estimation! pearl's paper 2012: how do you identify causal effect if you only have w? 220222

why we can use proxy to identify the causal effect?

2014: introduce 2 proxy ideas

# Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restriction

Afsaneh Mastouri * 1 Yuchen Zhu * 1 Limor Gultchin 2 3 Anna Korba 4 Ricardo Silva 1 Matt J. Kusner 1
Arthur Gretton † 1 Krikamol Muandet † 5

## Abstract

We address the problem of causal effect estimation in the presence of unobserved confounding, but where proxies for the latent confounder(s) are observed. We propose two kernel-based methods for nonlinear causal effect estimation in this setting: (a) a two-stage regression approach, and (b) a maximum moment restriction approach. We focus on the proximal causal learning setting, but our methods can be used to solve a wider class of inverse problems characterised by a Fredholm integral equation. In particular, we provide a unifying view of two-stage and moment restriction approaches for solving this problem in a nonlinear setting. We provide consistency guarantees for each algorithm, and demonstrate that these approaches achieve competitive results on synthetic data and data simulating a real-world task. In particular, our approach outperforms earlier methods that are not suited to leveraging proxy variables.

## 1 Introduction

Estimating average treatment effects (ATEs) is critical to answering many scientific questions. From estimating the effects of medical treatments on patient outcomes (Connors et al., 1996; Choi et al., 2002), to grade retention on cognitive development (Fruehwirth et al., 2016), ATEs are the key estimands of interest. From observational data alone, however, estimating such effects is impossible without further assumptions. This impossibility arises from potential unobserved confounding: one variable may seem to cause another, but this could be due entirely to an unobserved variable causing both of them, e.g., as was used by the tobacco industry to argue against the causal link between smoking and lung cancer (Cornfield et al., 1959).

unobserved confounding.

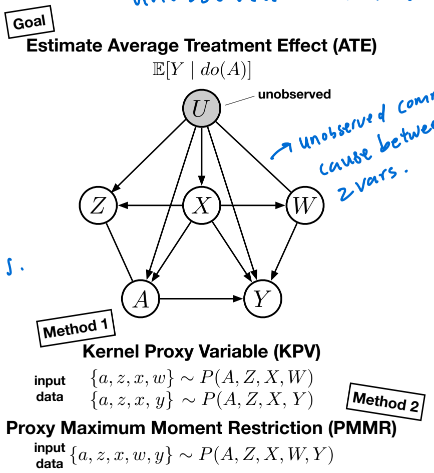

Goal
Estimate Average Treatment Effect (ATE)
$\mathbb{E}[Y \mid do(A)]$
unobserved
U
Z
X
W
A
Y
Method 1
Kernel Proxy Variable (KPV)
input data $\{a, z, x, w\} \sim P(A, Z, X, W)$
$\{a, z, x, y\} \sim P(A, Z, X, Y)$
Method 2
Proxy Maximum Moment Restriction (PMMR)
input data $\{a, z, x, w, y\} \sim P(A, Z, X, W, Y)$

Figure 1: The causal proxy estimation problem, and two methods we introduce to solve it.

Y: outcome

A: treatment

X: observed confounders.

unobserved common cause between 2 vars.

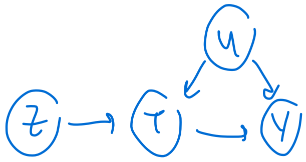
One of the most common assumptions to bypass this difficulty is to assume that no unobserved confounders exist (Imbens, 2004). This extremely restrictive assumption makes estimation easy: if there are also no observed confounders then the ATE can be estimated using simple regression, otherwise one can use backdoor adjustment (Pearl, 2000). Less restrictive is to assume observation of an *instrumental variable* (IV) that is independent of any unobserved confounders (Reiersøl, 1945). This independence assumption is often broken, however. For example, if medication requires payment, many potential instruments such as educational attainment will be confounded with the outcome through complex socioeconomic factors, which can be difficult to fully observe (e.g., different opportunities afforded by living in different neighborhoods). The same argument regarding unobserved confounding can be made for the grade retention and household expenditure settings.

This fundamental difficulty has inspired work to investigate the relaxation of this independence assumption. An increasingly popular class of models, called *proxy* models, does just this; and various recent studies incorporate proxies into

*Equal contribution †Equal contribution [1]University College London, London, United Kingdom [2]University of Oxford, Oxford, United Kingdom [3]The Alan Turing Institute, London, United Kingdom [4]ENSAE/CREST, Paris, France [5]Max Planck Institute for Intelligent Systems, Tübingen, Germany. Correspondence to: Afsaneh Mastouri <afs.mastouri@gmail.com>, Yuchen Zhu <yuchen.zhu.18@ucl.ac.uk>.

*Proceedings of the 38th International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

IV



no BD path from z to y going through u.

causal discovery and inference tasks to reduce the influence of confounding bias (Cai & Kuroki, 2012; Tchetgen Tchetgen, 2014; Schuemie et al., 2014; Sofer et al., 2016; Flanders et al., 2017; Shi et al., 2018). Consider the following example from Deaner (2018), described graphically in Figure 1: we wish to understand the effect (i.e., ATE) of holding children back a grade in school (also called 'grade retention') $A$ , on their math scores, $Y$. This relationship is confounded by an unobserved variable $U$ describing students' willingness to learn in school. Luckily we have access to a *proxy* of $U$, student scores from a cognitive and behavioral test $W$. Note that if $W = U$ we could use backdoor adjustment to estimate the effect of $A$ on $Y$ (Pearl, 2000). In general $W \neq U$, however, and this adjustment would produce a biased estimate of the ATE. In this case we can introduce a second proxy $Z$: the cognitive and behavioral test result *after* grade retention $A$. This allows us to form an integral equation similar to the IV setting. The solution to this equation is not the ATE (as it is in the IV case) but a function that, when adjusted over the distribution $P(W)$ (or $P(W, X)$ in the general case) gives the true causal effect (Kuroki & Pearl, 2014; Tchetgen Tchetgen et al., 2020). Building on Carroll et al. (2006) and Greenland & Lash (2011), Kuroki & Pearl (2014) were the first to demonstrate the possibility of identifying the causal effect given access to proxy variables. This was generalized by Miao & Tchetgen Tchetgen (2018), and Tchetgen Tchetgen et al. (2020) recently proved non-parametric identifiability for the general proxy graph (i.e., including $X$) shown in Figure 1.

The question of how to *estimate* the ATE in this graph for continuous variables is still largely unexplored, however, particularly in a non-linear setting, and with consistency guarantees. Deaner (2018) assume a sieve basis and describe a technique to identify a different causal quantity, the average treatment effect on the treated (ATT), in this graph (without $X$, but the work can be easily extended to include $X$). Tchetgen Tchetgen et al. (2020) assume linearity and estimate the ATE. The linearity assumption significantly simplifies estimation, but the ATE in principle can be identified without parametric assumptions (Tchetgen Tchetgen et al., 2020). At the same time, there have been exciting developments in using kernel methods to estimate causal effects in the non-linear IV setting, with consistency guarantees (Singh et al., 2019; Muandet et al., 2020b; Zhang et al., 2020). Kernel approaches to ATE, ATT, Conditional ATE, and causal effect estimation under distribution shift, have also been explored in various settings (Singh et al., 2020; Singh, 2020).

In this work, we propose two kernelized estimation procedures for the ATE in the proxy setting, with consistency guarantees: (a) a two-stage regression approach (which we refer to as Kernelized Proxy Variables, or KPV), and (b) a maximum moment restriction approach (which we refer to as Proxy Maximum Moment Restriction, or PMMR). Alongside consistency guarantees, we derive a theoretical connection between both approaches, and show that our methods can also be used to solve a more general class of inverse problems that involve a solution to a Fredholm integral equation. We demonstrate the performance of both approaches on synthetic data, and on data simulating real-world tasks.

## 2 Background

Throughout, a capital letter (e.g. $A$) denotes a random variable on a measurable space, denoted by a calligraphic letter (resp. $\mathcal{A}$). We use lowercase letters to denote the realization of a random variable (e.g. $A = a$).

### 2.1 Causal Inference with Proxy Variables

Our goal is to estimate the average treatment effect (ATE) of treatment $A$ on outcome $Y$ in the proxy causal graph of Figure 1 (throughout we will assume $Y$ is scalar and continuous; the discrete case is much simpler (Miao & Tchetgen Tchetgen, 2018)). To do so, we are given access to proxies of an unobserved confounder $U$: a treatment-inducing proxy $Z$, an outcome-inducing proxy $W$; and optionally observed confounders $X$. Formally, given access to samples from either the joint distribution $\rho(A, Z, X, W, Y)$ or from both distributions $\rho(A, Z, X, W)$ and $\rho(A, Z, X, Y)$, we aim to estimate the ATE $\mathbb{E}[Y \mid do(A = a)]$. Throughout we will describe causality using the structural causal model (SCM) formulation of Pearl (2000). Here, the causal relationships are represented as directed acyclic graphs. The crucial difference between these models and standard probabilistic graphical models is a new operator: the intervention $do(\cdot)$. This operator describes the process of forcing a random variable to take a particular value, which isolates its effect on downstream variables (i.e., $\mathbb{E}[Y \mid do(A = a)]$ describes the isolated effect of $A$ on $Y$). We start by introducing the assumptions that are necessary to identify this causal effect. These assumptions can be divided into two classes: (A) *structural assumptions* and (B) *completeness assumptions*.

(A) *Structural assumptions* via conditional independences:

**Assumption 1** $Y \perp\!\!\!\perp Z \mid A, U, X$.

**Assumption 2** $W \perp\!\!\!\perp (A, Z) \mid U, X$.

These assumptions are very general: they do not enforce restrictions on the functional form of the confounding effect, or indeed on any other effects. Note that we are not restricting the confounding structure, since we do not make any assumption on the additivity of confounding effect, or on the linearity of the relationship between variables.

(B) *Completeness assumptions* on the ability of proxy variables to characterize the latent confounder:

← Stronger assumption.

**Assumption 3** Let $l$ be any square integrable function. Then $\mathbb{E}[l(U) \mid a, x, z] = 0$ for all $(a, x, z) \in \mathcal{A} \times \mathcal{X} \times \mathcal{Z}$ , if and only if $l(U) = 0$ almost surely.

Other assumption is just smoothness

have BD adjustment

- $E[Y ; do(A=a)] = \underset{x,u}{E}[E[Y|u,x,a]] = \int_{x,w} h(a,x,w)\, p(x,w)\, dx\, dw$

BD formula.

- $E[Y|a,x,z] = \int_{w} p(w|a,x,z)\, E[Y|a,x,w,z]\, dw$

$E[g(x)] = 0$

$g(x) = 0$

**Assumption 4** Let $g$ be any square integrable function. Then $\mathbb{E}[g(Z) \mid a, x, w] = 0, \forall (a, x, w) \in \mathcal{A} \times \mathcal{X} \times \mathcal{W}$ if and only if $g(Z) = 0$ almost surely.

These assumptions guarantee that the proxies are sufficient to describe $U$ for the purposes of ATE estimation. For better intuition we can look at the discrete case: for categorical $U, Z, W$ the above assumptions imply that proxies $W$ and $Z$ have at least as many categories as $U$. Further, it can be shown that Assumption 4 along with certain regularity conditions (Miao et al., 2018, Appendix, Conditions (v)-(vii)) guarantees that there exists at least one solution to the following integral equation:

$$\mathbb{E}[Y \mid a, x, z] = \int_{\mathcal{W}} h(a, x, w)\rho(w|a, x, z)\, dw, \qquad (1)$$

which holds for all $(a, x, z) \in \mathcal{A} \times \mathcal{X} \times \mathcal{Z}$. We discuss the completeness conditions in greater detail in Appendix A.

Given these assumptions, it was shown by Miao & Tchetgen Tchetgen (2018) that the function $h(a, x, w)$ in (1) can be used to identify the causal effect $\mathbb{E}[Y \mid do(A = a)]$ as follows,

$$\mathbb{E}[Y \mid do(A = a)] = \int_{\mathcal{X},\mathcal{W}} h(a, x, w)\rho(x, w)\, dx dw. \quad (2)$$

While the causal effect can be identified, approaches for estimating this effect in practice are less well established, and include Deaner (2018) (via a method of sieves) and Tchetgen Tchetgen et al. (2020) (assuming linearity). The related IV setting has well established estimation methods, however the proximal setting relies on fundamentally different assumptions on the data generating process. None of the three key assumptions in the IV setting (namely the relevance condition, exclusion restriction, or unconfounded instrument) are required in proximal setting. In particular, we need a set of proxies which are complete for the latent confounder, i.e., dependent with the latent confounder, whereas a valid instrument is independent of the confounder. In this respect, the proximal setting is more general than the IV setting, including the recent "IVY" method of Kuang et al. (2020).

Before describing our approach to the problem of estimating the causal effect in (2), we give a brief background on reproducing kernel Hilbert spaces and the additional assumptions we need for estimation.

Set of function from F → R

$k(x,y) = k(y,x)$

### 2.2 Reproducing Kernel Hilbert Spaces (RKHS)

//

Symmetry

For any space $\mathcal{F} \in \{\mathcal{A}, \mathcal{X}, \mathcal{W}, \mathcal{Z}\}$, let $k : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be a positive semidefinite kernel. We denote by $\phi$ its associated canonical feature map $\phi(x) = k(x, \cdot)$ for any $x \in \mathcal{F}$, and $\mathcal{H}_{\mathcal{F}}$ its corresponding RKHS of real-valued functions on $\mathcal{F}$. The space $\mathcal{H}_{\mathcal{F}}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{F}}}$ and norm $\| \cdot \|_{\mathcal{H}_{\mathcal{F}}}$. It satisfies two important properties: (i) $k(x, \cdot) \in \mathcal{H}_{\mathcal{F}}$ for all $x \in \mathcal{F}$, (ii) the reproducing property: for all $f \in \mathcal{H}_{\mathcal{F}}$ and $x \in \mathcal{F}$, $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}_{\mathcal{F}}}$. We denote the tensor product and Hadamard product by $\otimes$ and $\odot$ respectively. For $\mathcal{F}, \mathcal{G} \in \{\mathcal{A}, \mathcal{X}, \mathcal{W}, \mathcal{Z}\}$, we will use $\mathcal{H}_{\mathcal{FG}}$ to denote the product space $\mathcal{H}_{\mathcal{F}} \times \mathcal{H}_{\mathcal{G}}$. It can be shown that $\mathcal{H}_{\mathcal{FG}}$ is isometrically isomorphic to $\mathcal{H}_{\mathcal{F}} \otimes \mathcal{H}_{\mathcal{G}}$. For any distribution $\rho$ on $\mathcal{F}$, $\mu_\rho := \int k(x, \cdot) d\rho(x)$ is an element of $\mathcal{H}_{\mathcal{F}}$ and is referred to as the kernel mean embedding of $\rho$ (Smola et al., 2007). Similarly, for any conditional distribution $\rho_{X|z}$ for each $z \in \mathcal{Z}$, $\mu_{X|z} := \int k(x, \cdot) d\rho(x|z)$ is a conditional mean embedding of $\rho_{X|z}$ (Song et al., 2009; 2013); see Muandet et al. (2017) for a review.

$K_x \in \mathbb{H}$ such that $eva_x(f) = < f, \mid$

$X \longrightarrow R$

### 2.3 Estimation Assumptions

To enable causal effect estimation in the proxy setting using kernels, we require the following additional assumptions.

**Assumption 5** (Regularity condition) $\mathcal{A}, \mathcal{X}, \mathcal{Y}, \mathcal{W}, \mathcal{Z}$ are measurable, separable Polish spaces. defined as :

some identification

Assumption 5 allows us to define the conditional mean embedding operator and a Hilbert–Schmidt operator.

$p_1 \neq p_2 \quad \mu_{p_1} \neq \mu_{p_2}$

**Assumption 6** $\exists c_Y < \infty, |Y| < c_Y$ a.s. and $\mathbb{E}[Y] < c_Y$. ↗ $Prob(x) \to \mathbb{H}$

$p \to \mu_p.$

**Assumption 7** (Kernels) (i) $k(w, \cdot)$ is a characteristic kernel. (ii) $k(a, \cdot)$, $k(x, \cdot)$, $k(w, \cdot)$ and $k(z, \cdot)$ are continuous, bounded by $\kappa > 0$, and their feature maps are measurable.

2 different $p$s can't give the same $\mu_p$.

The kernel mean embedding of any probability distribution is injective if a characteristic kernel is used (Sriperumbudur et al., 2011); this guarantees that a probability distribution can be uniquely represented in an RKHS.

**Assumption 8** The measure $\mathcal{P}_{\mathcal{AWX}}$ is a finite Borel measure with $\text{supp}[P_{AWX}] = \mathcal{A} \times \mathcal{W} \times \mathcal{X}$.

We will assume that the problem is well-posed.

**Assumption 9** Let $h$ be the function defined in (1). We assume that $h \in \mathcal{H}_{\mathcal{AXW}}$.

Finally, given assumption 9, we require the following completeness condition.

**Assumption 10** (Completeness condition in RKHS) For all $g \in \mathcal{H}_{\mathcal{AWX}}$: $\mathbb{E}_{AWX}[g(A, W, X)|A, Z, X] = 0\ \mathcal{P}_{\mathcal{AZX}}$-almost surely if and only if $g(a, w, x) = 0$, $\mathcal{P}_{\mathcal{AWX}}$−almost surely. probability = 1

This condition guarantees the uniqueness of the solution to the integral equation (1) in RKHS (see Lemma 10 in Appendix C).

## 3 Kernel Proximal Causal Learning

To solve the proximal causal learning problem, we propose two kernel-based methods, *Kernel Proxy Variable* (KPV) and *Proxy Maximum Moment Restriction* (PMMR). The KPV decomposes the problem of learning function $h$ in (1) into two stages: we first learn an empirical representation of $\rho(w|a, x, z)$, and then learn $h$ as a mapping from representation of $\rho(w|a, x, z)$ to $y$, with kernel ridge regression

$\mathop{E}_{AWX} [ y(A, W, X) \mid A, Z, X ] = 0 \qquad g(a, w, n) = 0$

$E[y \mid do(A=a)] = \int_{x,w} h'(a, x, w) p(w \mid a, x, z) dw \quad - E[y \mid do(A=a)] = \int_{x,w} h''(a, x, w) p(w \mid a, x, z) dw$

$0 = \int_w [h' - h''] p(w \mid a, x, z) dw$

$0 = E[g(A,w,x) | A,X,z]$

$h' - h'' = 0$

as the main apparatus of learning. This procedure is similar to Kernel IV regression (KIV) proposed by Singh et al. (2019). PMMR, on the other hand, employs the Maximum Moment Restriction (MMR) framework (Muandet et al., 2020a), which takes advantage of a closed-form solution for a kernelized conditional moment restriction. The structural function can be estimated in a single stage with a modified ridge regression objective. We clarify the connection between both approaches at the end of this section.

### 3.1 Kernel Proxy Variable (KPV)

To solve (1), the KPV approach finds $h \in \mathcal{H}_{\mathcal{AXW}}$ that minimizes the following risk functional:

$$\tilde{R}(h) = \mathbb{E}_{AXZY}\left[(Y - G_h(A,X,Z))^2\right], \tag{3}$$
$$G_h(a,x,z) := \int_{\mathcal{W}} h(a,x,w)\rho(w \mid a,x,z)dw$$

Let $\mu_{W|a,x,z} \in \mathcal{H}_{\mathcal{W}}$ be the conditional mean embedding of $\rho(W \mid a,x,z)$. Then, for any $h \in \mathcal{H}_{\mathcal{AXW}}$, we have:

$$G_h(a,x,z) = \left\langle h, \phi(a,x) \otimes \mu_{W|a,x,z} \right\rangle_{\mathcal{H}_{\mathcal{AXW}}} \tag{4}$$

where $\phi(a,x) = \phi(a) \otimes \phi(x)$. This result arises from the properties of the RKHS tensor space $\mathcal{H}_{\mathcal{AXW}}$ and of the conditional mean embedding. We denote by $\eta_{AXW}$ the particular function $h$ minimizing (3).

The procedure to solve (3) consists of two ridge regression stages. In the first stage, we learn an empirical estimate of $\mu_{W|a,x,z}$ using samples from $P_{AXZW}$. Based on the first-stage estimate $\widehat{\mu}_{W|a,x,z}$, we then estimate $\eta_{AXW}$ using samples from $P_{AXZY}$. The two-stage learning approach of KPV offers flexibility: we can estimate causal effects where samples from the full joint distribution of $\{(a,x,y,z,w)_i\}_{i=1}^n$ are not available, and instead one only has access to samples $\{(a,x,z,w)_i\}_{i=1}^{m_1}$ and $\{(\tilde{a},\tilde{x},\tilde{z},\tilde{y})_j\}_{j=1}^{m_2}$. The ridge regressions for these two stages are given in (5) and (6). The reader may refer to appendix B for a detailed derivation of the solutions.

**Stage 1.** From the first sample $\{(a,x,z,w)_i\}_{i=1}^{m_1}$, learn the conditional mean embedding of $\rho(W|a,x,z)$, i.e., $\widehat{\mu}_{W|a,x,z} := \widehat{C}_{W|A,X,Z}\left(\phi(a) \otimes \phi(x) \otimes \phi(z)\right)$ where $\widehat{C}_{W|A,X,Z}$ denotes the conditional mean embedding operator. We obtain $\widehat{C}_{W|A,X,Z}$ as a solution to:

$$\widehat{C}_{W|A,X,Z} = \underset{C \in \mathcal{H}_\Gamma}{\operatorname{argmin}}\ \widehat{E}(C), \text{ with} \tag{5}$$
$$\widehat{E}(C) = \tfrac{1}{m_1}\sum_{i=1}^{m_1} \|\phi(w_i) - C\phi(a_i,x_i,z_i)\|^2_{\mathcal{H}_{\mathcal{W}}} + \lambda_1\|C\|^2_{\mathcal{H}_\Gamma},$$

where $\mathcal{H}_\Gamma$ is the vector-valued RKHS of operators mapping $\mathcal{H}_{\mathcal{AXZ}}$ to $\mathcal{H}_{\mathcal{W}}$. It can be shown that $\widehat{C}_{W|A,X,Z} = \Phi(W)(\mathcal{K}_{AXZ} + m_1\lambda_1)^{-1}\Phi^T(A,X,Z)$ where $\mathcal{K}_{AXZ} = K_{AA} \odot K_{XX} \odot K_{ZZ}$ and $K_{AA}, K_{XX}$ and $K_{ZZ}$ are $m_1 \times m_1$ kernel matrices and $\Phi(W)$ is a vectors of $m_1$ columns, with $\phi(w_i)$ in its $i$th column (Song et al., 2009; Grünewälder et al., 2012; Singh et al., 2019). Consequently, $\widehat{\mu}_{W|a,x,z} = \Phi(W)(\mathcal{K}_{AXZ}+m_1\lambda_1)^{-1}\mathcal{K}_{axz}$ with $\mathcal{K}_{axz} = K_{Aa}\odot K_{Xx}\odot K_{Zz}$, where $K_{Aa}$ is a $m_1 \times 1$ vector denoting $k(a_s,a)$ evaluated at all $a_s$ in sample 1.

**Stage 2.** From the second sample $\{(\tilde{a},\tilde{x},\tilde{z},\tilde{y})_j\}_{j=1}^{m_2}$, learn $\hat{\eta}$ via empirical risk minimization (ERM):

$$\hat{\eta}_{AXW} = \underset{\eta \in \mathcal{H}_{\mathcal{AXW}}}{\operatorname{argmin}}\ \widehat{L}(\eta), \text{ where} \tag{6}$$
$$\widehat{L}(\eta) = \tfrac{1}{m_2}\sum_{j=1}^{m_2}(\tilde{y}_j - \eta[\phi(\tilde{a}_j,\tilde{x}_j) \otimes \widehat{\mu}_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j}])^2 + \lambda_2\|\eta\|^2_{\mathcal{H}_{\mathcal{AXW}}}.$$

where $\eta[\phi(\tilde{a}) \otimes \phi(\tilde{x}) \otimes \widehat{\mu}_{W|\tilde{a},\tilde{x},\tilde{z}}] = \langle \eta, \phi(\tilde{a}) \otimes \phi(\tilde{x}) \otimes \widehat{\mu}_{W|\tilde{a},\tilde{x},\tilde{z}} \rangle_{\mathcal{H}_{\mathcal{AXW}}}$ since $\eta \in \mathcal{H}_{\mathcal{AXW}}$. The estimator $\hat{\eta}_{AXW}$ given by (6) has a closed-form solution (Caponnetto & De Vito, 2007; Smale & Zhou, 2007).

**Theorem 1.** *For any $\lambda_2 > 0$, the solution of (6) exists, is unique, and is given by $\hat{\eta}_{AXW} = (\hat{\boldsymbol{T}}_2 + \lambda_2)^{-1}\hat{g}_2$ where*

$$\hat{\boldsymbol{T}}_2 = \tfrac{1}{m_2}\sum_{j=1}^{m_2}\left[\widehat{\mu}_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j} \otimes \phi(\tilde{a}_j,\tilde{x}_j)\right] \otimes \left[\widehat{\mu}_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j} \otimes \phi(\tilde{a}_j,\tilde{x}_j)\right]$$
$$\hat{g}_2 = \frac{1}{m_2}\sum_{j=1}^{m_2}\left[\widehat{\mu}_{W|\tilde{a}_j,\tilde{x}_j,\tilde{z}_j} \otimes \phi(\tilde{a}_j,\tilde{x}_j)\right]\tilde{y}_j.$$

**Remark 1.** *In the first stage, we learn the functional dependency of an outcome-induced proxy on the cause and an exposure induced proxy. Intuitively, one can interpret the set of $(a,x,z)$ (and not the individual elements of it) as the instrument set for $w$, with $\mu_{W|a,x,z}$ as a pseudo-structural function capturing the dependency between an instrument set and the target variable. It can be shown that for any fixed $a,x \in \mathcal{A}\times\mathcal{X}$, $\mu_{W|a,x,z}$ represents the dependency between $W$ and $Z$ due to the common confounder $U$ which is not explained away by $a$ and $x$. If $a = u$ for any $a,u \in \mathcal{A}\times\mathcal{U}$, $\rho(W|a,x,z) = \rho(W|a,x)$ and subsequently, $\mu_{W|a,x,z} = \mu_{W|a,x}$ (Miao & Tchetgen Tchetgen, 2018).*

Theorem 1 is the precise adaptation of Deaner (2018, eq. 3.3, 2021 version) to the case of infinite feature spaces, including the use of ridge regression in Stages 1 and 2, and the use of tensor product features. As we deal with infinite feature spaces, however, we cannot write our solution in terms of explicit feature maps, but we must express it in terms of feature inner products (kernels), following the form required by the representer theorem (see Proposition 2 below).

As an alternative solution to kernel proximal causal learning, one might consider using the Stage 1 estimate of $\widehat{\mu}_{W,A|A,X,Z}\left(\phi(a) \otimes \phi(x) \otimes \phi(z)\right) := \widehat{\mathbb{E}}\left(\phi(W) \otimes \phi(A)|a,x,z\right)$ as an input in Stage 2, which would allow an unmodified use of the KIV algorithm (Singh et al., 2019) in the proxy setting. Unfortunately regression from $\phi(a)$ to $\phi(a)$ is in population limit the identity mapping

$I_{\mathcal{H}_{\mathcal{A}}}$, which is not Hilbert-Schmidt for characteristic RKHS, violating the well-posedness assumption for consistency of Stage 1 regression (Singh et al., 2019). In addition, predicting $\phi(a)$ via ridge regression from $\phi(a)$ introduces bias in the finite sample setting, which may impact performance in the second stage (see Appendix B.7 for an example).

The KPV algorithm benefits from theoretical guarantees under well-established smoothness assumptions. The main assumptions involve well-posedness (i.e., minimizers belong to the RKHS search space) and source conditions on the integral operators for stage 1 and 2, namely Assumptions 12 to 15 in Appendix B. Specifically, $c_1$ and $c_2$ characterize the smoothness of the integral operator of Stage 1 and 2, respectively, while $b_2$ characterizes the eigenvalue decay of the Stage 2 operator.

**Theorem 2.** *Suppose Assumptions 5 to 7 and 12 to 15 hold. Fix $\zeta > 0$ and choose $\lambda_1 = m_1^{\frac{1}{c_1+1}}$ and $m_1 = m_2^{\frac{\zeta(c_1+1)}{(c_1-1)}}$.*

1. *If $\zeta \leq \frac{b_2(c_2+1)}{b_2c_2+1}$, choose $\lambda_2 = m_2^{-\frac{\zeta}{c_2+1}}$. Then $\tilde{R}(\widehat{\eta}_{AXW}) - \tilde{R}(\eta_{AXW}) = O_p\left(m_2^{-\frac{\zeta c_2}{c_2+1}}\right)$.*

2. *If $\zeta \geq \frac{b_2(c_2+1)}{b_2c_2+1}$, choose $\lambda_2 = m_2^{-\frac{b_2}{b_2c_2+1}}$. Then $\tilde{R}(\widehat{\eta}_{AXW}) - \tilde{R}(\eta_{AXW}) = O_p\left(m_2^{-\frac{b_2c_2}{b_2c_2+1}}\right)$.*

Our proof is adapted from Szabó et al. (2016); Singh et al. (2019) and is given in Appendix B. This leads to the following non-asymptotic guarantees for the estimate of the causal effect (2) at test time.

**Proposition 1.** *Consider a sample at test time $\{(x, w)_i\}_{i=1}^{n_t}$. Denote by $\widehat{\mu}_{XW} = \frac{1}{n_t}\sum_{i=1}^{n_t}[\phi(x_i) \otimes \phi(w_i)]$ the empirical mean embedding of $(X, W)$. For any $a \in \mathcal{A}$, the KPV estimator of the causal effect (2) is given by $\widehat{\beta}(a) = \widehat{\eta}_{AXW}[\widehat{\mu}_{XW} \otimes \phi(a)]$. Suppose the assumptions of Theorem 2 hold. The estimation error at test time is bounded by*

1. *If $\zeta \leq \frac{b_2(c_2+1)}{b_2c_2+1}$, $|\widehat{\beta}(a) - \beta(a)| \leq O_p(n_t^{-\frac{1}{2}} + m_2^{-\frac{\zeta c_2}{c_2+1}})$.*

2. *If $\zeta \geq \frac{b_2(c_2+1)}{b_2c_2+1}$, $|\widehat{\beta}(a) - \beta(a)| \leq O_p(n_t^{-\frac{1}{2}} + m_2^{-\frac{b_2c_2}{b_2c_2+1}})$.*

A proof of Proposition 1 is provided in Appendix B. Following Singh et al. (2020), this combines Theorem 2 with a probabilistic bound for the difference between $\widehat{\mu}_{XW}$ and its population counterpart.

From (4), a computable kernel solution follows from the representer theorem (Schölkopf et al., 2001),

$$\eta_{AXW}\left(\phi(\widetilde{a}_q, \widetilde{x}_q) \otimes \widehat{\mu}_{W|\widetilde{a},\widetilde{x},\widetilde{z}}\right) = \sum_{i,s=1}^{m_1}\sum_{j=1}^{m_2} \alpha_{ij} A_{ijsq}$$

with $A_{ijsq} = \{K_{w_i w_s}\left[\mathcal{K}_{AXZ} + m_1\lambda_1\right]^{-1}\mathcal{K}_{\widetilde{a}_q\widetilde{x}_q\widetilde{z}_q}\}\overline{\mathcal{K}}_{\widetilde{a}_q\widetilde{x}_q}$ where $\mathcal{K}_{\widetilde{a}_q\widetilde{x}_q\widetilde{z}_q} = K_{A\widetilde{a}_q} \odot K_{X\widetilde{x}_q} \odot K_{Z\widetilde{z}_q}$ and $\overline{\mathcal{K}}_{\widetilde{a}_q\widetilde{x}_q} = K_{\widetilde{a}_j\widetilde{a}_q} \odot K_{\widetilde{x}_j\widetilde{x}_q}$ (note that a correct solution requires $m_1 \times m_2$ coefficients, and cannot be expressed by $m_2$ coefficients alone; see Appendix B.3). Hence, the problem of learning $\widehat{\eta}_{AXW}$ amounts to learning $\widehat{\alpha} \in \mathbb{R}^{m_1\times m_2}$, for which classical ridge regression closed form solutions are available, i.e.,

$$\widehat{\alpha} = \underset{\alpha\in\mathbb{R}^{m_1\times m_2}}{\operatorname{argmin}} \frac{1}{m_2}\sum_{q=1}^{m_2}\left(\widetilde{y}_q - \sum_{i,s=1}^{m_1}\sum_{j=1}^{m_2}\alpha_{ij}A_{ijsq}\right)^2 + \lambda_2\sum_{i,r=1}^{m_1}\sum_{j,t=1}^{m_2}\alpha_{ij}\,\alpha_{rt}\,B_{ijrt}, \tag{7}$$

with $B_{ijrt} = K_{w_i w_r} K_{\widetilde{a}_j\widetilde{a}_t} K_{\widetilde{x}_j\widetilde{x}_t}$. Obtaining $\widehat{\alpha}$ involves inverting a matrix of dimension $m_1m_2 \times m_1m_2$, which has a complexity of $\mathcal{O}((m_1m_2)^3)$. Applying the Woodbury matrix identity on a vectorized version of (7), we get a cheaper closed-form solution for $\hat{\nu} = vec(\hat{\alpha})$ below.

**Proposition 2.** *Consider the optimization problem of (7). Let $\nu = vec(\alpha)$, and $\hat{\nu}$ the solution to the vectorized ERM in (7). We can show that for $\forall p, q \in \{1, \ldots, m_2\}$:*

$$\hat{\nu} = \left(\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})}\overline{\otimes}I_{m_2\times m_2}\right)(m_2\lambda_2 + \Sigma)^{-1}y \;\in \mathbb{R}^{m_1m_2},$$

$$\Sigma = \left(\Gamma^T_{(\widetilde{a}_q,\widetilde{x}_q,\widetilde{z}_q)}K_{WW}\Gamma_{(\widetilde{a}_p,\widetilde{x}_p,\widetilde{z}_p)}\right)(K_{\widetilde{a}_q\widetilde{a}_p}K_{\widetilde{x}_q\widetilde{x}_p}),$$

$$\Gamma_{(\widetilde{A},\widetilde{X},\widetilde{Z})} = (\mathcal{K}_{AXZ} + m_1\lambda_1)^{-1}(K_{A\widetilde{A}} \odot K_{X\widetilde{X}} \odot K_{Z\widetilde{Z}}),$$

*where $\overline{\otimes}$ represents tensor product of associated columns of matrices with the same number of columns.*

The details of the derivation are deferred to appendix B. Following these modifications, we only need to invert an $m_2 \times m_2$ matrix.

### 3.2 Proxy Maximum Moment Restriction (PMMR)

The PMMR relies on the following result, whose proof is given in Appendix C.

**Lemma 1.** *A measurable function $h$ on $\mathcal{A} \times \mathcal{W} \times \mathcal{X}$ is the solution to (1) if and only if it satisfies the conditional moment restriction (CMR) : $\mathbb{E}[Y - h(A, W, X) \,|\, A, Z, X] = 0$, $\mathbb{P}(A, Z, X)$-almost surely.*

iff $\mathbb{E}[(Y-h(A,W,X)]g(A,Z,X)] = 0$ for all measurable g

By virtue of Lemma 1, we can instead solve the integral equation (1) using tools developed to solve the CMR (Newey, 1993). By the law of iterated expectation, if $\mathbb{E}[Y - h(A, W, X)|A, Z, X] = 0$ holds almost surely, it implies that $\mathbb{E}[(Y - h(A, W, X))g(A, Z, X)] = 0$ for any measurable function $g$ on $\mathcal{A} \times \mathcal{X} \times \mathcal{Z}$. That is, the CMR gives rise to a continuum of conditions which $h$ must satisfy.

A maximum moment restriction (MMR) framework (Muandet et al., 2020a) requires that the moment restrictions hold presenter theorem !

intuition: we want to minimize the maximum value

*uniformly* over all functions $g$ that belong to a certain class of RKHS. Based on this framework, Zhang et al. (2020) showed, in the context of IV regression, that $h$ can be estimated consistently. In this section, we generalize the method proposed in Zhang et al. (2020) to the proxy setting by restricting the space of $g$ to a unit ball of the RKHS $\mathcal{H}_{\mathcal{AZX}}$ endowed with the kernel $k$ on $\mathcal{A} \times \mathcal{Z} \times \mathcal{X}$.

Before proceeding, we note that Miao & Tchetgen Tchetgen (2018) and Deaner (2018) also consider the CMR-based formulation in the proxy setting, but the techniques employed to solve it are different from ours. The MMR-IV algorithm of Zhang et al. (2020) also resembles other recent generalizations of GMM-based methods, notably, Bennett et al. (2019), Dikkala et al. (2020), and Liao et al. (2020); see Zhang et al. (2020, Sec. 5) for a detailed discussion.

**Objective.** To solve the CMR, the PMMR finds the function $h \in \mathcal{H}_{\mathcal{AXW}}$ that minimizes the MMR objective:

$$R_k(h) = \sup_{\substack{g \in \mathcal{H}_{\mathcal{AXZ}} \\ \|g\| \le 1}} \left(\mathbb{E}[(Y - h(A,W,X))g(A,Z,X)]\right)^2$$

Similarly to Zhang et al. (2020, Lemma 1), $R_k$ can be computed in closed form, as stated in the following lemma; see Appendix C for the proof.

**Lemma 2.** *Assume that*

$$\mathbb{E}[(Y - h(A,W,X))^2 k((A,Z,X),(A,Z,X))] < \infty$$

*and denote by $V'$ an independent copy of the random variable $V$. Then,* $R_k(h) = \mathbb{E}[(Y - h(A,W,X))(Y' - h(A',W',X'))k((A,Z,X),(A',Z',X'))]$.

Unlike Zhang et al. (2020), in this work the domains of $h$ and $g$ are not completely disjoint. That is, $(A,X)$ appears in both $h(A,W,X)$ and $g(A,Z,X)$. Next, we also require that $k$ is integrally strictly positive definite (ISPD).

**Assumption 11** The kernel $k : (\mathcal{A} \times \mathcal{Z} \times \mathcal{X})^2 \to \mathbb{R}$ is continuous, bounded, and is integrally strictly positive definite (ISPD), i.e., for any function $f$ that satisfies $0 < \|f\|_2^2 < \infty$, we have $\iint f(v)k(v,v')f(v')\,dvdv' > 0$ where $v := (a,z,x)$ and $v' := (a',z',x')$.

**A single-step solution.** Under Assumption 11, Zhang et al. (2020, Theorem 1) guarantees that the MMR objective preserves the consistency of the estimated $h$; $R_k(h) = 0$ if and only if $\mathbb{E}[Y - h(A,W,X)|A,Z,X] = 0$ almost surely. Hence, our Lemma 1 implies that any solution $h$ for which $R_k(h) = 0$ will also solve the integral equation (1).

Motivated by this result, we propose to learn $h$ by minimizing the empirical estimate of $R_k$ based on an i.i.d. sample $\{(a,z,x,w,y)_i\}_{i=1}^n$ from $P(A,Z,X,W,Y)$. By Lemma 2, this simply takes the form of a V-statistic (Serfling, 1980),

$$\boxed{\widehat{R}_V(h) := \frac{1}{n^2}\sum_{i,j=1}^{n}(y_i - h_i)(y_j - h_j)k_{ij},} \qquad (8)$$

allow a plug-in estimator.

where $k_{ij} := k((a_i,z_i,x_i),(a_j,z_j,x_j))$ and $h_i := h(a_i,w_i,x_i)$. When $i \neq j$, the samples $(a_i,w_i,x_i,y_i)$ and $(a_j,w_j,x_j,y_j)$ are indeed i.i.d. as required by Lemma 2. However, when $i = j$, the samples are dependent. Thus, $\widehat{R}_V$ is a biased estimator of $R_k$. The PMMR solution can then be obtained as a regularized solution to (8),

$$\hat{h}_\lambda = \operatorname*{argmin}_{h \in \mathcal{H}_{\mathcal{AXW}}} \widehat{R}_V(h) + \lambda\|h\|^2_{\mathcal{H}_{\mathcal{AXW}}}. \qquad (9)$$

Solution will be in terms of the kernel s.

Similarly to KPV, PMMR also comes with theoretical guarantees under a regularity assumption on $h$, characterized by a parameter $\gamma$, and a well-chosen regularization parameter.

**Theorem 3.** *Assume that Assumptions 6, 7 and 11 hold. If $n^{\frac{1}{2}-\frac{1}{2}\min\left(\frac{2}{\gamma+2},\frac{1}{2}\right)}$ is bounded away from zero and $\lambda = n^{-\frac{1}{2}\min\left(\frac{2}{\gamma+2},\frac{1}{2}\right)}$, then*

$$\underbrace{\|\hat{h}_\lambda - h\|}_{\text{consistency}}{}_{\mathcal{H}_{\mathcal{AXW}}} = O_p\left(n^{-\frac{1}{2}\min\left(\frac{2}{\gamma+2},\frac{1}{2}\right)}\right) \qquad (10)$$

This result is new, Zhang et al. (2020) did not provide a convergence rate for their solutions. The proof shares a similar structure with that of Theorem 2. There follows a bound for the estimate of the causal effect (2) at test time.

**Proposition 3.** *Consider a sample $\{(x,w)_i\}_{i=1}^{n_t}$ at test time. For any $a \in \mathcal{A}$, the PMMR estimator of the causal effect (2) is given by $\widehat{\beta}(a) = n^{-1}\sum_{i=1}^{n}\hat{h}(a,w_i,x_i)$. Suppose the assumptions of Theorem 3 hold. Then, the estimation error at test time is bounded by*

$$|\widehat{\beta}(a) - \beta(a)| \le O_p\left(n_t^{-\frac{1}{2}} + n^{-\frac{1}{2}\min\left(\frac{2}{\gamma+2},\frac{1}{2}\right)}\right),$$

where $\gamma$ can intuitively be thought of as a regularity parameter characterising the smoothness of $h$; we provide a formal characterisation in Appendix C Def. 4. Similar to Proposition 1, the proof of Proposition 3 combines Theorem 3 and a probabilistic bound on the difference between the empirical mean embedding of $(X,W)$ and its population version. The proofs are provided in Appendix C.

**Closed-form solution and comparison with MMR-IV.** Using the representer theorem (Schölkopf et al., 2001), we can express any solution of (9) as $\hat{h}(a,w,x) = \sum_{i=1}^{n}\alpha_i k((a_i,w_i,x_i),(a,w,x))$ for some $(\alpha_i)_{i=1}^n \in \mathbb{R}^n$. Substituting it back into (9) and solving for $\boldsymbol{\alpha} := (\alpha_1,\ldots,\alpha_n)$ yields $\boldsymbol{\alpha} = (LWL + \lambda L)^{-1}LW\boldsymbol{y}$ where $\boldsymbol{y} := (y_1,\ldots,y_n)^\top$ and $L, W$ are kernel matrices defined by $L_{ij} = k((a_i,w_i,x_i),(a_j,w_j,x_j))$ and $W_{ij} = k\left((a_i,z_i,x_i),(a_j,z_j,x_j)\right)$. Thus, PMMR has time complexity of $\mathcal{O}(n^3)$, which can be reduced via the usual Cholesky or Nyström techniques. Moreover, whilst we note that PMMR is similar to its predecessor, MMR-IV, we discuss their differences in Appendix C.1, Remark 5; specifically, we discuss an interpretation of the bridge function $h$, the difference between noise assumptions, and generalisation to a wider class of problems.

Question: why these kernels perform better than kernel ridge ?

just the same presenter theorem idea.

### 3.3 Connection Between the Two Approaches

From a non-parametric instrumental variable (NPIV) perspective, KPV and PMMR approaches are indeed similar to KIV (Singh et al., 2019) and MMR-IV (Zhang et al., 2020), respectively. We first clarify the connection between these two methods in the simpler setting of IV. In this setting, we aim to solve the Fredholm integral equation of the first kind:

$$\mathbb{E}[Y \mid z] = \int_{\mathcal{X}} f(x)\, \rho(x|z) dx, \quad \forall z \in \mathcal{Z}, \tag{11}$$

where, with an abuse of notation, $X$, $Y$, and $Z$ denote endogenous, outcome, and instrumental variables, respectively. A two-stage approach for IV proceeds as follows. In Stage 1, an estimate of the integral on the rhs of (11) is constructed. In Stage 2, the function $f$ is learned to minimize the discrepancy between the LHS and the estimated RHS of (11). Formally, this can be formulated via the unregularized risk

$$\begin{aligned} L(f) :=& \mathbb{E}_Z[(\mathbb{E}[Y|Z] - \mathbb{E}[f(X) \mid Z])^2] & (12)\\ \leq& \mathbb{E}_{YZ}[(Y - \mathbb{E}[f(X) \mid Z])^2] =: \tilde{L}(f) & (13)\end{aligned}$$

The risk (13) is considered in DeepIV (Hartford et al., 2017), KernelIV (Singh et al., 2019), and DualIV (Muandet et al., 2020b). Both DeepIV and KernelIV directly solve the surrogate loss (13), whereas DualIV solves the dual form of (13). Instead of (11), an alternative starting point to NPIV is the conditional moment restriction (CMR) $\mathbb{E}[\varepsilon \mid Z] = \mathbb{E}[Y - f(X) \mid Z] = 0$. Muandet et al. (2020a) showed that a RKHS for $h$ is sufficient in the sense that the inner maximum moment restriction (MMR) preserves all information about the original CMR. Although starting from the CMR perspective, Zhang et al. (2020) used the MMR in the IV setting to minimise loss as (12), as shown in Liao et al. (2020, Appendix F).

**The connection.** Both DeepIV (Hartford et al., 2017) and KernelIV (Singh et al., 2019) (resp. KPV) solve an objective function that is an upper bound of the objective function of MMR (Zhang et al., 2020) (resp. PMMR) and other GMM-based methods such as AGMM (Dikkala et al., 2020) and DeepGMM (Bennett et al., 2019). This observation reveals the close connection between modern nonlinear methods for NPIV, as well as between our KPV and PMMR algorithms, which minimize $R$ and $\tilde{R}$, respectively, where

$$\begin{aligned} R(h) &= \mathbb{E}_{AXZ}[(\mathbb{E}[Y|A,X,Z] - \mathbb{E}[h(A,X,W) \mid A,X,Z])^2] & (14)\\ &\leq \mathbb{E}_{AXZY}[(Y - \mathbb{E}[h(A,X,W) \mid A,X,Z])^2] = \tilde{R}(h) & (15)\end{aligned}$$

We formalize this connection in the following result.

**Proposition 4.** *Assume there exists $h \in L^2_{P_{AXW}}$ such that $\mathbb{E}[Y|A,X,Z] = \mathbb{E}[h(A,X,W)|A,X,Z]$. Then, (i) $h$ is a minimizer of $R$ and $\tilde{R}$. (ii) Assume $k$ satisfies Assumption 11. Then, $R_k(h) = 0$ iff $R(h) = 0$. (iii) Assume $k$ satisfies Assumption 7. Suppose that $\mathbb{E}[f(W)|A,X,Z = \cdot] \in \mathcal{H}_{\mathcal{AXZ}}$ for any $f \in \mathcal{H}_{\mathcal{W}}$ and that Assumption 9 holds. Then, $h$ is given by the KPV solution, and is a minimizer of $R$ and $\tilde{R}$.*

See Appendix E for the proof of Proposition 4. The assumptions needed for the third result ensures that the conditional mean embedding operator is well-defined and characterizes the full distribution $P_{A,X,W|A,X,Z}$, and that the problem is well-posed.

**Remark 2.** *The KPV and PMMR approaches minimise risk in different ways, and hence they offer two different representations of $h$. KPV minimises $\tilde{R}$ by first estimating the empirical conditional mean embedding $\widehat{\mu}_{W|a_i,z_i,x_i}$ from a sub-sample $\{(a,w,x)_i\}_{i=1}^{m_1}$, and then estimating $h$ by minimising $\frac{1}{m_2}\sum_{j=1}^{m_2}(y_j - \mathbb{E}_{W|a_j,x_j,z_j}(h))^2$ over a sub-sample $\{(a,z,x)_j\}_{j=1}^{m_2}$ using $\widehat{\mu}_{W|a_i,z_i,x_i}$ obtained from the first stage. Hence, $h^{KPV}(a,w,x) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} \alpha_{ij} k(a_j,a)k(x_j,x)k(w_i,w)$ for some $(\alpha_{i,j})_{i,j=1}^{m_1,m_2} \in \mathbb{R}^{m_1\times m_2}$. In contrast, PMMR directly minimises $R_k$, resulting in the estimator $\hat{h}^{PMMR}(a,w,x) = \sum_{i=1}^{n}\alpha_i k((a_i,w_i,x_i),(a,w,x))$ for some $(\alpha_i)_{i=1}^{n} \in \mathbb{R}^n$, and over the joint distribution of $\{(a,x,w,z)_i\}_{i=1}^{n}$.*

## 4 Experiments

We evaluate KPV and PMMR against the following baselines: (1) *KernelRidge*: kernel ridge regression $Y \sim A$. (2) *KernelRidge-W*: kernel ridge regression $Y \sim (A,W)$, adjusted over $W$, i.e., $\int \mathbb{E}[Y|A,W]d\rho(W)$. (3)*KernelRidge-W,Z*: kernel ridge regression $Y \sim (A,W,Z)$, adjusted over $W$ and $Z$, i.e., $\int \mathbb{E}[Y|A,W,Z]d\rho(W,Z)$. Methods (2) and (3) are implemented in accordance with (Singh et al., 2020). (4) *Deaner18*: a two-stage method with a finite feature dictionary (Deaner, 2018), and (5) *Tchetgen-Tchetgen20*: a linear two-stage method consistent consistent with Miao & Tchetgen Tchetgen (2018). The mean absolute error with respect to the true causal effect $\mathbb{E}[Y|do(A)]$, *c-MAE*, is our performance metric. Without loss of generality, we assume that $X$ is a null set. Both KPV and PMMR are capable of handling non-null sets of covariates, $X$. Our codes are publicly available at: https://github.com/yuchen-zhu/kernel_proxies.

The satisfaction of Assumptions 1- 4 will guarantee identification of $h$ on the support of data, but we still need $(A,W)$ to satisfy an overlap/positivity assumption (similar to classic positivity in causal inference, e.g. Westreich & Cole (2010)) to guarantee empirical identification of $\mathbb{E}[Y|do(a)]$. As we infer the causal effect from $h$ by $\int \hat{h}(a,w)d\rho(w)$, we need $\rho(w|a) > 0$ whenever $\rho(w) > 0$ for $h$ to be well-identified for the marginal support of $W$.

### 4.1 Hyperparameter Tuning

For both KPV and PMMR, we employ an exponentiated quadratic kernel for continuous variables, as it is continu-

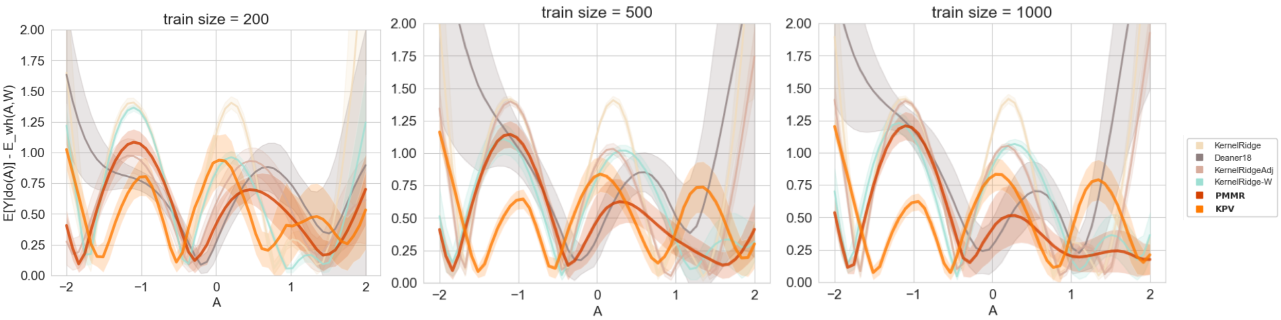


Figure 2: Mean absolute error of ATE estimation for each A of (16). The lower the mean absolute error the stronger is the performance of the model. Tchetgen-Tchetgen20 has error out of the range so we omit it for clearer plots.

ous, bounded, and characteristic, thus meeting all required assumptions. KPV uses a leave-one-out approach with grid search for the kernel bandwidth parameters in both stages and the regularisation parameters (See appendix D.2.1). PMMR uses a leave-M-out with gradient descent approach to tune for the $L-$bandwidth parameters and the regularisation parameter $\lambda$, where the bandwidth parameters are initialised using the median distance heuristic and the regularisation parameters are initialised randomly. The approach is based on Zhang et al. (2020). We defer the hyperparameter selection details to Appendix D.

## 4.2 Synthetic Experiment

First, we demonstrate the performance of our methods on a synthetic simulation with non-linear treatment and outcome response functions. In our generative process, $U, W$ and $Z \in \mathbb{R}^2$ and $A$ and $Y$ are scalar. We have defined the latent confounder $U$ such that $U_1$ is dependent on $U_2$. Appendix D fig. 7 demonstrates the relationship between $U_1$ and $U_2$. Given $U_2$, we know the range of $U_1$, but the reverse does not hold: knowning $U_1 \in [0, 1]$, then $U_2$ is with equal probability in one of the intervals $[-1, 0]$ or $[1, 2]$. In design of the experiment, we also have chosen $W$ such that its first dimension is highly correlated with $U_1$ (less informative dimension of $U$) with small uniform noise, and its second dimension is a view of $U_2$ with high noise. With this design, it is guaranteed that $\{W\}$ is not a sufficient proxy set for $U$. See eq. (16) for details.

$$
\begin{aligned}
U &:= [U_1, U_2], \quad U_2 \sim \text{Uniform}[-1, 2] \\
U_1 &\sim \text{Uniform}[0, 1] - \mathbb{1}[0 \leq U_2 \leq 1] \\
W &:= [W_1, W_2] = [U_1 + \text{Uniform}[-1, 1], U_2 + \mathcal{N}(0, \sigma^2)] \\
Z &:= [Z_1, Z_2] = [U_1 + \mathcal{N}(0, \sigma^2), U_2 + \text{Uniform}[-1, 1]] \\
A &:= U_2 + \mathcal{N}(0, \beta^2) \\
Y &:= U_2 \cos(2(A + 0.3U_1 + 0.2))
\end{aligned}
\tag{16}
$$

where $\sigma^2 = 3$ and $\beta^2 = 0.05$.

We use training sets of size 500 and 1000, and average results over 20 seeds affecting the data generation. The generative distribution is presented in Appendix D, fig. 7.

| Method | c-MAE(n=200) | c-MAE(n=500) | c-MAE(n=1000) |
|---|---|---|---|
| **KPV** | **0.499 ± 0.310** | **0.490 ± 0.285** | **0.491 ± 0.290** |
| **PMMR** | **0.533 ± 0.314** | **0.494 ± 0.330** | **0.472 ± 0.358** |
| KernelRidgeAdj | 0.569 ± 0.317 | 0.577 ± 0.352 | 0.607 ± 0.379 |
| KernelRidge-W | 0.635 ± 0.428 | 0.695 ± 0.460 | 0.716 ± 0.476 |
| KernelRidge | 0.840 ± 0.782 | 0.860 ± 0.709 | 0.852 ± 0.654 |
| Deaner18 | 0.681 ± 0.477 | 1.030 ± 1.020 | 1.050 ± 0.867 |
| Tchetgen-Tchetgen20 | 1.210 ± 1.070 | 17.60 ± 85.50 | 1.100 ± 1.460 |

Table 1: Results comparing our methods to baselines on the simulation studies described in eq. (16)

Table 1 summarizes the results of our experiment with the synthetic data. Both KPV and PMMR, as methodologies designed to estimate unbiased causal effect in proximal setting, outperform other methods by a large margin, and have a narrow variance around the results. As expected, the backdoor adjustment for $(W, Z)$, the current common practice to deal with latent confounders (without considering the nuances of the proximal setting), does not suffice to unconfound the causal effect. Related methods, *KernelRidge-(W,Z)* and *KernelRidge-W*, underperform our methods by large margins. As fig. 2 shows, they particulary fail to identify the functional form of the causal effect. *Tchetgen-Tchetgen20* imposes a strong linearity assumption, which is not suitable in this nonlinear case, hence its bad performance. The underperformance of *Deaner18* is largely related to it only using a finite dictionary of features, whereas the kernel methods use an infinite dictionary.

## 4.3 Case studies

In the next two experiments, our aim is to study the performance of our approaches in dealing with real world data. To have a real causal effect for comparison, we fit a generative model to the data, and evaluate against simulations from the model. See D for further discussion and for the full procedure. Consequently, we refrain from making any policy recommendation on the basis of our results. In both experiments, we sample a training set of size 1500, and average results over 10 seeds affecting the data generation.

### 4.3.1 Legalized abortion and crime

We study the data from Donohue & Levitt (2001) on the impact of legalized abortion on crime. We follow the data preprocessing steps from Woody et al. (2020), removing the

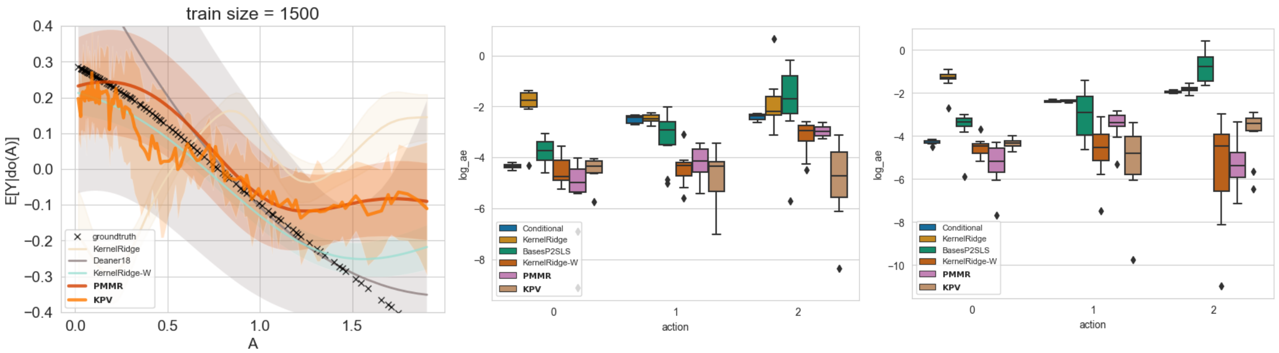


Figure 3: (Right: Abortion and Criminality), ATE comparison of adjustment on W with ground truth and direct regression (Middle: Maths, Left: Reading) Grade retention and cognitive outcome; log MAE $\log|\beta(a) - \hat{\beta}(a)|$ over action values $a = 0, 1, 2$; lower is better.

| **Metric** | **c-MAE** | | |
|---|---|---|---|
| **Method/Dataset** | Abort. & Crim. | Grd Ret., Maths | Grd. Ret., Reading |
| KPV | $0.129 \pm 0.105$ | $\mathbf{0.023 \pm 0.020}$ | $0.027 \pm 0.022$ |
| PMMR | $0.137 \pm 0.101$ | $0.032 \pm 0.022$ | $\mathbf{0.023 \pm 0.022}$ |
| Conditional | - | $0.062 \pm 0.036$ | $0.083 \pm 0.053$ |
| KernelRidge | $0.330 \pm 0.186$ | $0.200 \pm 0.631$ | $0.190 \pm 0.308$ |
| KernelRidge-W | $\mathbf{0.056 \pm 0.053}$ | $0.031 \pm 0.026$ | $0.024 \pm 0.021$ |
| Deaner18 | $0.369 \pm 0.284$ | $0.137 \pm 0.223$ | $0.240 \pm 0.383$ |

Table 2: Results comparing our methods to baselines on the real-world examples described in 4.3.

state and time variables. We choose the effective abortion rate as treatment ($A$), murder rate as outcome ($Y$), "generosity to aid families with dependent children" as treatment-inducing proxy ($Z$), and beer consumption per capita, log-prisoner population per capita and concealed weapons law as outcome-inducing proxies ($W$). We collect the rest of the variables as the unobserved confounding variables ($U$).

**Results.** Table 2 includes all results. Both KPV and PMMR beat KernelRidge and BasesP2SLS by a large margin, highlighting the advantage of our methods in terms of deconfounding and function space flexibility. KernelRidge-W is the best method overall, beating the second best by a wide margin. We find this result curious, as Figure 3 shows that adjustment over $W$ is sufficient for identifying the causal effect in this case, however it is not obvious how to conclude this from the simulation. We leave as future work the investigation of conditions under which proxies provide a sufficient adjustment on their own.

### 4.3.2 Grade retention and cognitive outcome

We use our methods to study the effect of grade retention on long-term cognitive outcome using data the ECLS-K panel study (Deaner, 2018). We take cognitive test scores in Maths and Reading at age 11 as outcome variables ($Y$), modelling each outcome separately, and grade retention as the treatment variable ($A$). Similar to Deaner (2018), we take the average of 1st/2nd and 3rd/4th year elementary scores as the treatment-inducing proxy ($Z$), and the cognitive test scores from Kindergarten as the outcome-inducing proxy ($W$). See Appendix D for discussion on data.

**Results.** Results are in Table 2. For both Math grade retention and Reading grade retention, our proposed methods outperform alternatives: KPV does better on the Math outcome prediction, while PMMR exceeds others in estimation for the Reading outcome. KernelRidge-W is still a strong contender, but all other baselines result in large errors.

## 5 Conclusion

In this paper, we have provided two kernel-based methods to estimate the causal effect in a setting where proxies for the latent confounder are observed. Previous studies mostly focused on characterising identifiability conditions for the proximal causal setting, but lack of methods for estimation was a barrier to wider implementation. Our work is primarily focused on providing two complementary approaches for causal effect estimation in this setting. This will hopefully motivate further studies in the area.

Despite promising empirical results, the hyperparameter selection procedure for both methods can be improved. For KPV, the hyperparameter tuning procedure relies on the assumption that optimal hyperparameters in the first and second stage can be obtained independently, while they are in fact interdependent. For PMMR, there is no systematic way of tuning the hyperparameter of the kernel $k$ that defines the PMMR objective, apart from the median heuristic. Developing a complete hyperparameter tuning procedure for both approaches is an important future research direction. Beyond this, both methods can be employed to estimate causal effect in wider set of problems, where the Average Treatment on the Treated, or Conditional Average Treatment Effect are the quantity of interests.

## Acknowledgments

RS was partially funded by a ONR grant, award number N62909-19-1-2096; YZ was funded by EPSRC with grant number EP/S021566/1.

## References

Baker, C. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186: 273–289, 1973.

Bennett, A., Kallus, N., and Schnabel, T. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems 32*, pp. 3564–3574. Curran Associates, Inc., 2019.

Cai, Z. and Kuroki, M. On identifying total effects in the presence of latent variables and selection bias. *arXiv preprint arXiv:1206.3239*, 2012.

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

Carrasco, M., Florens, J.-P., and Renault, E. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. In Heckman, J. and Leamer, E. (eds.), *Handbook of Econometrics*, volume 6B, chapter 77. Elsevier, 1 edition, 2007.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

Choi, H. K., Hernán, M. A., Seeger, J. D., Robins, J. M., and Wolfe, F. Methotrexate and mortality in patients with rheumatoid arthritis: a prospective study. *The Lancet*, 359(9313):1173–1177, 2002.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., et al. The effectiveness of right heart catheterization in the initial care of critically iii patients. *Jama*, 276(11):889–897, 1996.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22 (1):173–203, 1959.

Deaner, B. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2018.

Dikkala, N., Lewis, G., Mackey, L., and Syrgkanis, V. Minimax estimation of conditional moment models. *CoRR*, abs/2006.07201, 2020.

Donohue, John J., I. and Levitt, S. D. The Impact of Legalized Abortion on Crime*. *The Quarterly Journal of Economics*, 116(2):379–420, 05 2001. ISSN 0033-5533. doi: 10.1162/00335530151144050. URL https://doi.org/10.1162/00335530151144050.

Flanders, W. D., Strickland, M. J., and Klein, M. A new method for partial correction of residual confounding in time-series and other observational studies. *American journal of epidemiology*, 185(10):941–949, 2017.

Fruehwirth, J. C., Navarro, S., and Takahashi, Y. How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects. *Journal of Labor Economics*, 34(4):979–1021, 2016.

Fukumizu, K., Bach, F., and Jordan, M. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

Fukumizu, K., Gretton, A., and Bach, F. Statistical convergence of kernel cca. In *Advances in Neural Information Processing Systems*, volume 18, pp. 387–394. MIT Press, 2006.

Greenland, S. and Lash, T. L. Bias analysis. *International Encyclopedia of Statistical Science*, 2:145–148, 2011.

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. Conditional mean embeddings as regressors. *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Grunewalder, S., Lever, G., Baldassarre, L., Pontil, M., and Gretton, A. Modelling transition dynamics in mdps with rkhs embeddings. *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1414–1423. PMLR, 2017.

Imbens, G. W. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Kuang, Z., Sala, F., Sohoni, N., Wu, S., Córdova-Palomera, A., Dunnmon, J., Priest, J., and Ré, C. Ivy: Instrumental variable synthesis for causal inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 398–410. PMLR, 2020.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Liao, L., Chen, Y., Yang, Z., Dai, B., Wang, Z., and Kolar, M. Provably efficient neural estimation of structural equation model: An adversarial approach. In *Advances in Neural Information Processing Systems 33*. 2020.

Macedo, H. D. and Oliveira, J. N. Typing linear algebra: A biproduct-oriented approach, 2013.

Mann, H. B. and Wald, A. On stochastic limit and order relationships. *Ann. Math. Statist.*, 14(3):217–226, 09 1943. doi: 10.1214/aoms/1177731415.

Miao, W. and Tchetgen Tchetgen, E. A confounding bridge approach for double negative control inference on causal effects (supplement and sample codes are included). *arXiv preprint arXiv:1808.04945*, 2018.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

Muandet, K., Jitkrittum, W., and Kübler, J. Kernel conditional moment test via maximum moment restriction. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 41–50. PMLR, 2020a.

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. Dual instrumental variable regression. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc., 2020b.

Nashed, M. Z. and Wahba, G. Convergence rates of approximate least squares solutions of linear integral and operator equations of the first kind. *Mathematics of Computation*, 28(125):69–80, 1974.

Newey, W. 16 efficient estimation of models with conditional moment restrictions. *Handbook of Statistics*, 11: 419–454, 1993.

Newey, W. K. and McFadden, D. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pp. 2111 – 2245. Elsevier, 1994. doi: https://doi.org/10.1016/S1573-4412(05)80005-4.

Pearl, J. *Causality*. Cambridge university press, 2000.

Petersen, K. B. and Pedersen, M. S. The matrix cookbook, 2008. URL http://www2.imm.dtu.dk/pubdb/p.php?3274. Version 20081110.

Reiersøl, O. *Confluence analysis by means of instrumental sets of variables*. PhD thesis, Almqvist & Wiksell, 1945.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In Helmbold, D. and Williamson, B. (eds.), *Computational Learning Theory*, pp. 416–426, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

Schuemie, M. J., Ryan, P. B., DuMouchel, W., Suchard, M. A., and Madigan, D. Interpreting observational studies: why empirical calibration is needed to correct p-values. *Statistics in medicine*, 33(2):209–218, 2014.

Sejdinovic, D. and Gretton, A. What is an rkhs? 2014. URL http://www.stats.ox.ac.uk/~sejdinov/teaching/atml14/Theory_2014.pdf.

Serfling, R. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

Shi, X., Miao, W., Nelson, J. C., and Tchetgen Tchetgen, E. J. Multiply robust causal inference with double negative control adjustment for categorical unmeasured confounding. *arXiv preprint arXiv:1808.04906*, 2018.

Singh, R. Kernel methods for unobserved confounding: Negative controls, proxies, and instruments. *arXiv preprint arXiv:2012.10315*, 2020.

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, pp. 4595–4607, 2019.

Singh, R., Xu, L., and Gretton, A. Reproducing kernel methods for nonparametric and semiparametric treatment effects. *arXiv preprint arXiv:2010.04855*, 2020.

Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 08 2007. doi: 10.1007/s00365-006-0659-y.

Smola, A. J., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, pp. 13–31. Springer-Verlag, 2007.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen Tchetgen, E. J. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31(3):348, 2016.

Song, L., Huang, J., Smola, A., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 961–968, 2009.

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

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (7), 2011.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Sutherland, D. J. Fixing an error in caponnetto and de vito (2007). *arXiv preprint arXiv:1702.02982*, 2017.

Szabó, Z., Gretton, A., Póczos, B., and Sriperumbudur, B. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, pp. 948–957. PMLR, 2015.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

Tchetgen Tchetgen, E. The control outcome calibration approach for causal inference with unobserved confounding. *American journal of epidemiology*, 179(5):633–640, 2014.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. Minimax estimation of kernel mean embeddings. *J. Mach. Learn. Res.*, 18(1):3002–3048, January 2017. ISSN 1532-4435.

Van der Vaart, A. *Asymptotic Statistics*. Cambridge University Press, 2000.

Westreich, D. and Cole, S. R. Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, 171 (6):674–677, 02 2010. ISSN 0002-9262. doi: 10.1093/aje/kwp436. URL https://doi.org/10.1093/aje/kwp436.

Woody, S., Carvalho, C. M., Hahn, P., and Murray, J. S. Estimating heterogeneous effects of continuous exposures using bayesian tree ensembles: revisiting the impact of abortion rates on crime. *arXiv: Applications*, 2020.

Zhang, R., Imaizumi, M., Schölkopf, B., and Muandet, K. Maximum moment restriction for instrumental variable regression. *arXiv preprint arXiv:2010.07684*, 2020.