Learning Nonlinear Causal Effects via Kernel Anchor Regression

Wenqi Shi[1] **Wenkai Xu**[2]

[1]Department of Industrial Engineering,
Tsinghua University
[2]Department of Statistics,
University of Oxford

Abstract

Learning causal effects is a fundamental problem in science. Anchor regression has been developed to address this problem for a large class of causal graphical models, though the relationships between the variables are assumed to be linear. In this work, we tackle the nonlinear setting by proposing kernel anchor regression (KAR). Beyond a classic two-stage least square (2SLS) estimator, we also study an improved variant that involves nonparametric kernel regression in three separate stages. We provide convergence results for the proposed KAR estimators and the identifiability conditions for KAR to learn the nonlinear structural equation models (SEM). Experimental results demonstrate the superior performances of the proposed KAR estimators over existing baselines.

1 INTRODUCTION

Understanding the causal effects can be a key ingredient in many scientific studies. For instance, medical practitioners need to know how effective a treatment is to the target disease in clinical trials; econometricians ask how much change a particular purchasing behaviour drives the Consumer Price Index (CPI); epidemiologists want to understand to what extent a government policy can alleviate the pandemic. While the goal of revealing causal effects remains the same, the focus in the notion of causality may differ due to specific applications. To describe different aspects of the causal notions and design corresponding statistical procedures for inferring the causal effects, various frameworks have been developed including the potential outcomes framework [Rubin, 2004, 2005], counterfactual distributions [Chernozhukov et al., 2013] and Pearl's causal graphical models [Pearl et al., 2000, 2016]. A succinct yet comprehensive introduction can be found in Peters et al. [2017].

Causality has also been an emerging field in machine learning community and machine learning techniques have been studied to improve the procedures for learning the causal effects. In particular, independence [Gretton et al., 2005] and conditional independence [Fukumizu et al., 2007] measures have been exploited to infer causal graphical models [Colombo et al., 2012, Mooij et al., 2009], especially for the additive noise setting [Hoyer et al., 2008, Peters et al., 2014]. Independent Component Analysis (ICA) [Hyvärinen, 2013, Hyvarinen and Morioka, 2017] has been employed to identify causal relationships in both linear [Hyvärinen et al., 2010, Shimizu et al., 2006, 2011] and non-linear settings [Monti et al., 2020, Khemakhem et al., 2021]. Score matching [Hyvärinen and Dayan, 2005] has also been considered [Rolland et al., 2022] for non-linear causal learning. Moreover, kernel methods, that utilize rich representation of reproducing kernel Hilbert space (RKHS), have been applied to tackle nonparametric estimation [Muandet et al., 2021, Singh et al., 2019] and regression [Singh et al., 2019, Zhu et al., 2022] problems with causal implications. Deep neural networks have also been attempted for learning treatment effect [Johansson et al., 2020, Kallus, 2020, Louizos et al., 2017] or useful causal representations [Besserve et al., 2019, Schölkopf et al., 2021, Xu et al., 2020, 2021].

Recently, an elegant and statistically robust approach formulates causality as an invariant risk minimization (IRM), see for example [Bühlmann, 2018, Peters et al., 2016]. The causal structure is thought to be invariant across the environment and robust under intervention. The IRM learning procedure [Arjovsky et al., 2019] on the observational data is then formulated as a regularized empirical risk minimization (ERM) to achieve both in-distribution performance and out-of-distribution generalization. In particular, anchor regression [Rothenhäusler et al., 2018], closely related to K-class estimators [Jakobsen and Peters, 2022], has been developed under the IRM framework to tackle a very general class of causal graphical models with the confounders being partly (but not fully) observed. By choosing different regularization parameter, anchor regression is able to unify

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

the ordinary least square (OLS) regression, partialling out (PA) regression, and instrumental variable (IV) regression. While existing works mostly considered linear cases [Oberst et al., 2021, Rothenhäusler et al., 2018], we explore the non-linear setting for anchor regression [Kook et al., 2022]. Specifically, we consider the nonparametric estimation to tackle non-linear features via RKHS functions. Although nonlinear anchor regression may not perform well in terms of generalization [Christiansen et al., 2021], we show that the approach is valuable as it can identify nonlinear causal effects under certain conditions when confounders are only partially observed, and in certain setting, it can outperform other nonlinear methods in terms of MSE.

This paper is structured as follows. In Section 2, we review useful concepts including instrumental variable (IV), anchor regression (AR), and reproducing kernel Hilbert space (RKHS). Then we develop two versions of kernel anchor regression (KAR) estimators in Section 3. Theoretical analysis on the estimators and the causal interpretation with nonlinear SEM are provided in Section 4. Experimental results for synthetic data and real-world applications are shown in Section 5 followed by concluding discussion and future directions in Section 6. Code for the experiments is available at `https://github.com/Swq118/Kernel-Anchor-Regression`.

2 BACKGROUND

Directed Acyclic Graph (DAG) is a power class of graphical model for characterising conditional dependency structures and has been widely used for probabilistic modelling such as hidden Markov models [Rabiner and Juang, 1986], latent variable models [Bishop, 1998] and topic models [Blei, 2012]. By enforcing certain Markov and faithfulness assumptions [Peters et al., 2011], as well as noise structures [Hoyer et al., 2008], DAG models the causal relationships [Glymour et al., 2019, Spirtes et al., 2013] and the learning procedures have been developed [Colombo et al., 2012, Spirtes et al., 2000, Zhang et al., 2018].

From Instrumental Variable to Anchor Regression Instrumental variable (IV) has been developed to incorporate endogenous explanatory variables in econometrics [Bowden and Turkington, 1990] and then applied for estimating causal effect [Angrist et al., 1996]. Consider the linear regression problem $Y = X\beta + \epsilon$. OLS assumes independence between noise ϵ and explanatory X (the exogenous variable) and β is estimated via minimizing

$$\beta^{OLS} = \arg\min_{\beta} \mathbb{E}_{train}[\|Y - X\beta\|^2]. \tag{1}$$

The IV setting assumes explicit dependency between X and ϵ via instrumental variable Z, i.e. $X = Z\theta + \varepsilon$ where $Z \perp \varepsilon$. The two-stage least squares (2SLS) procedure, widely used in economics, tackles the linear IV estimation by first regressing Z over X to get conditional means $\bar{X}(z) := \mathbb{E}[X|Z = z]$ and secondly regressing outputs Y on these conditional means[1]. This corresponds to minimizing the projected least square objective,

$$\beta^{IV} = \arg\min_{\beta} \mathbb{E}_{train}[\|P_Z(Y - X\beta)\|^2]. \tag{2}$$

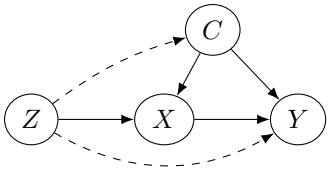

Figure 1: DAG representations for IV regression (solid lines only) and anchor regression (with dashed lines).

Let P_Z denote the projection to Z where $P_{Z=z}(X) = \mathbb{E}[X|Z = z] = \bar{X}(z)$. 2SLS works well when the underlying assumptions hold. The corresponding DAG is shown in Figure 1 with only solid lines. In practice, the relation between Y and X may not be linear, nor may be the relation between X and Z. Non-parametric IV has been explored through moment estimations [Dikkala et al., 2020] as well as using deep neural networks [Bennett et al., 2019, Centorrino et al., 2019, Hartford et al., 2017, Singh et al., 2019, Xu et al., 2020, Zhu et al., 2022].

However, Y's dependency on Z may not be solely through X, i.e. as the dashed lines from Z to Y in Figure 1 indicate, Y may depend on Z directly, and the strength of such dependency may remain unknown. The latent confounder C may not be independent of Z, as indicated by dashed line from Z to C in Figure 1. Incorporating such dependency structures tackles a much more general class of DAG, e.g. IV is a special case. To estimate β, anchor regression has been proposed [Rothenhäusler et al., 2018] to effectively combines Equation (1) and Equation (2). For regularization parameter γ and identity operator $Id(Z) := Z$,

$$\beta^{\gamma} = \arg\min_{\beta} \mathbb{E}_{train}[\|(Id - P_Z)(Y - X\beta)\|^2] \tag{3}$$

$$+\gamma\mathbb{E}_{train}[\|P_Z(Y - X\beta)\|^2]. \tag{4}$$

Here, $\gamma \geq 0$ can be thought of the level of direct dependency of Y on Z variable[2]. By setting different γ values, anchor regression recovers classical settings, i.e. $\gamma = 1$ corresponds to OLS, $\beta^1 = \beta^{OLS}$; $\gamma \to \infty$ corresponds to IV, $\beta^{\to\infty} := \lim_{\gamma\to\infty} \beta^{\gamma} = \beta^{IV}$; $\gamma = 0$ corresponds to the "partialling out" setting where only residuals between regression of Z to X and Y are of interest.

[1] Writing $Y = X\beta + \epsilon = (Z\theta)\beta + (\varepsilon\beta + \epsilon)$ where $Z\theta = \mathbb{E}[X|Z]$. the regressor is independent of noise and OLS apply.

[2] The smaller γ value dashed line, the stronger the dependency, i.e. the more solid dashed line from Z to Y.

Kernel-based Methods Functions in RKHS has been employed to tackle various statistical and machine learning tasks with nonlinear features [Hofmann et al., 2008], e.g. kernel ridge regression, support vector machine, etc. RKHS functions have also been utilized to represent and characterize distributions, via kernel mean embedding [Muandet et al., 2017]. For probability measure p and kernel k associated with RKHS $\mathcal{H}$, the mean embedding denoted by $\mu_p := \int k(x, \cdot)dp(x) \in \mathcal{H}$ has been widely used to compare distributions, e.g. via maximum-mean-discrepancy (MMD) [Gretton et al., 2012]. Conditional mean embedding [Song et al., 2009] has also been considered for learning in regression problems [Fukumizu et al., 2007, Grünewälder et al., 2012]. With the rich representations, RKHS functions are also applicable of learning distribution directly via distribution regression [Szabó et al., 2015, 2016].

3 KERNEL ANCHOR REGRESSION

To capture the non-linear features in the DAG, we kernelize the anchor regression framework by utilizing the rich feature representation of RKHS functions. The kernelizing procedure is inspired from kernel instrumental variable (KIV) [Singh et al., 2019] where the operators are learned for conditional mean embedding in two separate regression stages. The DAG representation is illustrated in Figure 2[3]. In our setting, Z is observable covariates called anchor which may or may not have effects on target X or Y. All unobservable latent confounders are denoted by C.

Let $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be measurable positive definite kernels corresponding to RKHS $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Z}}$. Denote the feature maps $\psi : \mathcal{X} \to \mathcal{H}_{\mathcal{X}}, x \to k_{\mathcal{X}}(x, \cdot)$ and $\phi : \mathcal{Z} \to \mathcal{H}_{\mathcal{Z}}, z \to k_{\mathcal{Z}}(z, \cdot)$. Let $P_{\phi(Z)}$ and Id denote the L_2-projection on the linear span from the components of $\phi(Z)$ and the identity operator, respectively. Denote $H : \mathcal{H}_{\mathcal{X}} \to \mathcal{Y}$ as the conditional operator we aim to learn. Then for $\gamma \geq 0$, define the population-level kernel anchor regression operator H^γ as

$$H^\gamma = \arg\min_H \mathbb{E}[\|(Id - P_{\phi(Z)})(Y - H\psi(X))\|^2] + \gamma \mathbb{E}[\|P_{\phi(Z)}(Y - H\psi(X))\|^2]. \quad (5)$$

To unravel $P_{\phi(Z)}$, both IV and AR estimators applied the two-stage procedure, where the first stage is to estimate the projection operator $P_{\phi(Z)}$ and the second stage is to perform the projection-adjusted regression.

3.1 PROJECTION STAGE

The projection stage aims to tackle $P_{\phi(Z)}$ by transforming the problem of learning $P_{\phi(Z)}\psi(X)$ and $P_{\phi(Z)}Y$ into two separate kernel ridge regressions. Let operators $E_X : \mathcal{H}_{\mathcal{Z}} \to \mathcal{H}_{\mathcal{X}}$ and $E_Y : \mathcal{H}_{\mathcal{Z}} \to \mathcal{Y}$ be the projections to learn; $\alpha_1, \alpha_2 > 0$ be regularization parameters. We note that due to the explicit dependency from Z to Y, $P_{\phi(Z)}Y$ needs to be treated separately from $P_{\phi(Z)}\psi(X)$. This is different from the IV setting where $P_{\phi(Z)}Y = Y$. The objectives regularized by Hilbert-Schmidt (HS) norm are

$$\mathcal{E}_{\alpha_1}(E_X) = \mathbb{E}\|\psi(X) - E_X\phi(Z)\|^2_{\mathcal{H}_{\mathcal{X}}} + \alpha_1\|E_X\|^2_{HS}, \quad (6)$$

$$\mathcal{E}_{\alpha_2}(E_Y) = \mathbb{E}\|Y - E_Y\phi(Z)\|^2_{\mathcal{Y}} + \alpha_2\|E_Y\|^2_{HS}. \quad (7)$$

Denote the optimal operators for the population risks as $E^p_{\alpha_1,X} = \arg\min_{E_X} \mathcal{E}_{\alpha_1}(E_X)$, and $E^p_{\alpha_2,Y} = \arg\min_{E_Y} \mathcal{E}_{\alpha_2}(E_Y)$. We then consider two variants of empirical risks and their corresponding estimations.

3.1.1 Disjoint sample sets projection

Firstly, we treat two ridge regressions in Equation (6) and Equation (7) independently, by using two *disjoint* sets of samples $\mathbb{S}_1 = \{(x_i, z_i)\}_{i\in[n_1]}$ and $\mathbb{S}_2 = \{(y_j, z_j)\}_{j\in[n_2]}$. The empirical forms for Equation (6) and Equation (7) are

$$\frac{1}{n_1}\sum_{i\in[n_1]} \|\psi(x_i) - E_X\phi(z_i)\|^2_{\mathcal{H}_{\mathcal{X}}} + \alpha_1\|E_X\|^2_{HS}, \quad (8)$$

$$\frac{1}{n_2}\sum_{j\in[n_2]} \|y_j - E_Y\phi(z_j)\|^2_{\mathcal{Y}} + \alpha_2\|E_Y\|^2_{HS}. \quad (9)$$

Denote $\Phi_{1,Z} = (\phi(z_1), \ldots, \phi(z_{n_1}))$, $\{z_i\}_{i\in[n_1]} \subset \mathbb{S}_1$; $\Phi_{2,Z} = (\phi(z_1), \ldots, \phi(z_{n_2})), \{z_j\}_{j\in[n_2]} \subset \mathbb{S}_2$; their gram matrices $K_{1,ZZ} = \Phi^\top_{1,Z}\Phi_{1,Z} \in \mathbb{R}^{n_1\times n_1}$ and $K_{2,ZZ} = \Phi^\top_{2,Z}\Phi_{2,Z} \in \mathbb{R}^{n_2\times n_2}$. Denote $\Psi_{1,X} = (\psi(x_1), \ldots, \psi(x_{n_1})), \{x_i\}_{i\in[n_1]} \subset \mathbb{S}_1$ and $Y_2 = (y_1, \ldots, y_{n_2}), \{y_j\}_{j\in[n_2]} \subset \mathbb{S}_2$. By the standard regression formula, the optimal operators to minimize Equation (8) and Equation (9) are

$$E^{n_1}_{\alpha_1,X} = \Psi_{1,X}(K_{1,ZZ} + n_1\alpha_1 I)^{-1}\Phi^\top_{1,Z}, \quad (10)$$

$$E^{n_2}_{\alpha_2,Y} = Y_2(K_{2,ZZ} + n_2\alpha_2 I)^{-1}\Phi^\top_{2,Z}, \quad (11)$$

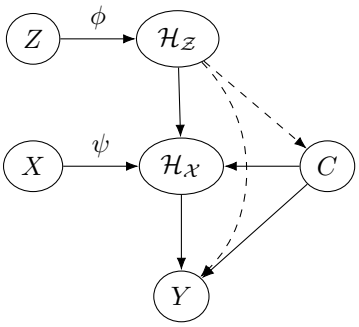

Figure 2: DAG representation for kernel anchor regression.

[3]We note that, as opposed to Figure 1, there is no edges between Z, X and Y as the learning is not based on the original data space, yet using the feature space $\phi(X) \in \mathcal{H}_X$ and $\psi(Z) \in \mathcal{H}_Z$.

where the superscripts n_1, n_2 explicitly reveal sample sizes. We note that the projections $P_{\phi(Z)}$ are estimated differently for $P_{\phi(Z)}\psi(X)$ and $P_{\phi(Z)}Y$, through $(K_{1,ZZ} + n_1\alpha_1 I)^{-1}$ and $(K_{2,ZZ} + n_2\alpha_2 I)^{-1}$, respectively. $K_{1,ZZ}$ and $K_{2,ZZ}$ are independent due to the disjoint i.i.d. sample sets of Z.

3.1.2 Joint sample set projection

On the other hand, we can also consider the projection analogous to [Rothenhäusler et al., 2018] where we jointly consider the samples used for both projections, i.e. projecting onto the same $\phi(Z)$ subspace. Setting $n = n_1 + n_2$ and $\alpha = \alpha_1 = \alpha_2$, we consider the joint sample set $\mathbb{S} = \{(x_i, y_i, z_i)\}_{i\in[n]}$ and the empirical risks

$$\frac{1}{n}\sum_{i\in[n]} \|\psi(x_i) - E_X\phi(z_i)\|^2_{\mathcal{H}_\mathcal{X}} + \alpha\|E_X\|^2_{HS}, \tag{12}$$

$$\frac{1}{n}\sum_{i\in[n]} \|y_i - E_Y\phi(z_i)\|^2_{\mathcal{Y}} + \alpha\|E_Y\|^2_{HS}. \tag{13}$$

Denote $K_{ZZ} \in \mathbb{R}^{n\times n}$ as the gram matrix from $\{z_i\}_{i\in[n]} \subset \mathbb{S}$; $\Phi_Z = (\phi(z_1), \ldots, \phi(z_n)), \{z_i\}_{i\in[n]} \subset \mathbb{S}$; $\Psi_X = (\psi(x_1), \ldots, \psi(x_n)), \{x_i\}_{i\in[n]} \subset \mathbb{S}$ and $Y = (y_1, \ldots, y_n), y_i \in \mathbb{S}$. Then we have

$$E^n_{\alpha,X} = \Psi_X(K_{ZZ} + n\alpha I)^{-1}\Phi_Z^\top, \tag{14}$$

$$E^n_{\alpha,Y} = Y(K_{ZZ} + n\alpha I)^{-1}\Phi_Z^\top. \tag{15}$$

By setting the same level of regularization, we can see that the estimates of $P_{\phi(Z)}$ projection, through $(K_{ZZ} + n\alpha I)^{-1}$, are the same for $P_{\phi(Z)}\psi(X)$ and $P_{\phi(Z)}Y$.

3.2 REGRESSION STAGE

With the learned projections $P_{\phi(Z)}\psi(X)$ and $P_{\phi(Z)}Y$, we can now tackle the overall objective in Equation (5).

Denote $\mathcal{E}(E_X)$ and $\mathcal{E}(E_Y)$ as the unregularized version of Equation (6) and Equation (7); E^p_X and E^p_Y their corresponding optimal operators, respectively. For given γ, define the transformed input and output as

$$\psi_\gamma(X) = \psi(X) - E^p_X\phi(Z) + \sqrt{\gamma}E^p_X\phi(Z) \in \mathcal{H}_\mathcal{X}, \tag{16}$$

$$Y_\gamma = Y - E^p_Y\phi(Z) + \sqrt{\gamma}E^p_Y\phi(Z) \in \mathcal{Y}. \tag{17}$$

Proposition 1 (Equivalence) *Let $H : \mathcal{H}_\mathcal{X} \to \mathcal{Y}$, and consider the regression of transformed output in Equation* (17) *on transformed input in Equation* (16)

$$\mathcal{E}^\gamma(H) = \mathbb{E}_{(Z,X,Y)}\|Y_\gamma - H\psi_\gamma(X)\|^2_{\mathcal{Y}}. \tag{18}$$

The solution to Equation (18) *is equivalent to the KAR estimator in Equation* (5)*, i.e. $H^\gamma = \arg\min_H \mathcal{E}^\gamma(H)$.*

The proof is by expanding the projection E^p_X and E^p_Y, which is similar to the linear case in Rothenhäusler et al. [2018].

With regularization parameter $\xi \geq 0$, Equation (18) has the kernel ridge regression form defined as

$$\mathcal{E}^\gamma_\xi(H) = \mathbb{E}_{(Z,X,Y)}\|Y_\gamma - H\psi_\gamma(X)\|^2_{\mathcal{Y}} + \xi\|H\|^2_{HS}.$$

The regression stage is regardless of how the projections are estimated in Section 3.1. For the empirical estimation for operators $\widehat{E}_X \in \{E^{n_1}_{\alpha_1,X}, E^n_{\alpha,X}\}$ and $\widehat{E}_Y \in \{E^{n_2}_{\alpha_2,Y}, E^n_{\alpha,Y}\}$, we use sample set $\mathbb{S}^m = \{(x_l, y_l, z_l)\}_{l\in[m]}$, which is disjoint to the set $\mathbb{S}$ used in the previous stages, and compute the transformed inputs and outputs as

$$\widehat{\psi}_{\gamma,l}(x) = \psi(x_l) + (\sqrt{\gamma} - 1)\widehat{E}_X\phi(z_l) \in \mathcal{H}_\mathcal{X},$$

$$\widehat{y}_{\gamma,l} = y_l + (\sqrt{\gamma} - 1)\widehat{E}_Y\phi(z_l) \in \mathcal{Y}.$$

The empirical risk has the form

$$\widehat{\mathcal{E}}^{\gamma,m}_\xi(H) = \frac{1}{m}\sum_{l\in[m]} \|\widehat{y}_{\gamma,l} - H\widehat{\psi}_{\gamma,l}(x)\|^2_{\mathcal{Y}} + \xi\|H\|^2_{HS},$$

$$\widehat{H}^{\gamma,m}_\xi = \arg\min \widehat{\mathcal{E}}^{\gamma,m}_\xi(H).$$

Denote $\widehat{\Psi}_\gamma = (\widehat{\psi}_{\gamma,1}(x), \ldots, \widehat{\psi}_{\gamma,m}(x))$ and its gram matrix $K_{\widehat{\Psi}_\gamma\widehat{\Psi}_\gamma} = \widehat{\Psi}^\top_\gamma\widehat{\Psi}_\gamma \in \mathbb{R}^{m\times m}$; $\widehat{Y}_\gamma = (\widehat{y}_{\gamma,1}, \ldots, \widehat{y}_{\gamma,m})$. Again, by standard regression formula,

$$\widehat{H}^{\gamma,m}_\xi = \widehat{Y}_\gamma(K_{\widehat{\Psi}_\gamma\widehat{\Psi}_\gamma} + m\xi I)^{-1}\widehat{\Psi}^\top_\gamma. \tag{19}$$

3.3 KAR ESTIMATOR

Given observational data of size N, $\{(x_i, y_i, z_i)\}_{i\in[N]}$, the KAR procedure can be performed in two ways based on the two variants in the projection stage.

Three-stage KAR To apply the disjoint sample sets projection in Section 3.1.1, we *randomly* split the data set of size N into three disjoint sets of sample size n_1, n_2, m where $N = n_1 + n_2 + m$ and re-index them in $[N]$. The first two sets of data $\{(x_i, z_i)\}_{i\in\{1:n_1\}}$ and $\{(y_j, z_j)\}_{j\in\{n_1+1:n_1+n_2\}}$ are used for learning the projection operators in Equation (10) and Equation (11). We note that samples $\{y_i\}_{i\in\{1:n_1\}}$ and $\{x_j\}_{j\in\{n_1+1:n_1+n_2\}}$ are not used. The third set $\{(x_l, y_l, z_l)\}_{l\in\{n_1+n_2+1:N\}}$ is used in regression stage to learn $\widehat{H}^{\gamma,m}_\xi$ in Equation (19). This procedure, termed **KAR**, includes solving three different regression problems, and differs from the two-stage settings used in linear anchor regression [Rothenhäusler et al., 2018].

Two-stage KAR For the joint sample set projection in Section 3.1.2, we only split the data of size N into two disjoint sets *randomly* of size n and m where $N = n + m$ and re-index them such that $\{(x_i, y_i, z_i)\}_{i\in\{1:n\}}$ and $\{(x_l, y_l, z_l)\}_{l\in\{n+1:N\}}$. The first set is then grouped into $\{(x_i, z_i)\}_{i\in\{1:n\}}$ and $\{(y_i, z_i)\}_{j\in\{1:n\}}$ to learn the projection operators in Equation (10) and Equation (11). In this manner, $\{z_i\}_{i\in\{1:n\}}$ are used twice. The second set $\{(x_l, y_l, z_l)\}_{l\in\{n+1:N\}}$ is used for regression stage to learn $\widehat{H}_\xi^{\gamma,m}$ in Equation (19), which is the same as the three-stage procedure above. This procedure, termed **KAR.2**, replicates the 2SLS used in KIV [Singh et al., 2019] and linear anchor regression [Rothenhäusler et al., 2018]. Note that the KIV procedure in [Singh et al., 2019] can be seen as a special case of our **KAR.2** by choosing $\gamma = \infty$.

4 ANALYSIS OF KAR ESTIMATORS

4.1 CONSISTENCY

We first focus on the three-stage KAR procedure with disjoint sample sets for projection in Section 3.1.1. The closed form solutions and convergence rates of the estimators are extended from the analysis of 2SLS in KIV [Singh et al., 2019]. We follow the integral operator notations in [Singh et al., 2019]. Define the projection stage operators as

$$\begin{aligned} S_1^* &: \mathcal{H}_\mathcal{Z} \to L^2(\mathcal{Z}, \rho_\mathcal{Z}), \quad l \to \langle l, \phi(\cdot)\rangle_{\mathcal{H}_\mathcal{Z}}, \\ S_1 &: L^2(\mathcal{Z}, \rho_\mathcal{Z}) \to \mathcal{H}_\mathcal{Z}, \quad \tilde{l} \to \int \phi(z)\tilde{l}(z)d\rho_\mathcal{Z}(z), \end{aligned}$$

where ρ denotes the joint distribution of (Z, X, Y). $L^2(\mathcal{Z}, \rho_\mathcal{Z})$ denotes the space of square integrable functions from $\mathcal{Z}$ to $\mathcal{Y}$ with respect to measure $\rho_\mathcal{Z}$, where $\rho_\mathcal{Z}$ is the restriction of ρ to $\mathcal{Z}$. $T_1 = T_2 = S_1 \circ S_1^*$ are then uncentered covariance operators. We define the power of operator T_1 with respect to its eigendecomposition. Let $\mathcal{H}_\Gamma = \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Z}$, $\mathcal{H}_\Theta = \mathcal{Y} \otimes \mathcal{H}_\mathcal{Z}$ and $\mathcal{H}_\Omega = \mathcal{Y} \otimes \mathcal{H}_\mathcal{X}$ be the relevant tensor product spaces for the operators.

Condition 1 *(i) $\mathcal{X}$ and $\mathcal{Z}$ are Polish, i.e. separable and completely metrizable topological spaces. (ii) $k_\mathcal{X}$ and $k_\mathcal{Z}$ are continuous and bounded:* $\sup_{x\in\mathcal{X}} \|\psi(x)\|_{\mathcal{H}_\mathcal{X}} \leq Q_1$, $\sup_{z\in\mathcal{Z}} \|\phi(z)\|_{\mathcal{H}_\mathcal{Z}} \leq \kappa$. *(iii) ψ and ϕ are measurable. (iv) $k_\mathcal{X}$ is characteristic. (v)* $E_X^p \in \mathcal{H}_\Gamma$ *s.t.* $\mathcal{E}(E_X^p) = \inf_{E_X\in\mathcal{H}_\Gamma} \mathcal{E}(E_X)$. *(vi) Fix $\zeta_1 < \infty$. For $c_1 \in (1, 2]$, define the prior $\mathcal{P}(\zeta_1, c_1)$ as the set of probability distributions ρ on $\mathcal{X} \times \mathcal{Z}$ s.t. $\exists G_1 \in \mathcal{H}_\Gamma$ s.t. $E_X^p = T_1^{(c_1-1)/2} \circ G_1$ and $\|G_1\|_{\mathcal{H}_\Gamma}^2 \leq \zeta_1$.*

Condition 1 is adapted from [Singh et al., 2019] to bound the approximation error of the regularized estimator $E_{\alpha_1,X}^{n_1}$. Parameter c_1 suggests the smoothness of conditional operator $E_{\alpha_1,X}^{n_1}$. A larger c_1 corresponds to a smoother operator.

Lemma 1 *$\forall\alpha_1 > 0$, the solution $E_{\alpha_1,X}^{n_1}$ of the regularized empirical objective in Equation* (8) *exists and is unique. With* $\mathbf{T}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \phi(z_i) \otimes \phi(z_i)$ *and* $\mathbf{g}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \phi(z_i) \otimes \psi(x_i)$, *the estimator in Equation* (10) *has the form* $E_{\alpha_1,X}^{n_1} = (\mathbf{T}_1 + \alpha_1)^{-1} \circ \mathbf{g}_1$. *Under Condition 1 and* $\alpha_1 = n_1^{-1/(c_1+1)}$, *we have:* $\|E_{\alpha_1,X}^{n_1} - E_X^p\|_{\mathcal{H}_\Gamma} = O_p(n_1^{-\frac{c_1-1}{2(c_1+1)}})$.

Lemma 1 follows from [Singh et al., 2019], and shows that the efficient rate of α_1 is $n_1^{-1/(1+c_1)}$. Note that the convergence rate of $E_{\alpha_1,X}^{n_1}$ is calibrated by c_1, which measures the smoothness of the conditional expectation operator E_X.

For the disjoint set projection in Section 3.1.1, the closed form solution and convergence rate for learning $P_{\phi(Z)}Y$ estimator is similar to that of learning $P_{\phi(Z)}\psi(X)$ due to the independent estimation procedure and further requires the following conditions.

Condition 2 (i) *$\mathcal{Y}$ is a Polish space.* (ii) *Y is bounded:* $\sup_{y\in\mathcal{Y}} \|y\|_\mathcal{Y} \leq Q_2$. (iii) $E_Y^p \in \mathcal{H}_\Theta$ *s.t.* $\mathcal{E}(E_Y^p) = \inf_{E_Y\in\mathcal{H}_\Theta} \mathcal{E}(E_Y)$. (iv) *Fix $\zeta_2 < \infty$. For $c_2 \in (1, 2]$, define the prior $\mathcal{P}(\zeta_2, c_2)$ as the set of probability distributions ρ on $\mathcal{Y} \times \mathcal{Z}$ s.t. $\exists G_2 \in \mathcal{H}_\Theta$ s.t. $E_Y^p = T_2^{(c_2-1)/2} \circ G_2$ and $\|G_2\|_{\mathcal{H}_\Theta}^2 \leq \zeta_2$.*

Lemma 2 *$\forall\alpha_2 > 0$, the solution $E_{\alpha_2,Y}^{n_2}$ of the regularized empirical objective in Equation* (9) *exists and is unique. With* $\mathbf{T}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \phi(z_j) \otimes \phi(z_j)$ *and* $\mathbf{g}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \phi(z_j)y_j$, *the estimator in Equation* (10) *has the form* $E_{\alpha_2,Y}^{n_2} = (\mathbf{T}_2 + \alpha_2)^{-1} \circ \mathbf{g}_2$. *Under Condition 1– 2 and* $\alpha_2 = n_2^{-1/(c_2+1)}$, *we have:* $\|E_{\alpha_2,Y}^{n_2} - E_Y^p\|_{\mathcal{H}_\Theta} = O_p(n_2^{-\frac{c_2-1}{2(c_2+1)}})$.

Similar to learning projection $P_{\phi(Z)}\psi(X)$, the efficient rate of α_2 is $n_2^{-1/(1+c_2)}$, where c_2 measures the smoothness of the conditional expectation operator E_Y.

Let $L^2(\mathcal{H}_\mathcal{X}, \rho_{\mathcal{H}_\mathcal{X}})$ denote the space of square integrable functions from $\mathcal{H}_\mathcal{X}$ to $\mathcal{Y}$ with respect to measure $\rho_{\mathcal{H}_\mathcal{X}}$, where $\rho_{\mathcal{H}_\mathcal{X}}$ is the extension of ρ to $\mathcal{H}_\mathcal{X}$. Define the regression stage operator as

$$\begin{aligned} S^* &: \mathcal{H}_\Omega \to L^2(\mathcal{H}_\mathcal{X}, \rho_{\mathcal{H}_\mathcal{X}}), \quad H \to \Omega_{(\cdot)}^* H, \\ S &: L^2(\mathcal{H}_\mathcal{X}, \rho_{\mathcal{H}_\mathcal{X}}) \to \mathcal{H}_\Omega, \\ &\quad \tilde{H} \to \int \Omega_{\psi_\gamma} \circ \tilde{H}\psi_\gamma d\rho_{\mathcal{H}_\mathcal{X}}(\psi_\gamma), \end{aligned}$$

where $\Omega_{\psi_\gamma} : \mathcal{Y} \to \mathcal{H}_\Omega$ defined by $y \to \Omega(\cdot, \psi_\gamma)y$ is the point evaluator of [Micchelli and Pontil, 2005]. Define $T_{\psi_\gamma} = \Omega_{\psi_\gamma} \circ \Omega_{\psi_\gamma}^*$ and covariance operator $T = S \circ S^*$. Define the power of operator T with respect to its eigendecomposition. Condition 3 below extends hypothesis 7–9 in [Singh et al., 2019], and is sufficient to bound the excess error of $\widehat{H}_\xi^{\gamma,m}$ with the error propagated from the estimators in the projection stage.

Condition 3 (i) *The $\{\Omega_{\psi_\gamma}\}$ operator family is uniformly bounded in Hilbert-Schmidt norm: $\exists B$ s.t. $\forall \psi_\gamma$, $\|\Omega_{\psi_\gamma}\|^2_{L_2(\mathcal{Y},\mathcal{H}_\Omega)} = Tr(\Omega^*_{\psi_\gamma} \circ \Omega_{\psi_\gamma}) \leq B$.*

(ii) *The $\{\Omega_{\psi_\gamma}\}$ operator family is is Hölder continuous in operator norm: $\exists L > 0, \iota \in (0,1]$ s.t. $\forall \psi_\gamma, \psi'_\gamma, \|\Omega_{\psi_\gamma} - \Omega_{\psi'_\gamma}\|_{L(\mathcal{Y},\mathcal{H}_\Omega)} \leq L\|\psi_\gamma - \psi'_\gamma\|^\iota_{\mathcal{H}_\mathcal{X}}$.*

(iii) *$H^\gamma \in \mathcal{H}_\Omega$, then $\mathcal{E}^\gamma(H^\gamma) = \inf_{H\in\mathcal{H}_\Omega} \mathcal{E}^\gamma(H)$.*

(iv) *Y_γ is bounded, i.e. $\exists C < \infty$ s.t. $\|Y_\gamma\|_\mathcal{Y} \leq C$.*

(v) *Fix $\zeta < \infty$. For given $b_\gamma \in (1,\infty]$ and $c_\gamma \in (1,2]$, define the prior $\mathcal{P}(\zeta, b_\gamma, c_\gamma)$ as the set of probability distributions ρ on $\mathcal{H}_\mathcal{X} \times \mathcal{Y}$ s.t.*

(a) *range space assumption is satisfied: $\exists G \in \mathcal{H}_\Omega$ s.t. $H^\gamma = T^{\frac{(c_\gamma-1)}{2}} \circ G$ and $\|G\|^2_{\mathcal{H}_\Omega} \leq \zeta$;*

(b) *the eigenvalues from spectral decomposition $T = \sum_{k=1}^\infty \lambda_k e_k \langle \cdot, e_k\rangle_{\mathcal{H}_\Omega}$, where $\{e_k\}_{k=1}^\infty$ is basis of $Ker(T)^\perp$, satisfy $\alpha \leq k^{b_\gamma}\lambda_k \leq \beta$ for $\alpha, \beta > 0$.*

We note that all parameters mentioned in Condition 3 depend on γ, though the function representations are not explicit. We set subscript γ especially for parameters b_γ and c_γ to emphasize their dependency on γ. Parameter b_γ measures the decay of eigenvalues of the covariance operator T, and larger b_γ suggests smaller effective input dimension. A larger c_γ corresponds to a smoother operator H^γ.

Lemma 3 *$\forall \xi > 0$, the solution $\widehat{H}^{\gamma,m}_\xi$ to $\widehat{\mathcal{E}}^{\gamma,m}_\xi$ exists and is unique for each γ. Let $\widehat{\mathbf{T}} = \frac{1}{m}\sum_{l=1}^m T_{\widehat{\psi}_{\gamma,l}}$, $\widehat{\mathbf{g}} = \frac{1}{m}\sum_{l=1}^m \Omega_{\widehat{\psi}_{\gamma,l}}\widehat{y}_{\gamma,l}$. Equation* (19) *has the form*

$$\widehat{H}^{\gamma,m}_\xi = (\widehat{\mathbf{T}} + \xi)^{-1} \circ \widehat{\mathbf{g}}.$$

Condition 4 *For c_1, c_2 set in Conditions 1– 2 and ι satisfying Condition 3, assume $n_2 \geq n_1^{\frac{\iota(c_1-1)(c_2+1)}{(c_1+1)(c_2-1)}}$.*

Remark 1 *Condition 4 is sufficient but not necessary to ensure that the error propagates to regression stage from estimating E^p_Y is smaller than that from estimating E^p_X in disjoint sample sets projection.*

The main challenge of extending the convergence rate of KIV estimator [Singh et al., 2019] to KAR estimator is that in our case, the excess error depends not only on the accuracy of E^p_X estimator but also on the accuracy of E^p_Y estimator. However, by proposing Condition 4, we ensure the error from estimating E^p_Y is dominated by that of E^p_X, and manage to illustrate the optimal convergence rate for KAR as shown in Thereom 1. In this way, the three-stage procedure can guarantee the same convergence rate as the two-stage procedure in KIV.

Theorem 1 *Under Condition 1– 4, let $d_1, d_2 > 0$ and choose $\alpha_1 = n_1^{-\frac{1}{c_1+1}}$, $\alpha_2 = n_2^{-\frac{1}{c_2+1}}$, $n_1 = m^{\frac{d_1(c_1+1)}{\iota(c_1-1)}}$, $n_2 = m^{\frac{d_2(c_2+1)}{\iota(c_2-1)}}$, we have:*

(i) If $d_1 \leq \frac{b_\gamma(c_\gamma+1)}{b_\gamma c_\gamma+1}$, then $\mathcal{E}^\gamma(\widehat{H}^{\gamma,m}_\xi) - \mathcal{E}^\gamma(H^\gamma) = O_p(m^{-\frac{d_1 c_\gamma}{c_\gamma+1}})$ with $\xi = m^{-\frac{d_1}{c_\gamma+1}}$.

(ii) If $d_1 > \frac{b_\gamma(c_\gamma+1)}{b_\gamma c_\gamma+1}$, then $\mathcal{E}^\gamma(\widehat{H}^{\gamma,m}_\xi) - \mathcal{E}^\gamma(H^\gamma) = O_p(m^{-\frac{b_\gamma c_\gamma}{b_\gamma c_\gamma+1}})$ with $\xi = m^{-\frac{b_\gamma}{b_\gamma c_\gamma+1}}$.

At $d_1 = b_\gamma(c_\gamma+1)/(b_\gamma c_\gamma+1) < 2$, the convergence rate of KAR estimator $m^{-b_\gamma c_\gamma/(b_\gamma c_\gamma+1)}$ is optimal. This statistically efficient rate is calibrated by b_γ, the effective input dimension, together with c_γ, the smoothness of the operator H^γ. The condition $d_1 = b_\gamma(c_\gamma+1)/(b_\gamma c_\gamma+1) < 2$ also suggests that $n_1 > m$. We provide additional discussion on the two-stage KAR estimator, and show that only a lower convergence rate can be guaranteed (see Section A.3 in supplementary material).

4.2 CAUSAL EFFECT AND TARGET KAR ESTIMATE

In this section, we discuss the scenarios assuming that the data are generated from a structural causal model with nonlinear features as shown below,

$$\begin{pmatrix} C \\ \psi(X) \\ Y \end{pmatrix} = B \begin{pmatrix} \phi(Z) \\ C \\ \psi(X) \\ Y \end{pmatrix} + \begin{pmatrix} \epsilon_C \\ \epsilon_X \\ \epsilon_Y \end{pmatrix}, \tag{20}$$

where we write operator B in the following matrix form

$$B = \begin{pmatrix} B_{CZ} & 0 & 0 & 0 \\ B_{XZ} & B_{XC} & 0 & 0 \\ B_{YZ} & B_{YC} & B_{YX} & 0 \end{pmatrix}.$$

We note that each operator $B_{\triangle\square}$ represents an operator that takes an element from $\square$-related space to $\triangle$-related space, e.g. $B_{XZ} : \mathcal{H}_\mathcal{Z} \to \mathcal{H}_\mathcal{X}$ and $B_{YZ} : \mathcal{H}_\mathcal{Z} \to \mathcal{Y}$. The noise variables ϵ_Z, ϵ_C, ϵ_X and ϵ_Y are independent of each other. Let Σ_Z, Σ_C, Σ_X and Σ_Y denote the covariance of ϵ_Z, ϵ_C, ϵ_X and ϵ_Y, respectively. Here each operator in B represents a line in the model shown in Figure 2. For instance, B_{CZ} stands for the line from $\mathcal{H}_\mathcal{Z}$ to C; B_{YX} corresponds to the line from $\mathcal{H}_\mathcal{X}$ to Y. B_{YX} in Equation (20) reflects the causal effect we are interested in. We study the identifiability scenarios where operator B_{YX} can be learned via KAR estimator H^γ.

Theorem 2 *An operator B_{XZ} is a zero operator written by $B_{XZ} = 0$, if $\langle \psi(x), B_{XZ}\phi(z)\rangle_{\mathcal{H}_\mathcal{X}} = 0$, $\forall \psi(X) \in \mathcal{H}_\mathcal{X}, \phi(z) \in \mathcal{H}_\mathcal{Z}$. Operator $B_{CZ} = 0$ if $c^\top B_{CZ}\phi(z) = 0$, $\forall c \in \mathcal{C}, \phi(z) \in \mathcal{H}_\mathcal{Z}$. A matrix-valued operator, e.g. $B_{YC} = 0$ if all entries are 0. For data generation process following Equation* (20)*, we have $H^\gamma = B_{YX}$ in following cases.*

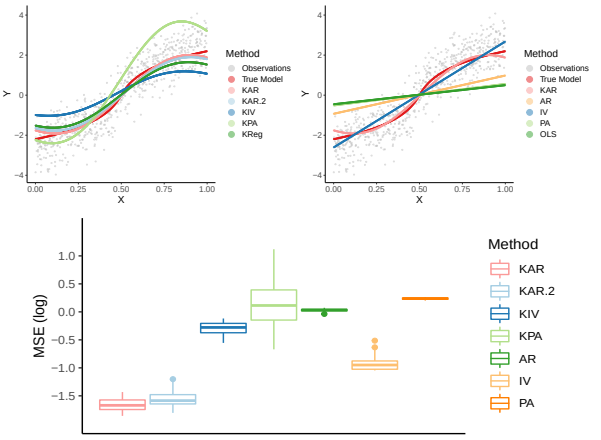

Figure 3: Synthetic example: fitted (top left) nonlinear models; (top right) linear models; (bottom): log MSE.

(i) $B_{YC} = 0$ *and* $\gamma = 0$*, i.e. no latent confounder.*

(ii) $B_{YZ} + B_{YC}B_{CZ} = 0$ *and* $\gamma = \infty$*, where kernel IV is a special case, i.e. both* $B_{YZ} = 0$ *and* $B_{CZ} = 0$*.*

(iii) $B_{YC} = 0$*,* $B_{YZ} + B_{YC}B_{CZ} = 0$*, and* $\gamma \geq 0$*.*

(iv) $\Sigma^{||}_{YX} = -a\Sigma^{\perp}_{YX}$ *for some* $a > 0$*, and* $\gamma = 1/a$*.*

$\Sigma^{||}_{YX} = (B_{YZ} + B_{YC}B_{CZ})\Sigma_Z(B_{ZX} + B_{ZC}B_{CX})$ *denotes the covariance between* $\psi(X)$ *and* Y *projected on the linear span from the components of* $\phi(Z)$*; and* $\Sigma^{\perp}_{YX} = B_{YC}\Sigma_C B_{CX}$ *denotes the covariance between the residuals of* $\psi(X)$ *and* Y*.*

Thereom 2 (i) suggests that KPA (kernel partialling out regression) is optimal when there is no unobserved confounder; (ii) suggests that KIV identifies the causal effect under a generalized condition including KIV assumption, i.e. $B_{YZ} = 0$ and $B_{YC} = 0$; (iii) shows the KAR estimator identifies the causal relation from X to Y regardless of γ with generalized KIV condition in (ii) and no latent confounder in (i). The condition $\gamma \geq 0$ in (iii) actually implies that H^{γ} is constant over all $\gamma \geq 0$, which coincides with the definition of anchor stability in [Rothenhäusler et al., 2018]. Thereom 2 (iv) shows the KAR identifiability condition with appropriate choice of γ when $\Sigma^{||}_{XY}$ and $\Sigma^{\perp}_{XY}$ are in the flipped direction. To further illustrate the identifiability condition, consider the linear case and assume that X, Z and C have only one dimension. In this case, it's not trivial for $\Sigma^{||}_{YX} = -a\Sigma^{\perp}_{YX}$ holds for some $a > 0$, which ensures KAR to identify the causal effect. We stress that the identifiability condition in Theorem 2 (iv) does not include no hidden confounding ($B_{YC} = 0$) nor valid instrument variable ($B_{YZ} = 0$ and $B_{YC} = 0$).

In the next section, we show the empirical results for KAR estimators compared with relevant baseline methods.

5 EMPIRICAL RESULTS

5.1 SYNTHETIC EXPERIMENTS

We consider the data generating process of the following nonlinear structural equation,

$$Y = 0.75C - 0.25Z + \ln(|16X - 8| + 1)sgn(X - 0.5),$$

where $sgn(x) \in \{-1, 0, +1\}$ denotes the sign of x. The explanatory variables X, Z, C are generated from

$$\begin{pmatrix} C \\ V \\ W \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1, 0.3, 0.2 \\ 0.3, 1, 0 \\ 0.2, 0, 1 \end{pmatrix}\right),$$
$$X = F\left(\frac{W + V}{\sqrt{2}}\right), \quad Z = F(W) - 0.5,$$

where F denote the c.d.f of standard normal distribution. For our learning procedure, Z, X and Y are available, and C is unobservable.

We generate $\{(x_i, y_i, z_i)\}_{i\in[N]}$ with $N = 700$. To perform the data-splitting procedures described in Section 3, we set $n_1 = n_2 = 250$ and $n = 500$ for a fair comparison in the projection stage ; and $m = 200$ in the regression stage . We set regularizers as $\alpha_1 = 1.5n_1^{-0.5}$, $\alpha_2 = 1.5n_2^{-0.5}$, $\alpha = 1.5n^{-0.5}$ and $\xi = 1.5m^{-0.5}$.

Fitting methods We consider estimations via the three-stage kernel anchor regression with disjoint data set projection (**KAR**) and two-stage kernel anchor regression with joint data set projection (**KAR.2**). The baseline approaches include the kernel-based nonlinear methods: kernel instrument variable regression (**KIV**), kernel partialling out regression (**KPA**), kernel ridge regression (**KReg**); and the linear models: linear anchor regression (**AR**), linear instrument variable regression (**IV**), linear partialling out regression (**PA**) and ordinary least square (**OLS**). We use Gaussian kernel for all kernel-based methods, where the median heuristic is used for choosing the bandwidth [Gretton et al., 2012]. We show that the median heuristic is a good choice by achieving close-to-optimal cross-validation error (see Section B.3 in supplementary material). For the synthetic example, we set $\gamma = 2$ for all anchor regressions (**KAR**, **KAR.2** and **AR**).

For each algorithm, we implement 50 trials and calculate the mean squared error (MSE) with respect to the true causal model $\mathbb{E}(Y|do(x))$[4], which can be computed from the structural model. A trial is shown in Figure 3 as a visual example. We can see that the **KAR** produces a closest estimation to the true model among all methods and outperforms

[4]Setting a particular value $X = x$ while ignoring other variables that may potentially changing the distribution of y, $p(y|X = x)$ is noted as $p(y|do(x)$ [Pearl, 2009, Peters et al., 2016]. $\mathbb{E}[Y|do(x)]$ is set us the mean over $p(y|do(x))$ averaging out different Z values in this case.

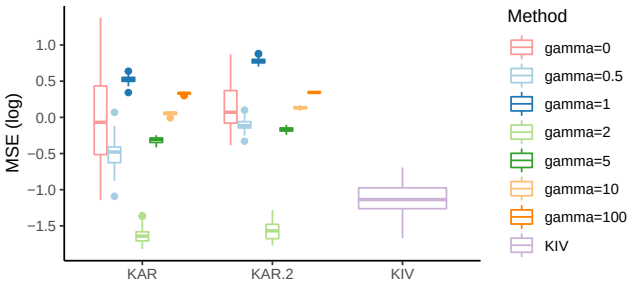

Figure 4: Effects of different γ choices on MSE.

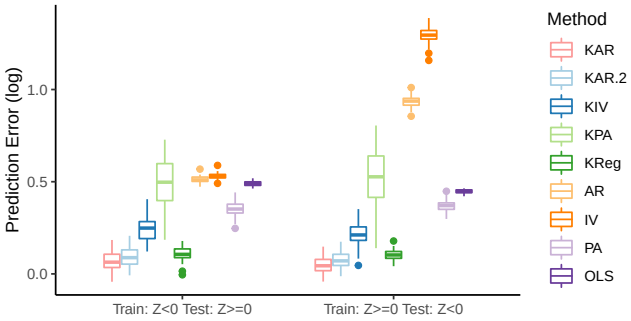

Figure 5: Prediction error with distributional intervention.

KAR.2. The comparison with linear models are also shown. IV model fits better than other linear models. We report $\log_{10}$(MSE) in the bottom of Figure 3, which shows that both KAR methods have smaller errors than others. **KAR** performs slightly better than **KAR.2** in this case. To check the robustness of KAR estimators, we study a less smooth variant of the data generating process and show the results in Section B.2 in supplementary material.

The effect of γ choices To investigate how the change of γ affects the estimator, we consider KIV as our baseline as the IV setting corresponds to $\gamma \to \infty$. We consider the data generating process used in the KIV paper Singh et al. [2019]. The $\log_{10}$(MSE) results of **KAR** and **KAR.2**, in comparison with **KIV**, are shown in Figure 4. For the simulation, we set $N = 1000$, $n_1 = n_2 = 200$, $n = n_1 + n_2 = 400$ and $m = 600$. From the result, we can see that both **KAR** and **KAR.2** achieves smaller error when choosing $\gamma = 2$. Data generation and model implementation details are included in Section B.1 supplementary material.

Intervention and Generalization To evaluate the robustness and generalization performance of both KAR estimators under distribution shift, as discussed in [Rothenhäusler et al., 2018], we intervene the anchor variable Z. We train the model on a subpopulation of samples with $Z < 0$ and test on the samples with $Z \geq 0$. The performance is measured by prediction error (PE) of fitted model with respect to $\mathbb{E}(Y|X = x, Z \geq 0)$, where the true conditional model is not known in closed form but estimated from samples.

We also exchange the training set and the testing set. As shown in Figure 5, our **KAR** estimator has the lowest PE among others, showing better out-of-distribution generalization performance. More importantly, by checking the two (flipped) scenarios, i.e. train on $Z < 0$ v.s. train on $Z \geq 0$, we also see that **KAR** is the most invariant in terms of PE. On the contrary, linear version of **AR** and **IV** achieves very different PE in both cases. Variances of PE for **KPA** are also very different in the two cases. Despite **KReg** achieving a relatively low PE in both cases, the distributions of PE can be found very different. In supplementary material (see Section B.1), we also illustrate that KAR can outperform other non kernel-based approaches, e.g. DeepIV or SmoothIV.

5.2 REAL-WORLD APPLICATION

We consider the smoking dataset extracted from National Medical Expenditure Survey (NMES) [Johnson et al., 2003] to study the effect of smoking amount on medical expenditure [Imai and Van Dyk, 2004][5].

The treatment variable X is the log of smoking amount, the outcome Y is the log of medical expenditure, and the anchor Z is set to be the last age for smoking. We use 1000 samples, randomly selected from 9708 available samples, to fit the model. We set $n_1 = n_2 = 300$, $n = 600$ and $m = 400$. We also set $\gamma = 2.9$ and apply Gaussian kernel with median heuristic bandwidth [Gretton et al., 2012] for all kernel methods. As shown in the upper part of Figure 6, KAR estimators show that the effect of X on Y is more significant when $X \in [-2, 1]$ compared to $X \in [1, 4]$. Our method can also be used in complement with the approaches finding causal directions, e.g. [Peters et al., 2016][6]. We run the CAM to ensure that there is a causal effect in the direction from X to Y and KAR procedure further learns the specific function representing such effect. However, existing work such as propensity score approaches [Imai and Van Dyk, 2004] did not manage to extract such causal relationship between smoking and medical expenditure.

To strengthen our finding, we quantify the performance of the estimators. Since we do not know the real generating process of the data, we cannot compare the MSE as Figure 3 and 4. Instead, it's feasible to evaluate the estimators' performance under distribution perturbation via PE, similar to Figure 5. We train models on male subjects and compute the prediction accuracy of fitted model on female subjects. As shown in the bottom of Figure 6, we see that both KAR approaches outperform other kernel-based approaches as well as the linear version of AR, suggesting a better learned

[5]The dataset is accessible through using the `R` package for "estimating causal dose response function" `causaldrf` [Galagate, 2016] `https://cran.r-project.org/web/packages/causaldrf/index.html`.

[6]Implementation with `R` package `CAM` can be found at `https://rdrr.io/cran/CAM/man/CAM.html`

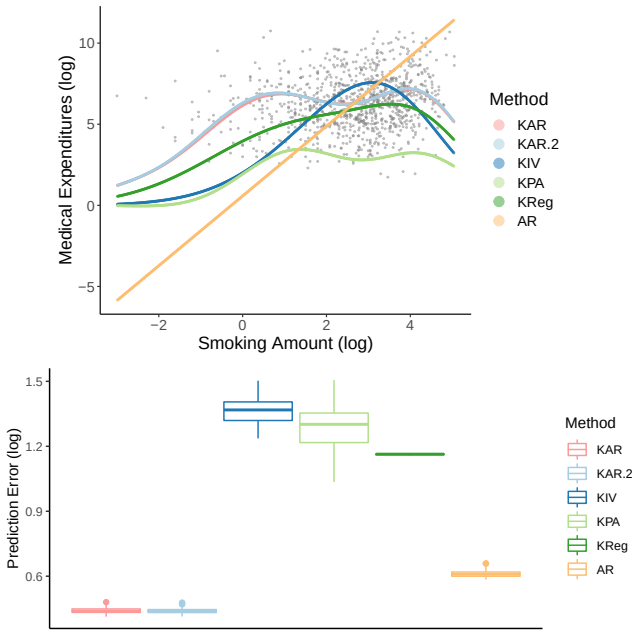

Figure 6: Fitted models (top) and prediction errors (bottom) for training on male subjects and testing on female subjects.

causal effect for smoking amount to medical expenditure.

6 CONCLUSION AND FUTURE WORKS

In this work, we consider learning a more general class of causal DAG in a nonlinear setting using kernelized anchor regression. By considering different data splitting strategies to estimate the projection operators, we show that the three-stage approach not only performs better empirically than baseline approaches as well as the 2SLS approach, but also achieves optimal rate under given conditions. Identifiability results for our approach are provided and are shown to generalize KIV and "no latent confounder" scenarios.

Our study opens several directions to better understand the nonlinear causal effect using the anchor regression framework. For the future, data adaptive choice of γ and its causal interpretation can be further to explore. Moreover, while we focus on effect variable Y in its original space in this work, anchor regression for the feature space of Y can be another interesting future study.

We also note that Rothenhäusler et al. [2018] studies the distribution generalisation property for the linear anchor regression. While, this work does not focus on the theoretical properties of generalisation using RKHS functions in the non-linear setting, the possibility and conditions to achieve distribution generalisation property for non-linear anchor regression is another interesting future direction.

Acknowledgements

The authors thank Ana Korba and Arthur Gretton for helpful discussions. W.X. acknowledges the support from EPSRC grant EP/T018445/1.

References

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

A. Bennett, N. Kallus, and T. Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019.

M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *International Conference on Learning Representations*, 2019.

C. M. Bishop. Latent variable models. In *Learning in graphical models*, pages 371–403. Springer, 1998.

D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

R. J. Bowden and D. A. Turkington. *Instrumental variables*. Number 8. Cambridge university press, 1990.

P. Bühlmann. Invariance, causality and robustness. *arXiv preprint arXiv:1812.08233*, 2018.

S. Centorrino, F. Fève, and J.-P. Florens. Nonparametric instrumental regressions with (potentially discrete) instruments independent of the error term. *arXiv preprint arXiv:1905.07812*, 2019.

V. Chernozhukov, I. Fernández-Val, and B. Melly. Inference on counterfactual distributions. *Econometrica*, 81(6): 2205–2268, 2013.

R. Christiansen, N. Pfister, M. E. Jakobsen, N. Gnecco, and J. Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2021.

D. Colombo, M. H. Maathuis, M. Kalisch, and T. S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

N. Dikkala, G. Lewis, L. Mackey, and V. Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. *Advances in neural information processing systems*, 20, 2007.

D. Galagate. *Causal inference with a continuous treatment and outcome: Alternative estimators for parametric dose-response functions with applications*. PhD thesis, University of Maryland, College Park, 2016.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

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

S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1803–1810, 2012.

J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The annals of statistics*, 36(3):1171–1220, 2008.

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

A. Hyvärinen. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371 (1984):20110534, 2013.

A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

A. Hyvarinen and H. Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

K. Imai and D. A. Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467): 854–866, 2004.

M. E. Jakobsen and J. Peters. Distributional robustness of k-class estimators and the pulse. *The Econometrics Journal*, 25(2):404–432, 2022.

F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

E. Johnson, F. Dominici, M. Griswold, and S. L. Zeger. Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics*, 112(1):135–151, 2003.

N. Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *International Conference on Machine Learning*, pages 5067–5077. PMLR, 2020.

I. Khemakhem, R. Monti, R. Leech, and A. Hyvarinen. Causal autoregressive flows. In *International conference on artificial intelligence and statistics*, pages 3520–3528. PMLR, 2021.

L. Kook, B. Sick, and P. Bühlmann. Distributional anchor regression. *Statistics and Computing*, 32(3):1–19, 2022.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

R. P. Monti, K. Zhang, and A. Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, pages 186–195. PMLR, 2020.

J. Mooij, D. Janzing, J. Peters, and B. Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual international conference on machine learning*, pages 745–752, 2009.

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

K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

M. Oberst, N. Thams, J. Peters, and D. Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pages 8260–8270. PMLR, 2021.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

J. Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 589–598, 2011.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

L. Rabiner and B. Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, and F. Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

D. B. Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2):161–170, 2004.

D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

R. Singh, M. Sahani, and A. Gretton. Kernel instrumental variable regression. *arXiv preprint arXiv:1906.00232*, 2019.

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

P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman. *Causation, prediction, and search*. MIT press, 2000.

P. L. Spirtes, C. Meek, and T. S. Richardson. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

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

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

L. Xu, Y. Chen, S. Srinivasan, N. de Freitas, A. Doucet, and A. Gretton. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*, 2020.

L. Xu, H. Kanagawa, and A. Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. *Advances in Neural Information Processing Systems*, 34:26264–26275, 2021.

K. Zhang, B. Schölkopf, P. Spirtes, and C. Glymour. Learning causality and causality-related learning: some recent progress. *National science review*, 5(1):26–29, 2018.

Y. Zhu, L. Gultchin, A. Gretton, M. J. Kusner, and R. Silva. Causal inference with treatment measurement error: a nonparametric instrumental variable approach. In *Uncertainty in Artificial Intelligence*, pages 2414–2424. PMLR, 2022.