

Theorem I.10 (Conditional expectation operator rate)**.** *Suppose Assumptions 2.1, E.1, E.2, and E.3 hold. Then w.p.* $1-\delta$

$$
\|\hat{E}_6 - E_6\|_{\mathcal{L}_2} \leq r_E(\delta, n, c_6) := \frac{\sqrt{\zeta_6}(c_6+1)}{4^{\frac{1}{c_6+1}}} \left(\frac{4\kappa_d \kappa_x (\kappa_y + \kappa_d \kappa_x \|E_6\|_{\mathcal{L}_2}) \ln(2/\delta)}{\sqrt{n\zeta_6}(c_6-1)}\right)^{\frac{c_6-1}{c_6+1}}
$$

Proof. Immediate from [129, Theorem 2], observing that

$$
E_6 : \mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{D}} \otimes \mathcal{H}_{\mathcal{X}}, \quad \|\phi(d)\|_{\mathcal{H}_{\mathcal{D}}} \leq \kappa_d, \quad \|\phi(x)\|_{\mathcal{H}_{\mathcal{X}}} \leq \kappa_x, \quad \|\phi(y)\|_{\mathcal{H}_{\mathcal{Y}}} \leq \kappa_y
$$

□

I.6.2 Unconditional mean embedding

See Theorem I.2.

I.6.3 Conditional mean embedding

See Theorem I.3.

I.6.4 Target parameter

Proposition I.15. *In summary*

$$
\begin{aligned}
\|\hat{E}_6 - E_6\|_{\mathcal{L}_2} &= O_p\left(n^{-\frac{1}{2}\frac{c_6-1}{c_6+1}}\right) \\
\|\hat{\mu}_x - \mu_x\|_{\mathcal{H}_{\mathcal{X}}} &= O_p\left(n^{-\frac{1}{2}}\right) \\
\|\hat{\mu}_x(d) - \mu_x(d)\|_{\mathcal{H}_{\mathcal{X}}} &= O_p\left(n^{-\frac{1}{2}\frac{c_1-1}{c_1+1}}\right)
\end{aligned}
$$

Proof of Theorem E.3. We consider each distribution embedding

1. $\check{\theta}_0^{DATE}$

$$
\begin{aligned}
&\hat{\theta}^{DATE}(d) - \check{\theta}_0^{DATE}(d) \\
&= \hat{E}_6^*[\phi(d) \otimes \hat{\mu}_x] - E_6^*[\phi(d) \otimes \mu_x] \\
&= \hat{E}_6^*[\phi(d) \otimes \{\hat{\mu}_x - \mu_x\}] + \{\hat{E}_6^* - E_6^*\}[\phi(d) \otimes \mu_x] \\
&= \{\hat{E}_6^* - E_6^*\}[\phi(d) \otimes \{\hat{\mu}_x - \mu_x\}] + E_6^*[\phi(d) \otimes \{\hat{\mu}_x - \mu_x\}] \\
&\quad + \{\hat{E}_6^* - E_6^*\}[\phi(d) \otimes \mu_x]
\end{aligned}
$$

Therefore for all $d \in \mathcal{D}$, w.p. $1 - 2\delta$

$$\begin{aligned}
&\|\hat{\theta}^{DATE}(d) - \check{\theta}_0^{DATE}(d)\|_{\mathcal{H}_\mathcal{Y}} \\
&\leq \|\hat{E}_6 - E_6\|_{\mathcal{L}_2}\|\phi(d)\|_{\mathcal{H}_\mathcal{D}}\|\hat{\mu}_x - \mu_x\|_{\mathcal{H}_\mathcal{X}} + \|E_6\|_{\mathcal{L}_2}\|\phi(d)\|_{\mathcal{H}_\mathcal{D}}\|\hat{\mu}_x - \mu_x\|_{\mathcal{H}_\mathcal{X}} \\
&\quad + \|\hat{E}_6 - E_6\|_{\mathcal{L}_2}\|\phi(d)\|_{\mathcal{H}_\mathcal{D}}\|\mu_x\|_{\mathcal{H}_\mathcal{X}} \\
&\leq \kappa_d \cdot r_E(n, \delta, c_6) \cdot r_\mu(n, \delta) + \kappa_d \cdot \|E_6\|_{\mathcal{L}_2} \cdot r_\mu(n, \delta) + \kappa_d\kappa_x \cdot r_E(n, \delta, c_6) \\
&= O_p\left(n^{-\frac{1}{2}\frac{c_6-1}{c_6+1}}\right)
\end{aligned}$$

2. $\check{\theta}_0^{DATT}$

$$\begin{aligned}
&\hat{\theta}^{DATT}(d, d') - \check{\theta}_0^{DATT}(d, d') \\
&= \hat{E}_6^*[\phi(d') \otimes \hat{\mu}_x(d)] - E_6^*[\phi(d') \otimes \mu_x(d)] \\
&= \hat{E}_6^*[\phi(d') \otimes \{\hat{\mu}_x(d) - \mu_x(d)\}] + \{\hat{E}_6^* - E_6^*\}[\phi(d') \otimes \mu_x(d)] \\
&= \{\hat{E}_6^* - E_6^*\}[\phi(d') \otimes \{\hat{\mu}_x(d) - \mu_x(d)\}] + E_6^*[\phi(d') \otimes \{\hat{\mu}_x(d) - \mu_x(d)\}] \\
&\quad + \{\hat{E}_6^* - E_6^*\}[\phi(d') \otimes \mu_x(d)]
\end{aligned}$$

Therefore for all $d, d' \in \mathcal{D}$, w.p. $1 - 2\delta$

$$\begin{aligned}
&\|\hat{\theta}^{DATT}(d, d') - \check{\theta}_0^{DATT}(d, d')\|_{\mathcal{H}_\mathcal{Y}} \\
&\leq \|\hat{E}_6 - E_6\|_{\mathcal{L}_2}\|\phi(d')\|_{\mathcal{H}_\mathcal{D}}\|\hat{\mu}_x(d) - \mu_x(d)\|_{\mathcal{H}_\mathcal{X}} \\
&\quad + \|E_6\|_{\mathcal{L}_2}\|\phi(d')\|_{\mathcal{H}_\mathcal{D}}\|\hat{\mu}_x(d) - \mu_x(d)\|_{\mathcal{H}_\mathcal{X}} \\
&\quad + \|\hat{E}_6 - E_6\|_{\mathcal{L}_2}\|\phi(d')\|_{\mathcal{H}_\mathcal{D}}\|\mu_x(d)\|_{\mathcal{H}_\mathcal{X}} \\
&\leq \kappa_d \cdot r_E(n, \delta, c_6) \cdot r_\mu^{ATT}(n, \delta, c_1) + \kappa_d \cdot \|E_6\|_{\mathcal{L}_2} \cdot r_\mu^{ATT}(n, \delta, c_1) + \kappa_d\kappa_x \cdot r_E(n, \delta, c_6) \\
&= O_p\left(n^{-\frac{1}{2}\frac{c_1-1}{c_1+1}} + n^{-\frac{1}{2}\frac{c_6-1}{c_6+1}}\right)
\end{aligned}$$

□

Proof of Theorem E.4. We prove the result for θ_0^{DATE}. The argument for θ_0^{DATE} is identical.

Fix d. By Theorem E.3

$$\|\hat{\theta}^{DATE}(d) - \check{\theta}_0^{DATE}(d)\|_{\mathcal{H}_\mathcal{Y}} = O_p\left(n^{-\frac{1}{2}\frac{c_6-1}{c_6+1}}\right)$$

Denote the samples constructed by Algorithm E.2 by $\{\tilde{y}_j\}_{j\in[m]}$. Then by [11, Section 4.2]

$$\left\|\hat{\theta}^{DATE}(d) - \frac{1}{m}\sum_{j=1}^{m}\phi(\tilde{y}_j)\right\|_{\mathcal{H}_\mathcal{Y}} = O(m^{-\frac{1}{2}})$$

Therefore by triangle inequality

$$\left\|\frac{1}{m}\sum_{j=1}^{m}\phi(\tilde{y}_j) - \check{\theta}_0^{DATE}(d)\right\|_{\mathcal{H}_\mathcal{Y}} = O_p\left(n^{-\frac{1}{2}\frac{c_6-1}{c_6+1}} + m^{-\frac{1}{2}}\right)$$

The desired result follows from [133], as quoted by [128, Theorem 1.1]. □

J Tuning

J.1 Simplified setting

In the present work, we propose a family of novel estimators that are inner products of kernel ridge regressions. As such, the same two kinds of hyperparameters that arise in kernel ridge regressions

arise in our estimators: ridge regression penalties and kernel hyperparameters. In this section, we describe practical tuning procedures for such hyperparameters. To simplify the discussion, we focus on the regression of Y on W. Recall that the closed form solution of the regression estimator using all observations is

$$\hat{f}(w) = K_{wW}(K_{WW} + n\lambda)^{-1}Y$$

J.2 Ridge penalty

It is convenient to tune λ by leave-one-out cross validation (LOOCV), since the validation loss has a closed form solution.

Algorithm J.1 (Ridge penalty tuning)**.** *Construct the matrices*

$$K_{WW} \in \mathbb{R}^{n\times n}, \quad H_\lambda := I - K_{WW}(K_{WW} + n\lambda)^{-1} \in \mathbb{R}^{n\times n}, \quad \tilde{H}_\lambda := diag(H_\lambda) \in \mathbb{R}^{n\times n}$$

where $\tilde{H}_\lambda$ has the same diagonal entries as H_λ and off-diagonal entries of 0. Then set

$$\lambda^* = \operatorname*{argmin}_{\lambda\in\Lambda} \frac{1}{n}\|\tilde{H}_\lambda^{-1}H_\lambda Y\|_2^2, \quad \Lambda \subset \mathbb{R}$$

Derivation. We prove that $\frac{1}{n}\|\tilde{H}_\lambda^{-1}H_\lambda Y\|_2^2$ is the LOOCV loss. By definition, the LOOCV loss is

$$\mathcal{E}(\lambda) := \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}_{-i}(x_i)]^2$$

where $\hat{f}_{-i}$ is the regression estimator using all observations except the i-th observation.

To lighten notation, let Φ be the matrix of features, with i-th row $\phi(x_i)^T$, and let $Q := \Phi^T\Phi + n\lambda$. By the regression first order condition

$$\begin{aligned}\hat{f} &= Q^{-1}\Phi^T Y \\ \hat{f}_{-i} &= \{Q - \phi(x_i)\phi(x_i)^T\}^{-1}\{\Phi^T Y - \phi(x_i)y_i\}\end{aligned}$$

Viewing $-\phi(x_i)\phi(x_i)^T$ as a rank-one update, the Sherman-Morrison formula gives

$$\hat{f}_{-i} = \hat{f} + \frac{Q^{-1}\phi(x_i)\phi(x_i)^T\hat{f} - Q^{-1}\phi(x_i)y_i}{1-\beta_i}, \quad \beta_i := \phi(x_i)^T Q^{-1}\phi(x_i)$$

i.e. $\hat{f}_{-i}$ can be expressed in terms of $\hat{f}$.

Substituting back into the LOOCV loss

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_{-i}(x_i)\right]^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}(x_i) - \frac{\phi(x_i)^T Q^{-1}\phi(x_i)\hat{f}(x_i) - \phi(x_i)^T Q^{-1}\phi(x_i)y_i}{1-\beta_i}\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}(x_i) - \frac{\beta_i\hat{f}(x_i) - \beta_i y_i}{1-\beta_i}\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\{y_i - \hat{f}(x_i)\}\left\{1 + \frac{\beta_i}{1-\beta_i}\right\}\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\{y_i - \hat{f}(x_i)\}\left\{\frac{1}{1-\beta_i}\right\}\right]^2 \\ &= \frac{1}{n}\|\tilde{H}_\lambda^{-1}\{Y - K_{WW}(K_{WW} + n\lambda)^{-1}Y\}\|_2^2 \\ &= \frac{1}{n}\|\tilde{H}_\lambda^{-1}H_\lambda Y\|_2^2\end{aligned}$$

□

J.3 Kernel

The Gaussian kernel is the most popular kernel among machine learning practitioners.

$$k(w, w') = \exp\left\{-\frac{1}{2}\frac{\|w - w'\|_{\mathcal{W}}^2}{\sigma^2}\right\}$$

Importantly, this kernel satisfies the required properties: it is continuous, bounded, and characteristic. See [134, 135] for its additional properties. Observe that the Gaussian kernel has a hyperparameter: the *lengthscale* σ. A convenient heuristic is to set the lengthscale equal to the median interpoint distance of $\{w_i\}_{i=1}^n$, where the interpoint distance between observations i and j is $\|w_i - w_j\|_{\mathcal{W}}$.

In our implementations, we use the Gaussian kernel. When the input W is multidimensional, we use the kernel obtained as the product of scalar kernels for each input dimension. For example, if $\mathcal{W} \subset \mathbb{R}^d$ then

$$k(w, w') = \prod_{j=1}^{d} \exp\left\{-\frac{1}{2}\frac{[w_j - w_j']^2}{\sigma_j^2}\right\}$$

Each lengthscale σ_j is set according to the median interpoint distance for that input dimension.

K Experiment details

K.1 Simulation: Continuous treatment effect

A single observation consists of the triple (Y, D, X) for outcome, treatment, and covariates where $Y, D \in \mathbb{R}$ and $X \in \mathbb{R}^{100}$. A single observation is generated is as follows. Draw unobserved noise as $\nu, \epsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Define the vector $\beta \in \mathbb{R}^{100}$ by $\beta_j = j^{-2}$. Define the matrix $\Sigma \in \mathbb{R}^{100\times 100}$ such that $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 1\{|i - j| = 1\}$ for $i \neq j$. Then draw $X \sim \mathcal{N}(0, \Sigma)$ and set

$$D = \Phi(3X'\beta) + 0.75\nu, \quad Y = 1.2D + 1.2D'\beta + D^2 + DX_1 + \epsilon$$

We implement our estimator $\hat{\theta}^{ATE}(d)$ (`MeanEmb`) described in Section 2, with the tuning procedure described in Appendix J. Specifically, we use ridge penalties determined by leave-one-out cross validation, and product Gaussian kernel with lengthscales set by the median heuristic. We implement [92] (`DML1`) using the default settings of the command `ctseff` in the `R` package `npcausal`. We implement [32] (`DML2`) using default settings in `Python` code shared by the authors. Specifically, we use random forest for prediction, with the suggested hyperparameter values. For the Nadaraya-Watson smoothing, we select bandwidth that minimizes out-of-sample MSE. For [32] (`DML2`), we also report results for its plug-in and generalized IPW components (`PlugIn`, `IPW`).

K.2 Simulation: Mediated effect

The mediated effect design [80] involves learning the counterfactual function

$$\theta_0^{ME}(d, d') = 0.3d' + 0.09d + 0.15dd' + 0.25 \cdot (d')^3$$

A single observation consists of the tuple (Y, D, M, X) for outcome, treatment, mediator, and covariates where $Y, D, M, X \in \mathbb{R}$. A single observation is generated is as follows. Draw unobserved noise as $u, v, w \overset{i.i.d.}{\sim} \mathcal{U}(-2, 2)$. Draw the covariate as $X \sim \mathcal{U}(-1.5, 1.5)$. Then set

$$\begin{aligned} D &= 0.3X + w \\ M &= 0.3D + 0.3X + v \\ Y &= 0.3D + 0.3M + 0.5DM \\ &\quad + 0.3X + 0.25D^3 + u \end{aligned}$$

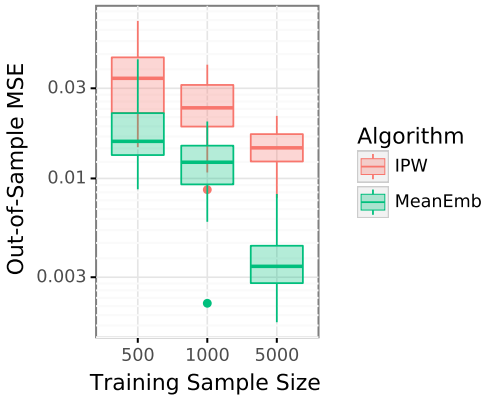


Figure 6: Mediated effect simulation

Note that [80] also present a simpler version of this design.

In addition to our estimator (`MeanEmb`), we implement [80] (`IPW`), which involves Nadaraya-Watson kernel density estimation en route to a generalized inverse propensity weighting estimate. We implement our estimator $\hat{\theta}^{ME}(d, d')$ (`MeanEmb`) described in Appendix B, with the tuning procedure described in Appendix J. Specifically, we use ridge penalties determined by leave-one-out cross validation, and product Gaussian kernel with lengthscales set by the median heuristic. We implement [80] (`IPW`) using the default settings of the command `medweightcont` in the `R` package `causalweight`.

K.3 Application: Continuous and heterogeneous treatment effects

We implement our estimators $\hat{\theta}^{ATE}(d)$ and $\hat{\theta}^{CATE}(d, v)$ described in Section 2, with the tuning procedure described in Appendix J. Specifically, we use ridge penalties determined by leave-one-out cross validation, and product Gaussian kernel with lengthscales set by the median heuristic.

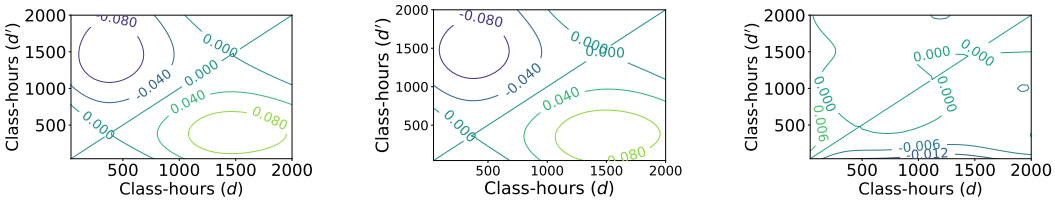
K.4 Application: Total, direct, and indirect effects

Next, we consider arrests by the police to be the outcome of interest. We consider employment to be a possible mechanism through which training hours affect arrests, following [48, 80]. In this setting, the outcome $Y \in \mathbb{R}$ is the number of times an individual is arrested by police; the treatment $D \in \mathbb{R}$ is total hours spent in academic academic or vocational classes; and the mediator $M \in \mathbb{R}$ is the proportion of weeks employed. The covariates $X \in \mathbb{R}^{40}$ are as before. We use the same sample as before.

Recall that the fundamental quantity of interest is the mediated effect $\theta_0^{ME}(d, d')$, which in turn produces total, direct, and indirect effects $(\theta_0^{TE}, \theta_0^{DE}, \theta_0^{IE})$. In particular, $\theta_0^{TE}(d, d')$ is the total effect of class-hours d' relative to class-hours d on arrests; $\theta_0^{DE}(d, d')$ is the direct effect of class-hours d' relative to class-hours d on arrests; and $\theta_0^{IE}(d, d')$ is the indirect effect of class-hours d' relative to class-hours d on arrests, as mediated by the mechanism of employment. We implement our estimator $\hat{\theta}^{ME}(d, d')$ described in Appendix B, with the tuning procedure described in Appendix J. Specifically, we use ridge penalties determined by leave-one-out cross validation, and product Gaussian kernel with lengthscales set by the median heuristic. Figure 7 visualizes results.

The total effect of training on arrests is negative. At best, the total effect of receiving 1,600 class-hours (40 weeks) versus 480 class-hours (12 weeks) is a reduction of 0.1 arrests. In Section 3, we recommend a policy of 12-14 weeks of classes to optimize the employment effect. We see that there would be no total effect of this proposed policy on arrests.

The direct effect of class-hours on arrests is negative, with the same magnitude as the total effect. The indirect effect of class-hours on arrests, as mediated through employment, is essentially zero. Our estimates have the same shape but are smoother than those of [80]. We conclude that the effect of class-hours on arrests is purely direct; class-hours decrease arrests, but *not* via the economic mechanism of increasing employment. From a policy perspective, there are benefits of the training program that are not explained by employment alone. By measuring direct and indirect treatment effects, our results can guide economic modeling.



(a) Total effect (b) Direct effect (c) Indirect effect

Figure 7: Total, direct, and indirect effect of class-hours on arrests