Kernel Single Proxy Control for Deterministic Confounding

Liyuan Xu **Arthur Gretton**
Gatsby Computational Neuroscience Unit

Abstract

We consider the problem of causal effect estimation with an unobserved confounder, where we observe a *single* proxy variable that is associated with the confounder. Although it has been shown that the recovery of an average causal effect is impossible in general from a single proxy variable, we show that causal recovery is possible if the outcome is generated deterministically. This generalizes existing work on causal methods with a single proxy variable to the continuous treatment setting. We propose two kernel-based methods for this setting: the first based on the two-stage regression approach, and the second based on a maximum moment restriction approach. We prove that both approaches can consistently estimate the causal effect, and we empirically demonstrate that we can successfully recover the causal effect on challenging synthetic benchmarks.

1 Introduction

Causal learning involves estimating the impact of our actions on the world. For instance, we might want to understand how flight ticket prices affect sales, or how a specific medication increases the chance of recovery. The action we take is referred to as a "treatment," which leads to a specific "outcome." Determining the causal effect solely from observational data is often challenging, however, because the observed relationship between treatment and outcome can be influenced by an unobserved factor known as a "confounder." In the case of the medical example, patient health might act as a confounder: if the patient is weaker, then he or she might require a higher dose of medicine, and yet have a lower chance of survival. Thus, from observing historical data, the dose may be *negatively correlated* with the chance of recovery; an instance of *Simpson's paradox*. Hence, it is necessary to address the bias caused by the confounder to assess the true effect.

Although many methods for causal inference methods assume that there are *no unmeasured confounders*, this can be too restrictive, and it is often difficult to determine how the confounder affects treatment assignments and outcomes. A milder assumption is that we have access to *proxy variables*, which are correlated to the confounder. It is known that we cannot in general recover the true causal effect with a single proxy variable (Pearl, 2010; Kuroki and Pearl, 2014): for this reason, studies have focused on the case where two proxy variables are available (Kuroki and Pearl, 2014; Miao et al., 2018; Deaner, 2018; Mastouri et al., 2021; Singh, 2020; Xu et al., 2021).

In this paper, we introduce a novel sufficient condition which characterizes when causal effect estimation from a *single* proxy variable is possible. Our key assumption is *deterministic confounding*, meaning that the outcome is deterministically generated given the treatment and the confounder. This generalizes *Single Proxy Control* (Tchetgen Tchetgen et al., 2023), which considers the setting where the confounder is the potential outcome (Rubin, 2005). As we will see, the deterministic confounding assumption can be also be understood with reference to a similar requirement for nonlinear independent component analysis (Hyvarinen et al., 2019).

Given the deterministic confounding assumption, we show that we can use the outcome itself as a "proxy variable" in the setting of Proxy Causal Learning (PCL) (Deaner, 2018; Miao et al., 2018). This insight enables us to use a technique adapted from kernel PCL (Singh, 2020; Mastouri et al., 2021) to estimate the causal effect. We further conduct a theoretical and empirical sensitivity analysis on the violation of the deterministic confounding assumption. We also derive a novel closed-form solution for kernel PCL methods with improved numerical stability, which may be of independent interest.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

This paper is structured as follows. After reviewing the related work, we present an overview of the problem setting, and of relevant background material on kernel methods in Section 2. In Section 3, we establish conditions under which the causal effect can be recovered from the observed variables. Then, in Section 4, we introduce our estimation procedures: namely, the two-stage least squares and moment matching approaches, as in Mastouri et al. (2021) , with consistency results for each. In Section 5, we report the empirical performance of the proposed method on synthetic causal data. We also explore the case where the deterministic confounding assumption is violated, and show that our proposed methods can still obtain a reasonable estimate of the causal relationship.

Related Work Despite the negative results shown by Kuroki and Pearl (2014), various attempts have been made to estimate causal effects from proxy variables. CEVAE (Louizos et al., 2017) uses a VAE (Kingma and Welling, 2014) to recover the distribution of confounders using a single proxy variable. Although it lacks theoretical justification and can be unstable (Rissanen and Marttinen, 2021), CEVAE has shown strong empirical performance in several settings.

Proxy causal learning (PCL) (Deaner, 2018; Miao et al., 2018) guarantees to recover the true causal effect given *two* of proxy variables, under existence and identifiability conditions. The original work of Miao et al. (2018) provides a closed-form solution in the case of categorical treatments and outcomes, and derives a Fredholm equation of the first kind to characterize the continuous case. Solutions for the latter case have been obtained where the causal relationships are modeled using a dictionary of smooth basis functions (Deaner, 2018), RKHS functions (Mastouri et al., 2021; Singh, 2020), and neural networks (Xu et al., 2021; Kompa et al., 2022; Kallus et al., 2021). Our main contribution is to show that we can use the outcome as a proxy in PCL given an additional deterministic assumption.

Negative control outcome only requires *one* proxy variable, but places stricter assumptions on the confounding structure. Tchetgen Tchetgen et al. (2023) introduced *Single Proxy Control*, which generalizes the *Difference-in-Difference (DiD)* method (Angrist and Pischke, 2008; Card and Krueger, 1994). Although the work relaxes the strict parallel trending assumption in DiD, they consider only the case of binary treatment, and require the proxy to reveal the potential outcome (Rubin, 2005) of the treated. Our method generalizes Single Proxy Control to continuous treatments and a more general confounding mechanism. A relation with nonlinear independent component analysis (NICA) (Hyvarinen et al., 2019) is described in the next section, once the required notation has been introduced.

2 Preliminaries

Notation. Throughout the paper, a capital letter (e.g. A) denotes a random variable, and we denote the set where a random variable takes values by the corresponding calligraphic letter (e.g. $\mathcal{A}$). The symbol $\mathbb{P}[\cdot]$ denotes the probability distribution of a random variable (e.g. $\mathbb{P}[A]$). We use a lowercase letter to denote a realization of a random variable (e.g. a). We denote the expectation over a random variable as $\mathbb{E}[\cdot]$ and $\|f\|_{\mathbb{P}[\cdot]}$ as the L^2-norm of a function f with respect to $\mathbb{P}[\cdot]$; i.e. $\|f\|_{\mathbb{P}[A]} = \sqrt{\mathbb{E}_A[f^2(A)]}$.

2.1 Problem Setting

We begin by describing the problem setting. We observe the treatment $A \in \mathcal{A}$ and corresponding outcome $Y \in \mathcal{Y}$. We assume that there exists an unobserved confounder $U \in \mathcal{U}$ that affects both A and Y. Our goal is to estimate the structural function f_{struct} defined as

$$f_{\text{struct}}(\tilde{a}) = \mathbb{E}_U\left[\mathbb{E}_Y\left[Y|A=\tilde{a},U\right]\right],$$

for a given test point $\tilde{a} \in \mathcal{A}$. This function is also known as the *Dose-Response Curve* or *Average Potential Outcome*. The challenge in estimating f_{struct} is that the confounder U is not observable — we cannot estimate the structural function from observations A and Y alone. To deal with this, we assume access to a proxy W, which satisfies the following two assumptions:

Assumption 2.1. We assume $W \perp\!\!\!\perp (A, Y)|U$.

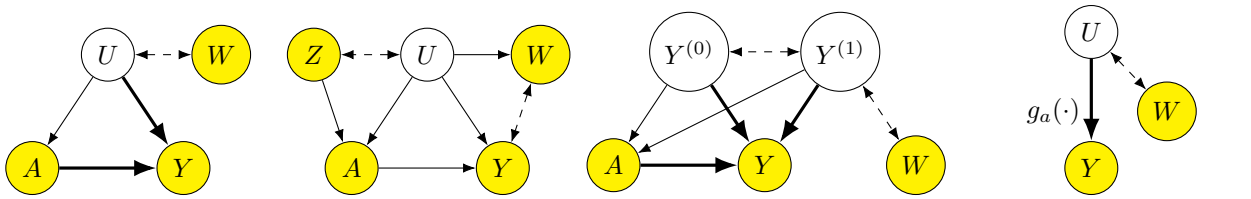
Assumption 2.2. For any square integrable function $l : \mathcal{U} \to \mathbb{R}$, the following conditions hold for all $a \in \mathcal{A}$:

$$\mathbb{E}\left[l(U) \mid A = a, W = w\right] = 0 \quad \forall w \in \mathcal{W}$$
$$\Leftrightarrow \quad l(u) = 0 \quad \mathbb{P}[U]\text{-a.e.}$$

Figure 1a shows the causal graph describing these relationships. We may additionally consider an observable confounder, which is discussed in Appendix A.

The assumptions on the proxy W are similar to those made in *Proxy Causal Learning (PCL)* (Deaner, 2018; Miao et al., 2018; Mastouri et al., 2021; Singh, 2020; Xu et al., 2021; Kallus et al., 2021). Our setting only assumes access to a single proxy W, however, while PCL requires an additional treatment-proxy Z that satisfies $Z \perp\!\!\!\perp Y|A, U$ as in Figure 1b. Instead of observing this proxy, we place deterministic confounding assumptions on the outcome Y, as follows.

Assumption 2.3. There exists a *deterministic* function $\gamma_0(a, u)$ such that $Y = \gamma_0(A, U)$ and $\|\gamma_0(a, U)\|_{\mathbb{P}[U]} \leq \infty$ for all $a \in \mathcal{A}$.



(a) Our setting (b) PCL setting (c) Single Proxy Control setting (d) Nonlinear ICA setting

Figure 1: Causal graphs with proxy variables. Here, the bidirectional arrows mean that we allow an arrow in either direction, or even a common ancestor variable; and the thick arrow indicates a deterministic relationship between the variables.

This deterministic confounding assumption is unique in our setting: we will further discuss its necessity in Section 3. Our main contribution is to show that we may use outcome Y as the treatment-proxy Z in PCL if we make Assumption 2.3. Further details can be found in Appendix B.

Applications One typical application is a time-evolving deterministic process. For example, in a chemical synthesis plant, the amount of the final product Y can be computed given perfect knowledge of the plant state U and the control signal A, since the reaction process can be modeled by a deterministic differential equation (Beeler et al., 2023).

Single Proxy Control (Tchetgen Tchetgen et al., 2023), whose causal graph is shown in Figure 1c, also considers deterministic confounding. From the definition of the potential outcome (Rubin, 2005), Assumption 2.3 holds with $\gamma_0(a, Y^{(0)}, Y^{(1)}) = \mathbb{1}(a = 0)Y^{(0)} + \mathbb{1}(a = 1)Y^{(1)}$. Single Proxy Control applies only to the binary treatment case, however, whereas our setting generalizes the approach to continuous treatments.

A related setting is nonlinear independent component analysis (NICA) (Hyvarinen et al., 2019), in which the observation Y is given by a *smooth, deterministic, invertible*, nonlinear function of unobserved latents U, as shown in Figure 1d. Denoting this function as g, Hyvarinen et al. (2019) prove that g can be recovered up to certain indeterminacies if we have an informative proxy variable W, and assume that the entries of U are mutually independent given W. While we share our *deterministic confounding* assumption and use of an informative proxy with NICA, our goal is different: we do not recover U, nor do we require assumptions concerning the dependence structure within U. Our aim is rather to estimate the causal effect $\mathbb{E}[g_a(U)]$ under intervention a and confounding U, where we have modeled a *family* of functions g_a indexed by a.

2.2 Reproducing Kernel Hilbert Space

For any space $\mathcal{F} \in \{\mathcal{A}, \mathcal{Y}, \mathcal{W}\}$, let $k_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be a positive semidefinite kernel. Here, we assume all kernel functions are characteristic Sriperumbudur et al. (2010) and bounded, $\max_{f,f' \in \mathcal{F}} k_{\mathcal{F}}(f, f') \leq \kappa$. We denote corresponding RKHS as $\mathcal{H}_{\mathcal{F}}$ and canonical feature as $\phi_{\mathcal{F}}$. The inner product and the norm in $\mathcal{H}_{\mathcal{F}}$ is denoted as $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{F}}}$ and $\| \cdot \|_{\mathcal{H}_{\mathcal{F}}}$, respectively. We also define the tensor product as $\otimes$, which induces the Hilbert-Schmidt operator $f \otimes g : \mathcal{H}_{\mathcal{F}_2} \to \mathcal{H}_{\mathcal{F}_1}$ for $f \in \mathcal{H}_{\mathcal{F}_1}, g \in \mathcal{H}_{\mathcal{F}_2}$, such that $(f \otimes g)h = \langle g, h \rangle_{\mathcal{H}_{\mathcal{F}_2}} f$ for $h \in \mathcal{H}_{\mathcal{F}_2}$. We denote the space of Hilbert-Schmidt operators of $\mathcal{H}_{\mathcal{F}_2} \to \mathcal{H}_{\mathcal{F}_1}$ as $\mathcal{H}_{\mathcal{F}_1 \mathcal{F}_2}$ with Hilbert-Schmidt norm $\| \cdot \|_{\mathcal{H}_{\mathcal{F}_1 \mathcal{F}_2}}$. We introduce the notation of $\phi_{\mathcal{F}_1 \mathcal{F}_2}(f_1, f_2) = \phi_{\mathcal{F}_1}(f_1) \otimes \phi_{\mathcal{F}_2}(f_2)$. Note that it can be shown that $\mathcal{H}_{\mathcal{F}_1 \mathcal{F}_2}$ is isometrically isomorphic to the product space $\mathcal{H}_{\mathcal{F}_1} \times \mathcal{H}_{\mathcal{F}_2}$.

3 Identification

In this section, we show that the structural functions can be estimated from observable variables (Y, A, W) given assumptions in Section 2.1. Similar to classical PCL, we consider the bridge function whose partial average corresponds to the structural function.

Theorem 3.1. *Given Assumptions 2.1 to 2.3 and Assumptions E.2 to E.3 presented in Appendix E.1, then there exists a bridge function $h_0(a, w)$ such that*

$$\mathbb{E}[h_0(a, W) | A = a, Y = y] = y \quad \forall (a, y) \in \mathcal{A} \times \mathcal{Y}. \tag{1}$$

Using the connection between our setting and PCL discussed in Appendix B, Theorem 3.1 can be proved by generalizing PCL identification (Miao et al., 2018). The proof details can be found in Appendix E.2.

Under the sufficient condition discussed in Theorem 3.3, we can obtain the structural function by taking the spatial average of bridge function

$$f_{\text{struct}}(\tilde{a}) = \mathbb{E}_W[h_0(\tilde{a}, W)]. \tag{2}$$

We first demonstrate this in the case where all variables follow linear structural equations and then move to a more general case. Finally, we conduct a theoretical analysis where the deterministic confounding assumption (Assumption 2.3) is violated.

Bridge function in the linear setting To illustrate our method and the implications of our assumptions, we first discuss the case where all the variables follow structural linear equations. Assume all variables are in $\mathbb{R}$ and follow:

$$U = \varepsilon_u,\ W = \alpha_{wu}U + \varepsilon_w,\ A = \alpha_{au}U + \varepsilon_a,$$
$$Y = \alpha_{ya}A + \alpha_{yu}U + \varepsilon_y,$$

where $\varepsilon_u, \varepsilon_w, \varepsilon_a, \varepsilon_y$ are independent random variables with zero mean, and we assume all coefficients are non-zero. This setting is discussed by Kuroki and Pearl (2014), who show that we cannot recover α_{ya} unless we can estimate $\alpha_{wu}^2 \mathrm{Var}[U]$. Our work avoids this requirement, however, by making the deterministic assumption.

Proposition 3.2. *If we assume deterministic confounding $\varepsilon_y = 0$, the function*

$$h_0(a, w) = \alpha_{ya}a + (\alpha_{yu}/\alpha_{wu})w$$

satisfies equation (1), and we have $\mathbb{E}[h_0(\tilde{a}, W)] = f_{\text{struct}}(\tilde{a}) = \alpha_{ya}\tilde{a}$.

Proof. Let bridge function be $h_0(a, w) = h_a a + h_w w$. Then, the left hand side of (1) becomes

$$\begin{aligned}
&\mathbb{E}[h_0(a, W)|A = a, Y = y] \\
&\quad = h_a a + h_w \mathbb{E}[W|A = a, Y = y] \\
&\quad = h_a a + h_w \alpha_{wu} \mathbb{E}[U|A = a, Y = y] \\
&\quad = h_a a + h_w \alpha_{wu}(y - \alpha_{ya}a)/\alpha_{yu},
\end{aligned}$$

where the last equation holds for deterministic confounding $\varepsilon_y = 0$. Hence, if we solve (1), we have $h_a = \alpha_{ya}, h_w = \alpha_{yu}/\alpha_{wu}$. Furthermore, we have $\mathbb{E}[h_0(a, W)] = \alpha_{ya}a$ since $\mathbb{E}[W] = 0$. □

From Proposition 3.2, we can see that the structural function f_{struct} can be given as the partial average of bridge function $h_0(a, w)$ if we make the deterministic assumption. Note that (1) only contains the observable variables, and thus we can estimate bridge function h_0 from samples (A, Y, W) even in non-linear settings, as discussed in Section 4.

Bridge function in the general case It is important to note that the solution of (1) is not unique in general. Let h_0 be the "true" bridge function that satisfies (2). Then all functions in

$$\mathcal{S} = \{h_0 + \delta(a, w) \mid \mathbb{E}[\delta(a, W)|A = a, Y = y] = 0\}$$

satisfy (1), and some $h \in \mathcal{S}$ might give a different partial average $\mathbb{E}[h(\tilde{a}, W)] \neq f_{\text{struct}}(\tilde{a})$. In the following theorem, we provide the sufficient condition that all partial averages in $\mathcal{S}$ correspond to the structural function.

Theorem 3.3. *For test point $\tilde{a} \in \mathcal{A}$, given either $W \perp\!\!\!\perp A|Y^{(\tilde{a})}$ or*

$$\begin{aligned}
&\mathbb{E}[l(U) \mid A = \tilde{a}, Y = y] = 0 \quad \forall y \in \mathcal{Y} \\
&\quad \Leftrightarrow l(u) = 0 \quad \mathbb{P}[U]\text{-a.e.},
\end{aligned} \tag{3}$$

then all functions in $\mathcal{S}$ satisfy $\mathbb{E}[h(\tilde{a}, W)] = f_{\text{struct}}(\tilde{a})$.

The assumption $W \perp\!\!\!\perp A|Y^{(\tilde{a})}$ is employed by Tchetgen Tchetgen et al. (2023), and ensures that the proxy W only affects the treatment through the confounder U whose mechanism can be captured by the potential output to be estimated $Y^{(\tilde{a})}$. In (Tchetgen Tchetgen et al., 2023), it is discussed that municipality-level birth rates before the Zika virus outbreak satisfy the assumption when the treatment A is the Zika epidemic status and the outcome Y is the post-epidemic birth rate. Our analysis extends the results of Tchetgen Tchetgen et al. (2023) to general treatments A, while only binary treatment $\mathcal{A} = \{0, 1\}$ is considered in the original analysis.

Another assumption in (3) is similar to the "informative treatment-proxy" assumption in PCL; hence, we call it the "informative outcome" condition. This implies that the outcome can reveal the information on the hidden confounder. The informative outcome assumption also holds in our chemical synthesis setting, since we can (in theory) recover the transient U from the final Y by solving the differential equation backward in time. Note that we do not need both $W \perp\!\!\!\perp A|Y^{(\tilde{a})}$ and (3); either one is sufficient for identification.

The condition in Theorem 3.3 is not testable in practice and might be violated. However, our method appears robust to the violation, as shown in the empirical experiments in Section 5.

Sensitivity Analysis on Deterministic Confounding One natural question is whether we can recover the structural function when the outcome is given stochastically. In such a case, we cannot recover the bridge function h_0 by solving (1), and thus the estimate of the structural function is unavoidably biased.[1] However, we can bound the bias if the noise is bounded and additive:

[1] We might be tempted to include the additive noise ε in $Y = \gamma(A, U) + \varepsilon$ as part of the confounder $\tilde{U} = (U, \varepsilon)$, however this violates our modeling assumptions since the proxy W is then not sufficiently informative for ε (i.e. Assumption 2.2 fails for $\tilde{U}$).

Theorem 3.4. *Assume that outcome is given as* $Y = \gamma_0(A, U) + \varepsilon$ *where* $\mathbb{E}[\varepsilon] = 0$, $|\varepsilon| \leq M$. *Suppose Assumption 2.1 holds, and that there exists* $\Xi \leq \infty$ *for all bounded function* $\ell(u)$,

$$\sup_u |\ell(u)| \leq \Xi \sup_{y \in \mathcal{Y}} \mathbb{E}[\ell(U)|A = a, Y = y] \tag{4}$$

for all $(a, y) \in \mathcal{A} \times \mathcal{Y}$. *If a bridge function* h *satisfying (1) exists, then for all* $a \in \mathcal{A}$, *we have*

$$|f_{\text{struct}}(a) - \mathbb{E}_W[h(a, W)]| \leq M\Xi$$

The proof can be found in Appendix E.3. Note that the condition in (4) is the sufficient condition for (3). One important difference compared to Theorem 3.3 is that the existence of the bridge function $h_0(a, w)$ is not guaranteed in Theorem 3.4. This is because we need Assumption 2.3 to prove the existence of the bridge function. Although Theorem 3.4 only upper-bounds the error, the experiments in Section 5 show promising empirical results. Further theoretical analysis on this stochastic case is left as future work.

4 Methods

In this section, we present our proposed methods to estimate the bridge function. We introduce two algorithms, *Single Kernel Proxy Variable (SKPV)* and *Single Proxy Maximum Moment Restriction (SPMMR)*, which estimate the bridge function using different approaches.

From Theorem 3.1, we can see that it is sufficient to estimate the bridge function to obtain the structural function, which can be done by solving functional equation (1) using samples $\{a_i, w_i, y_i\}_{i=1}^n$. This problem is ill-posed when we consider a sufficiently rich functional space (see discussion of Nashed and Wahba, 1974). Here, we assume that the bridge function h is in RKHS $\mathcal{H}_{\mathcal{AW}}$ and add a smoothness requirement. We will discuss the case where the bridge function h can be represented as neural network in Appendix D.

We propose two approaches to solve (1): A *Two-stage regression approach* (Singh, 2020; Xu et al., 2021; Mastouri et al., 2021), which minimizes the L^2-distance between both sides of (1); and a *Maximum Moment Restriction approach* (Mastouri et al., 2021; Kompa et al., 2022; Kallus et al., 2021), which minimizes the maximum moment of the deviation in (1). We also derive novel closed-form solutions for these approaches, which improve the numerical stability compared to the existing solutions.

4.1 Two-Stage Regression Approach

In two-stage regression, we solve (1) by minimizing the loss

$$\mathcal{L}^{\text{2SR}}(h) = \mathbb{E}_{AY}\left[(Y - \mathbb{E}[h(A, W)|A, Y])^2\right] + \eta\|h\|^2_{\mathcal{H}_{\mathcal{AW}}},$$

where $\eta\|h\|^2_{\mathcal{H}_{\mathcal{AW}}}$ is the regularizing term. This loss cannot be directly minimized as it involves the conditional expectation $\mathbb{E}[h(A, W)|A, Y]$. To deal with this, Mastouri et al. (2021) and Singh (2020) employ a two-stage regression approach, which first estimates conditional expectation, and then minimizes the loss. We apply this approach in our single proxy setting, and propose the *Single Kernel Proxy Variable (SKPV)* method. Specifically, if $h \in \mathcal{H}_{\mathcal{AW}}$, we have

$$\mathbb{E}[h(a, W)|A = a, Y = y] = \langle h, \phi_{\mathcal{A}}(a) \otimes \mu_{W|A,Y}(a, y)\rangle_{\mathcal{H}_{\mathcal{AW}}}.$$

Here, $\mu_{W|A,Y}(a, y)$ is known as a *conditional mean embedding* (Song et al., 2009; Grünewälder et al., 2012; Li et al., 2022; Park and Muandet, 2020), defined as $\mu_{W|A,Y}(a, y) = \mathbb{E}[\phi_{\mathcal{W}}(W)|A = a, Y = y]$. Under the regularity condition, there exists an operator $C_{W|A,Y}$ such that $\mu_{W|A,Y} = C_{W|A,Y}(\phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{Y}}(y))$. This operator can be estimated from the samples $\{w_i, a_i, y_i\}_{i=1}^n$ (Song et al., 2009; Grünewälder et al., 2012; Li et al., 2022; Park and Muandet, 2020) by minimizing the loss $\hat{C}_{W|A,Y} = \arg\min_{C \in \mathcal{H}_{\mathcal{W}(\mathcal{AY})}} \hat{\mathcal{L}}^{\text{cond}}(C)$ defined as

$$\hat{\mathcal{L}}^{\text{cond}}(C) = \frac{1}{n}\sum_{i=1}^n \|\phi_{\mathcal{W}}(w_i) - C\phi_{\mathcal{AY}}(a_i, y_i)\|^2_{\mathcal{H}_{\mathcal{W}}} + \lambda\|C\|^2_{\mathcal{H}_{\mathcal{W}(\mathcal{AY})}},$$

where $\phi_{\mathcal{AY}}(a, y) = \phi_{\mathcal{A}}(a) \otimes \phi_{\mathcal{Y}}(y)$ is the tensor feature, $\lambda > 0$ is the regularizing coefficient, $\mathcal{H}_{\mathcal{W}(\mathcal{AY})}$ denotes the space of Hilbert-Schmidt operators that maps $\mathcal{H}_{\mathcal{AY}}$ to $\mathcal{H}_{\mathcal{W}}$, and $\|\cdot\|_{\mathcal{H}_{\mathcal{W}(\mathcal{AY})}}$ is the norm of that space. This is known as *stage 1 regression* as we regress the feature $\phi_{\mathcal{W}}(w_i)$ on tensor product features $\phi_{\mathcal{A}}(a_i) \otimes \phi_{\mathcal{Y}}(y_i)$. The closed-from solution is

$$\hat{C}_{W|A,Y} = \hat{C}_{W,(A,Y)}(\hat{C}_{(A,Y),(A,Y)} + \lambda I)^{-1}$$

where $\hat{C}_{W,(A,Y)} = \frac{1}{n}\sum_{i=1}^n \phi_{\mathcal{W}}(w_i) \otimes \phi_{\mathcal{AY}}(a_i, y_i)$ and $\hat{C}_{(A,Y),(A,Y)} = \frac{1}{n}\sum_{i=1}^n \phi_{\mathcal{AY}}(a_i, y_i) \otimes \phi_{\mathcal{AY}}(a_i, y_i)$. Given this, we minimize $\mathcal{L}^{\text{2SR}}$ using the estimate in stage 1 regression. Specifically, we obtain the bridge function estimate $\hat{h}$ as $\hat{h} = \arg\min_{h \in \mathcal{H}_{\mathcal{AW}}} \hat{\mathcal{L}}^{\text{2SR}}(h)$ given stage 2 data $\{\dot{a}_i, \dot{y}_i\}_{i=1}^m$, where the loss is defined as

$$\hat{\mathcal{L}}^{\text{2SR}}(h) = \frac{1}{n}\sum_{i=1}^n \left(\dot{y}_i - \left\langle h, \phi_{\mathcal{A}}(\dot{a}_i) \otimes \hat{C}_{W|A,Y}\phi_{\mathcal{AY}}(\dot{a}_i, \dot{y}_i)\right\rangle_{\mathcal{H}_{\mathcal{AW}}}\right)^2 + \eta\|h\|^2_{\mathcal{H}_{\mathcal{AW}}}.$$

This regression is known as *stage 2 regression*. We may use the same samples in stage 1 and 2 regressions, but it is reported that splitting samples between stage 1 and 2 regression can reduce bias in the estimation (Angrist and Krueger, 1995). Furthermore, since we do not need samples of W in stage 2 regression, we may use an additional source of data for the samples $\{\dot{a}_i, \dot{y}_i\}$. We can obtain the closed-form solution of two-stage regression as follows:

Theorem 4.1. *Given stage 1 samples $\{w_i, a_i, y_i\}_{i=1}^n$ and stage 2 samples $\{\dot{a}_i, \dot{y}_i\}_{i=1}^m$, and regularizing parameter (λ, η), the minimizer of $\hat{\mathcal{L}}^{\mathrm{2SR}}$ is given as $\hat{h}(a, w) = \boldsymbol{\alpha}^\top \boldsymbol{k}(a, w)$ where*

$$\boldsymbol{\alpha} = (M + m\eta I)^{-1}\dot{\boldsymbol{y}}, \ M = K_{\dot{A}\dot{A}} \odot (B^\top K_{WW} B).$$

Here, $\dot{\boldsymbol{y}} = (\dot{y}_1, \ldots, \dot{y}_m)^\top \in \mathbb{R}^m$, and $\boldsymbol{k}(a, w) = \boldsymbol{k}_{\dot{A}}(a) \odot (B^\top \boldsymbol{k}_W(w)) \in \mathbb{R}^m$ where

$$B = (K_{AA} \odot K_{YY} + n\lambda I)^{-1}(K_{A\dot{A}} \odot K_{Y\dot{Y}}).$$

For $F \in \{A, W, Y\}$, we denote the stage 1 kernel matrix as $K_{FF} = (k_{\mathcal{F}}(f_i, f_j))_{ij} \in \mathbb{R}^{n\times n}$ where $k_{\mathcal{F}}$ and f_i are the corresponding space and stage 1 samples. Similarly, for $\dot{F} \in \{\dot{A}, \dot{Y}\}$, we denote the stage 2 kernel matrix as $K_{\dot{F}\dot{F}} = (k_{\mathcal{F}}(\dot{f}_i, \dot{f}_j))_{ij} \in \mathbb{R}^{n\times n}$ where $k_{\mathcal{F}}$ and f_i are the corresponding space and stage 2 samples. We further denote $K_{F\dot{F}} = (k_{\mathcal{F}}(f_i, \dot{f}_j))_{ij} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{k}_{\dot{A}}(a) = (k(\dot{a}_i, a))_i \in \mathbb{R}^m, \boldsymbol{k}_W(w) = (k(w_i, w))_i \in \mathbb{R}^n$

The proof is in Appendix F, where we further show that this solution is equivalent to the result of Mastouri et al. (2021, Proposition 2) , which involves learning nm parameters, while ours only requires learning n parameters. Our solution is moreover slightly different from Singh (2020, Algorithm 1), which takes the form $\boldsymbol{\alpha} = (MM + m\eta M)^{-1}M\dot{\boldsymbol{y}}$. Although these solutions coincide when M is invertible, we claim that our solution is more numerically stable, especially when M is close to singular, as shown in Appendix C. The computation requires $O(n^3 + m^3)$, which can be reduced via the usual Cholesky or Nystrom techniques.

Given estimated bridge function $\hat{h}(\tilde{a}, w)$, we can estimate the structural function as $\hat{f}_{\mathrm{struct}}(\tilde{a}) = \frac{1}{n}\sum_{i=1}^n \hat{h}(\tilde{a}, w_i)$.

4.2 Maximum Moment Restriction Approach

Another approach to solving (1) is to consider moment restriction, which minimizes the maximum moment of $\mathbb{P}[Y - h(a, W)|A, Y]$. We apply this approach to our setting and propose *Single Proxy Maximum Moment Restriction (SPMMR)*. Specifically, we consider the following minimax problem.

$$\operatorname*{arg\,min}_{h\in\mathcal{H}_{\mathcal{AW}}} \max_{g\in\mathcal{H}_{\mathcal{AY}}, \|g\|_{\mathcal{H}_{\mathcal{AY}}}\le 1} (\mathbb{E}[(Y - h(A, W))g(A, Y)])^2.$$

As shown by Mastouri et al. (2021, Lemma 1) , the unique solution to this minimax problem is the solution of (1). Furthermore, using a similar discussion to Mastouri et al. (2021, Lemma 2) , we can show that the maximum over g can be computed in closed form as

$$\mathcal{L}^{\mathrm{MMR}}(h) = \mathbb{E}\left[\Delta_{Y,A,W}(h)\Delta_{Y',A',W'}(h)k_{\mathcal{A}}(A, A')k_{\mathcal{Y}}(Y, Y')\right],$$

where $\Delta_{Y,A,W}(h) = (Y - h(A, W))$ and (W', A', Y') are independent copies of (W, A, Y), respectively. Empirically, we minimize $\mathcal{L}^{\mathrm{MMR}}$ with regularization as $\hat{h} = \arg\min_{h\in\mathcal{H}_{\mathcal{AW}}} \hat{\mathcal{L}}^{\mathrm{MMR}}$, where

$$\hat{\mathcal{L}}^{\mathrm{MMR}} = \sum_{i,j=1}^n \frac{\Delta_i\Delta_j}{n^2}k_{\mathcal{A}}(a_i, a_j)k_{\mathcal{Y}}(y_i, y_j) + \eta\|h\|^2_{\mathcal{H}_{\mathcal{AW}}},$$

and $\Delta_i = (y_i - h(a_i, w_i))$. The closed-form solution of $\hat{h}$ can be obtained as follows:

Theorem 4.2. *Given samples $\{w_i, a_i, y_i\}_{i=1}^n$, the minimizer of $\hat{\mathcal{L}}^{\mathrm{MMR}}$ is given as $\hat{h}(a, w) = \boldsymbol{\alpha}^\top \boldsymbol{k}(a, w)$, where weight $\boldsymbol{\alpha} \in \mathbb{R}^n$ is*

$$\boldsymbol{\alpha} = \sqrt{G}\left(\sqrt{G}L\sqrt{G} + n^2\eta I\right)^{-1}\sqrt{G}\boldsymbol{y},$$

and

$$L = K_{AA} \odot K_{WW}, \quad G = K_{AA} \odot K_{YY}.$$

Here, $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ and $\boldsymbol{k}(a, w) = (k_{\mathcal{A}}(a_i, a)k_{\mathcal{W}}(w_i, w))_{ij} \in \mathbb{R}^n$ with $K_{FF} = (k_{\mathcal{F}}(f_i, f_j))_{ij} \in \mathbb{R}^{n\times n}$ for $F \in \{W, A, Y\}$ and corresponding kernel function $k_{\mathcal{F}}$ and samples f_i, and $\sqrt{G}$ denotes the square root of $G = \sqrt{G}\sqrt{G}$.

The derivation is given in Appendix F. Again, this slightly differs from the original solution in Mastouri et al. (2021) , which is $\boldsymbol{\alpha} = (LGL + n^2\eta L)^{-1}LG\boldsymbol{y}$, but they are identical if L, G are non-singular. We claim that our solution is more numerically stable when the condition number of L is large. The computational complexity is $O(n^3)$, which can be reduced via the usual Cholesky or Nystrom techniques. We can use the empirical average of the estimated bridge function to obtain the structural function as in SKPV.

4.3 Consistency

Here, we state the consistency results for our estimates.

Proposition 4.3. *Assume Assumptions G.1 and G.2 in Appendix G. Given stage 1 samples $\{w_i, a_i, y_i\}_{i=1}^n$ and stage 2 samples $\{\dot{a}_i, \dot{y}_i\}_{i=1}^m$,*

$$\sup_{a,w} |h_0(a, w) - \hat{h}(a, w)| \to 0$$

with $n, m \to \infty$ by reducing regularizers (λ, η) at appropriate rates.

The proof is given in Appendix G, in which we also provide the rate of convergence and the optimal values of regularizers (λ, η). We can also show similar results for SPMMR.

Proposition 4.4. *Assume Assumption G.4 in Appendix G. Given samples $\{w_i, a_i, y_i\}$ size of n. Then,*

$$\sup_{a,w} |\hat{h}(a,w) - h_0(a,w)|_{\mathcal{H}_{\mathcal{AW}}} \to 0$$

as $n \to \infty$ by reducing the regularizer λ at an appropriate rate.

Appendix G provides further details on this consistency result.

5 Experiments

In this section, we present the empirical performance of our single proxy method in a synthetic setting. Details of the experiments are summarized in Appendix H [2].

Low-Dimensional Proxy We consider the following data generation processes: We generate hidden confounder as $U \sim \text{Unif}(-1, 1)$ and

$$A = \Phi(U) + \varepsilon_1, W = \exp(U) + \varepsilon_2,$$
$$Y = \sin(\pi U/2) + A^2 - 0.3. \quad (5)$$

Here, Φ is the Gaussian error function, and $\varepsilon_1 \sim \mathcal{N}(0, (0.1)^2), \varepsilon_2 \sim \mathcal{N}(0, (0.05)^2)$. The true structural function is $f_{\text{struct}}(a) = a^2 - 0.3$. Note that deterministic confounding Assumption 2.3 and informative outcome assumption (3) hold in this setting since $U = \frac{2}{\pi} \arcsin(Y - A^2 - 0.3)$.

To estimate the structural function, we apply SKPV and SPMMR. We used the Gaussian kernel, where the bandwidth is tuned using the median trick. We selected the regularizers using the procedure described in Appendix H. We split data evenly for stage 1 and 2 regression in SKPV. We compare proposed methods to CEVAE (Louizos et al., 2017), which uses a VAE (Kingma and Welling, 2014) to recover the distribution of confounder U from the "proxy" W. Although CEVAE does not always provide consistent causal effect estimates (Rissanen and Marttinen, 2021), it has shown strong empirical performance in a number of settings. The estimation loss is summarized in the first plot of Figure 2, in which we show the error of naive regression ignoring the confounding in the red line. In Figure 2, SKPV and SPMMR consistently outperform CEVAE when the data size is set to 1000 and 5000. In our experiment, SKPV performs better than SPMMR.

[2] Codes can be found at https://github.com/liyuan9988/KernelSingleProxy/

High-Dimensional Proxy We also consider the case where the proxy is high-dimensional. We used the same data generation process for (A, Y, U) in (5) with deterministic confounding, and replaced the proxy W with MNIST images, where the digit label is chosen as $\lfloor 5U + 5 \rfloor$. Results with data of size 1000 and 5000 are summarized in the second plot in Figure 2. Since the proxy W is high-dimensional, all methods perform worse than in the low-dim proxy setting. In particular, the performance of CEVAE is highly unstable, consistent with the observation of Rissanen and Marttinen (2021) that for high dimensional W, the autoencoder can prioritize quality of reconstruction of W over causal effect estimation. SKPV and SPMMR perform considerably better than CEVAE in this setting.

Sensitivity Analysis on Deterministic Assumption We further test performance in a case where deterministic confounding (Assumption 2.3) is violated. We added Gaussian noise $\mathcal{N}(0, \sigma^2)$ to the outcome Y in (5), where we varied the noise level in $\sigma \in \{0.1, 1.0\}$. We again generated 1000 samples and repeated each noise level 10 times. The mean squared error in estimating the structural function is presented in the third plot in Figure 2. We can see that SKPV and SPMMR can still correct confounding bias even if the outcome contains noise. Furthermore, SPMMR performs better in the large noise case $\sigma = 1.0$. We hypothesize that this is because the true structural function lies in support of data distribution once noise is added, as shown in Appendix H. Overall, we see that the effect of the additive error is minimal, as suggested by Theorem 3.4.

Sensitivity Analysis on Informative Outcome Assumption To test sensitivity on the informative outcome assumption (3), we consider two settings. One is to increase the frequency of the dependence on U,

$$Y = \sin(\beta U) + A^2 - 0.3,$$

where we consider $\beta = \{\pi/2, 3\pi/4, \pi\}$. For $\beta > \pi/2$, the informative outcome assumption (3) is violated since there are multiple U that can generate the same Y. The result is shown in the left plot of Figure 4. This shows that there is indeed a marginal increase in error for both methods.

We next approximate the sine curve with the staircase function,

$$Y = \widetilde{\sin}_N(\pi U/2) + A^2 - 0.3,$$

where $\widetilde{\sin}_N$ is the discretized sine curve as shown in Figure 3. Smaller N means a wider range of U corresponding to the same Y, hence the outcome Y becomes less informative. We investigated each of $N = \{3, 10, 30\}$ with 1000 samples; results are shown in the right plot

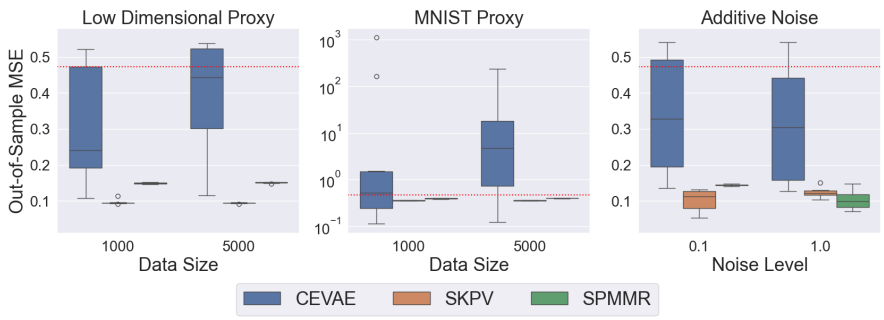


Figure 2: Results for experiments. The red dotted line shows the error of the regression $\|\mathbb{E}[Y|A] - f_{\text{struct}}(A)\|^2$.

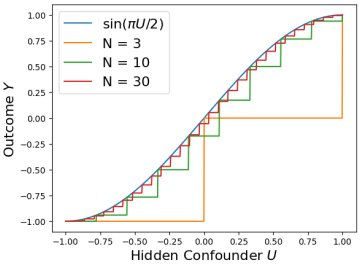


Figure 3: Plot of $\widetilde{\sin}_N(\pi U/2)$

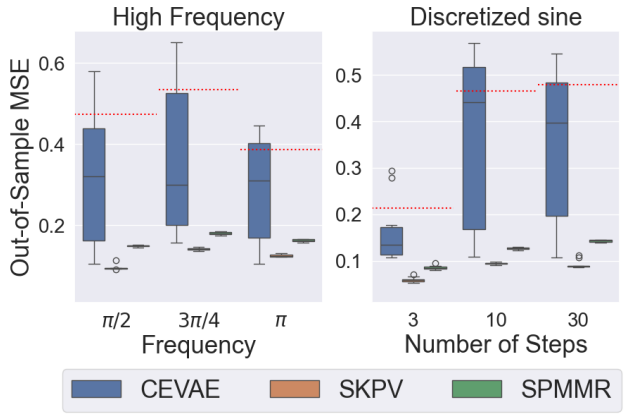


Figure 4: Result for sensitivity analysis on informative assumption.The red dotted line shows the error of the regression $\|\mathbb{E}[Y|A] - f_{\text{struct}}(A)\|^2$.

of Figure 4, in which we can see the counter-intuitive results that smaller N results in the better performance. This is because as we increase the number of steps, the confounding bias $\|\mathbb{E}[Y|A] - f_{\text{struct}}(A)\|^2$ decreases, as shown in the red dotted line in Figure 4. This demonstrates a larger principle: the existence of settings where the informative outcome assumption being "more violated" results in the outcome Y being "less dependent" on the confounder U, which results in a smaller confounding bias. Note, however, that the situation is more complicated when frequency β increases, with confounding bias first increasing, and then decreasing below that for the "informative" setting ($\beta = \pi/2$), as seen in Figure 4. In sum, a violation of the informativeness assumption (3) does not necessarily harm performance.

6 Conclusion

In this paper, we propose single proxy causal learning, which only requires a single proxy variable. Our main contribution is to extend Tchetgen Tchetgen et al. (2023) by drawing the connection with the original PCL, notably by identifying the outcome variable with the treatment proxy in the original PCL setting. This insight enables us to consider continuous treatments in the single proxy setting and use the alternative *informative outcome* assumption for the identification.

Following the theoretical work, we propose two methods, SKPV and SPMMR, to estimate the bridge function which are shown to converge to the true bridge function as the number of samples increases. We also empirically show that we can still apply these methods in the additive noise case, and demonstrate a bound on the performance penalty incurred through violation of the deterministic outcome and informative outcome assumptions.

Acknowledgments

This work was supported by the Gatsby Charitable Foundation. We thank Ben Deaner for helpful discussions.

References

Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business and Economic Statistics*, 13(2):225–235.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Beeler, C., Subramanian, S. G., Sprague, K., Chatti, N., Bellinger, C., Shahen, M., Paquin, N., Baula, M., Dawit, A., Yang, Z., Li, X., Crowley, M., and Tamblyn, I. (2023). Chemgymrl: An interactive framework for reinforcement learning for digital chemistry.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, 84(4):772–793.

Deaner, B. (2018). Proxy controls and panel data.

Fischer, S. and Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(1).

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

Hyvarinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ica using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 859–868.

Kallus, N., Mao, X., and Uehara, M. (2021). Causal inference under unmeasured confounding with negative controls: A minimax learning approach.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Kompa, B., Bellamy, D., Kolokotrones, T., Beam, A., et al. (2022). Deep learning methods for proximal inference via maximum moment restriction. *Advances in Neural Information Processing Systems*, 35:11189–11201.

Kress, R. (1999). *Linear Integral Equations*. Applied Mathematical Sciences. Springer New York.

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

Li, Z., Meunier, D., Mollenhauer, M., and Gretton, A. (2022). Optimal rates for regularized conditional mean embedding learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R. S., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 31.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. (2021). Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *Proceedings of The 38th International Conference on Machine Learning*.

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

Murphy, K. P. (2013). *Machine learning: a probabilistic perspective*. MIT Press.

Nashed, M. Z. and Wahba, G. (1974). Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations. *SIAM Journal on Mathematical Analysis*, 5(6):974–987.

Park, J. and Muandet, K. (2020). A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pages 21247–21259.

Pearl, J. (2010). On measurement bias in causal inference. In *UAI'10: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 425 – 432.

Rissanen, S. and Marttinen, P. (2021). A critical look at the consistency of causal estimation with deep latent variable models. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

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

Singh, R. (2020). Kernel methods for unobserved confounding: Negative controls, proxies, and instruments.

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

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561.

Tchetgen Tchetgen, E., Park, C., and Richardson, D. (2023). Single proxy control.

Xu, L., Kanagawa, H., and Gretton, A. (2021). Deep proxy causal learning and its application to confounded bandit policy evaluation. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Ues

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Kernel Single Proxy Control for Deterministic Confounding Supplementary Materials

A Observable Confounder

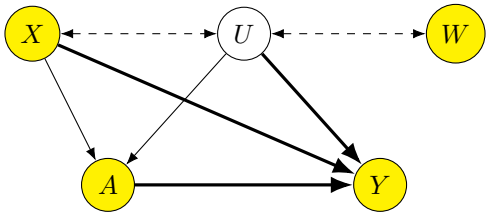


Figure 5: Typical causal graph for the single proxy variable with observable confounder X. Here, the bidirectional arrows mean that we allow an arrow in either direction, or even a common ancestor variable; and the thick arrow indicates a deterministic relationship between the variables.

In this appendix, we discuss the case where there exists the observable confounder X. The causal graph in this case is given in Figure 5. Now, the structural function is defined as

$$f_{\text{struct}}(\tilde{a}) = \mathbb{E}_{X,U}\left[\mathbb{E}\left[Y|A=\tilde{a},X,U\right]\right].$$

Given the $X \in \mathcal{X}$, we would solve the bridge function $h_0(a,x,w) : \mathcal{A}\times\mathcal{X}\times\mathcal{W}\to\mathcal{Y}$,

$$\mathbb{E}\left[h_0(a,x,W)|A=a,X=x,Y=y\right]=y \quad \forall a\in\mathcal{A}, y\in\mathcal{Y}, x\in\mathcal{X} \tag{6}$$

Such bridge function h_0 exists when the proxy satisfies similar conditions.

Assumption A.1. We assume $W \perp\!\!\!\perp (A,Y)|U,X$.

Assumption A.2. For any square integrable function $l:\mathcal{U}\to\mathbb{R}$, the following conditions hold for all $a\in\mathcal{A}$ and $x\in\mathcal{X}$:

$$\mathbb{E}\left[l(U)\mid A=a,W=w,X=x\right]=0 \quad \forall w\in\mathcal{W} \quad \Leftrightarrow \quad l(u)=0 \quad \mathbb{P}\left[U\right]\text{-a.e.}$$

Assumption A.3. There exists a *deterministic* function $\gamma_0(a,x,u)$ such that $Y=\gamma_0(A,U,X)$ and $\|\gamma_0(a,x,U)\|_{\mathbb{P}[U]}\le\infty$ for all $a\in\mathcal{A},x\in\mathcal{X}$.

We can show the existence using a similar discussion to Theorem 3.1. The structural function is obtained by the partial average over the bridge function as $\mathbb{E}\left[h(\tilde{a},X,W)\right]=f_{\text{struct}}(\tilde{a})$ if we further assume either of the following:

Assumption A.4 (Tchetgen Tchetgen et al., 2023, Assumption 2(iii))**.** For test point $\tilde{a}\in\mathcal{A}$, we have $W\perp\!\!\!\perp A|X,Y^{(\tilde{a})}$.

Assumption A.5. For any square integrable function $l:\mathcal{U}\to\mathbb{R}$, the following conditions hold for test point $\tilde{a}\in\mathcal{A}$ and for any $x\in\mathcal{X}$:

$$\mathbb{E}\left[l(U)\mid A=\tilde{a},Y=y,X=x\right]=0 \quad \forall y\in\mathcal{Y} \quad \Leftrightarrow \quad l(u)=0 \quad \mathbb{P}\left[U\right]\text{-a.e.}$$

We may use either two-stage regression or the maximum moment restriction approach to obtain the bridge function. We state the closed-form solution for each approach.

Theorem A.6. *Given stage 1 samples* $\{w_i, a_i, y_i, x_i\}_{i=1}^n$ *and stage 2 samples* $\{\dot{a}_i, \dot{y}_i, \dot{x}_i\}_{i=1}^m$*, and regularizing parameter* (λ, η)*, the solution of two-stage regression approach is given as*

$$\hat{h}(a, w, x) = \boldsymbol{\alpha}^\top \boldsymbol{k}(a, w, x), \quad \boldsymbol{\alpha} = (M + m\eta I)^{-1}\dot{\boldsymbol{y}}$$

where $\dot{\boldsymbol{y}} = (\dot{y}_1, \ldots, \dot{y}_m)^\top \in \mathbb{R}^m$*, and* $M \in \mathbb{R}^{m\times m}$ *and* $\boldsymbol{k}(a, w) \in \mathbb{R}^m$ *are defined as*

$$\begin{aligned}
&M = K_{\dot{A}\dot{A}} \odot (B^\top K_{WW} B) \odot K_{\dot{X}\dot{X}},\\
&\boldsymbol{k}(a, w, x) = \boldsymbol{k}_{\dot{A}}(a) \odot (B^\top \boldsymbol{k}_W(w)) \odot \boldsymbol{k}_{\dot{X}}(x)\\
&B = (K_{AA} \odot K_{YY} \odot K_{XX} + n\lambda I)^{-1}(K_{A\dot{A}} \odot K_{Y\dot{Y}} \odot K_{X\dot{X}}).
\end{aligned}$$

Here, for $F \in \{A, W, Y, X\}$*, we denote the stage 1 kernel matrix as* $K_{FF} = (k_{\mathcal{F}}(f_i, f_j))_{ij} \in \mathbb{R}^{n\times n}$ *where* $k_{\mathcal{F}}$ *and* f_i *are the corresponding space and stage 1 samples. Similarly, for* $\dot{F} \in \{\dot{A}, \dot{Y}, \dot{X}\}$*, we denote the stage 2 kernel matrix as* $K_{\dot{F}\dot{F}} = (k_{\mathcal{F}}(\dot{f}_i, \dot{f}_j))_{ij} \in \mathbb{R}^{n\times n}$ *where* $k_{\mathcal{F}}$ *and* f_i *are the corresponding space and stage 2 samples. We further denote* $K_{F\dot{F}} = (k_{\mathcal{F}}(f_i, \dot{f}_j))_{ij} \in \mathbb{R}^{n\times m}$ *and* $\boldsymbol{k}_{\dot{A}}(a) = (k(\dot{a}_i, a))_i \in \mathbb{R}^m, \boldsymbol{k}_W(w) = (k(w_i, w))_i \in \mathbb{R}^n, \boldsymbol{k}_{\dot{X}}(s) = (k(\dot{x}_i, x))_i \in \mathbb{R}^m$

Theorem A.7. *Given samples* $\{w_i, a_i, y_i, x_i\}_{i=1}^n$*, the solution of maximum moment restriction approach is given as* $\hat{h}(a, w) = \boldsymbol{\alpha}^\top \boldsymbol{k}(a, w)$*, where weight* $\boldsymbol{\alpha} \in \mathbb{R}^n$ *is given as*

$$\boldsymbol{\alpha} = \sqrt{G}\left(\sqrt{G}L\sqrt{G} + n^2\eta I\right)^{-1}\sqrt{G}\boldsymbol{y}.$$

Here, $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ *and*

$$\begin{aligned}
&L = K_{AA} \odot K_{WW} \odot K_{XX}, \quad G = K_{AA} \odot K_{YY} \odot K_{XX},\\
&\boldsymbol{k}(a, w) = (k_{\mathcal{A}}(a_i, a)k_{\mathcal{W}}(w_i, w)k_{\mathcal{X}}(x_i, x))_i \in \mathbb{R}^n
\end{aligned}$$

with $K_{FF} = (k_{\mathcal{F}}(f_i, f_j))_{ij} \in \mathbb{R}^{n\times n}$ *for* $F \in \{W, A, Y, X\}$ *and corresponding kernel function* $k_{\mathcal{F}}$ *and samples* f_i*, and* $\sqrt{G}$ *denotes the square root of* $G = \sqrt{G}\sqrt{G}$*.*

The consistency of each estimation can be proved by a similar discussion presented in Proposition G.3 and Proposition G.5.

B Connection to PCL

In Proxy Causal Learning (PCL), we assume access to a treatment-inducing proxy variable Z, and an outcome-inducing proxy variable W, which satisfy the following *structural assumption* and *completeness assumption*.

Assumption B.1 (Structural Assumption (Deaner, 2018; Mastouri et al., 2021))**.** We assume $Y \perp\!\!\!\perp Z|A, U$, and $W \perp\!\!\!\perp (A, Z)|U$.

Assumption B.2 (Completeness Assumption on Confounder (Deaner, 2018; Mastouri et al., 2021))**.** Let $l : \mathcal{U} \to \mathbb{R}$ be any square integrable function $\|l\|_{\mathbb{P}[U]} < \infty$. The following conditions hold for all $a \in \mathcal{A}$

$$\begin{aligned}
&\mathbb{E}\left[l(U) \mid A = a, W = w\right] = 0 \quad \forall w \in \mathcal{W} \quad \Leftrightarrow \quad l(u) = 0 \quad \mathbb{P}\left[U\right]\text{-a.e.}\\
&\mathbb{E}\left[l(U) \mid A = a, Z = z\right] = 0 \quad \forall z \in \mathcal{Z} \quad \Leftrightarrow \quad l(u) = 0 \quad \mathbb{P}\left[U\right]\text{-a.e.}
\end{aligned}$$

Here, A, Y, U are the treatment, the outcome, and the unobserved confounder, respectively. We show that given Assumptions 2.2, 2.3 and G.1 and informative outcome assumption (3), these PCL assumptions hold with $Z = Y$.

Since Y is deterministic given A, U (Assumption 2.3), we have

$$Y \perp\!\!\!\perp Y|A, U$$

since the deterministic variables are independent of the deterministic variables by definition. Also, from Assumption 2.1, we have $W \perp\!\!\!\perp (Z, Y)|U$. Therefore, we have Assumption B.1. Furthermore, the first line of Assumption B.2 is equivalent to Assumption 2.2, and the second line of Assumption B.2 corresponds to informative outcome assumption (3) with $Z = Y$.

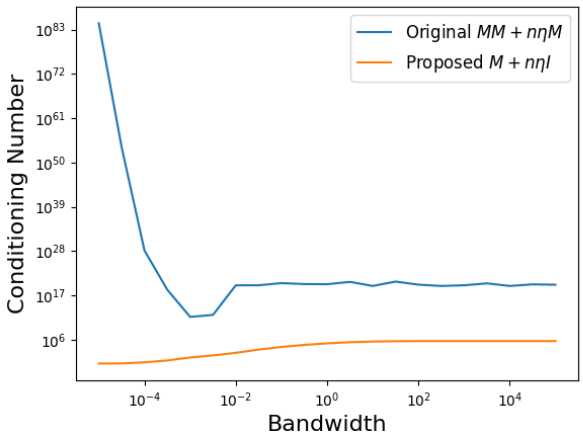


Figure 6: Conditioning number for different SKPV formulations. We tested on our deterministic low-dimensional proxy setting with 1000 samples.

From this, we show that the outcome Y can be used as the proxy Z in PCL. We use this connection to derive the functional equation for bridge function (1) and its relation to structural function f_{struct}. In PCL, the bridge function h_0 is defined as the solution of

$$\mathbb{E}\left[h_0(a, W)|A, Z\right] = \mathbb{E}\left[Y|A, Z\right],$$

and we have $f_{\text{struct}}(a) = \mathbb{E}_W\left[h_0(a, W)\right]$. Now, if we replace the proxy Z to the outcome Y, the right-hand side becomes just Y, and the left-hand side is $\mathbb{E}\left[h_0(a, W)|A, Y\right]$, which gives the definition in our bridge function (1). From this, we can derive the identification results (Theorem 3.3) by using a similar discussion to the PCL identification.

C Numerical Stability

To illustrate the numerical stability of our solution, we plot the conditioning numbers with respect to the various bandwidths of the Gaussian kernel in our synthetic experimental setting in Figure 6. This shows matrix to be inverted in our solution has a smaller conditioning number (orange) than that in the original solution (blue). We observed similar phenomena in SPMMR.

D Use of Neural Network Features

A number of works (Xu et al., 2021; Kompa et al., 2022; Kallus et al., 2021) proposed the deep learning methods for proxy causal learning, which can be applied to our settings by using the connection discussed in Appendix B. We expect these methods to achieve better empirical performances than kernel-based methods, especially when the data is high-dimensional as discussed in the original PCL setting (Xu et al., 2021; Kompa et al., 2022; Kallus et al., 2021). We first discuss the two-stage regression approach using neural nets, and then introduce the neural version of the maximum moment restriction approach.

Two-stage Regression Approach We can apply a similar method as DFPV (Xu et al., 2021) to derive the neural version of the two-stage regression approach. In DFPV, we model the bridge function $h_0(a, w)$ as

$$h(a, w) = \boldsymbol{u}^\top(\boldsymbol{\psi}_{\theta_{A(2)}}(a) \otimes \boldsymbol{\psi}_{\theta_W}(w)),$$

where $\boldsymbol{u}$ is weight parameters, $\otimes$ is the tensor product $\boldsymbol{a} \otimes \boldsymbol{b} = \text{vec}(ab^\top)$ and $\boldsymbol{\psi}_{\theta_{A(2)}}, \boldsymbol{\psi}_{\theta_W}$ are neural nets parametrized by $\theta_{A(2)}, \theta_W$, respectively. We minimize the loss $\mathcal{L}^{\text{2SR}}$ to learn these parameters.

$$\mathcal{L}^{\text{2SR}}(h) = \mathbb{E}_{AY}\left[(Y - \boldsymbol{u}^\top(\boldsymbol{\psi}_{\theta_{A(2)}}(a) \otimes \mathbb{E}\left[\boldsymbol{\psi}_{\theta_W}(W)|A, Y\right]))\right]$$

Similar to the SKPV, we need to model the conditional feature mean $\mathbb{E}\left[\boldsymbol{\psi}_{\theta_W}(W)|A,Y\right]$ to minimize the loss $\mathcal{L}^{\mathrm{2SR}}$. We model this using the different neural net feature maps.

$$\mathbb{E}_{W|A=a,Y=y}\left[\boldsymbol{\psi}_{\theta_W}(W)\right] = \boldsymbol{V}(\boldsymbol{\phi}_{\theta_{A(1)}}(a) \otimes \boldsymbol{\phi}_{\theta_Y}(y)),$$

where $\boldsymbol{V}$ are parameters, and $\boldsymbol{\phi}_{\theta_{A(1)}}, \boldsymbol{\phi}_{\theta_Z}$ are neural nets parametrized by $\theta_{A(1)}, \theta_Z$, respectively. Note we may use different neural nets in the treatment features $\boldsymbol{\phi}_{\theta_{A(1)}}$ and $\boldsymbol{\psi}_{\theta_{A(2)}}$.

As in SKPV, we learn $\mathbb{E}_{W|a,y}\left[\boldsymbol{\psi}_{\theta_W}(W)\right]$ in stage 1 regression and $h_0(a,w)$ in stage 2 regression, but in addition to the weights $\boldsymbol{u}$ and $\boldsymbol{V}$, we also learn the parameters of the feature maps.

Specifically, in stage 1, we learn $\boldsymbol{V}$ and parameters $\theta_{A(1)}, \theta_Z$ by minimizing the following empirical loss:

$$\hat{\mathcal{L}}^{\mathrm{cond}}(\boldsymbol{V}, \theta_{A(1)}, \theta_Z) = \frac{1}{m}\sum_{i=1}^{m} \left\| \boldsymbol{\psi}_{\theta_W}(w_i) - \boldsymbol{V}\left(\boldsymbol{\phi}_{\theta_{A(1)}}(a_i) \otimes \boldsymbol{\phi}_{\theta_Y}(y_i)\right)\right\|^2 + \lambda\|\boldsymbol{V}\|^2.$$

Although $\hat{\mathcal{L}}^{\mathrm{cond}}$ depends on θ_W, at this stage, we do not optimize θ_W with $\hat{\mathcal{L}}^{\mathrm{cond}}$ as $\boldsymbol{\psi}_{\theta_W}(w)$ is the "target variable" in stage 1 regression. Given the minimizers $(\hat{\boldsymbol{V}}, \hat{\theta}_{A(1)}, \hat{\theta}_Z) = \arg\min \hat{\mathcal{L}}_1$, we learn weights $\boldsymbol{u}$ and parameters $\theta_W, \theta_{A(2)}$ by minimizing the empirical stage 2 loss,

$$\hat{\mathcal{L}}^{\mathrm{2SR}}(\boldsymbol{u}, \theta_W, \theta_{A(2)}) = \frac{1}{n}\sum_{i=1}^{n}\left(\dot{y}_i - \boldsymbol{u}^\top\left(\boldsymbol{\psi}_{\theta_{A(2)}}(\dot{a}_i) \otimes \hat{\boldsymbol{V}}\left(\boldsymbol{\phi}_{\hat{\theta}_{A(1)}}(\dot{a}_i) \otimes \boldsymbol{\phi}_{\hat{\theta}_Y}(\dot{y}_i)\right)\right)\right)^2 + \eta\|\boldsymbol{u}\|^2.$$

Although the expression of $\hat{\mathcal{L}}^{\mathrm{2SR}}$ does not explicitly contain θ_W, it implicitly depends on θ_W as $(\hat{\boldsymbol{V}}, \hat{\theta}_{A(1)}, \hat{\theta}_Z)$ is the solution of a minimization problem involving θ_W. To deal with this implicit dependency, we may use the method proposed in Xu et al. (2021), in which we ignore the dependency of θ_W on parameters $\hat{\theta}_{A(1)}, \hat{\theta}_Y$, and compute the gradient via the closed-form solution of $\hat{V}$.

This gives the following learning procedure. First, we fix parameters in the adaptive feature maps $(\theta_{A(1)}, \theta_Y, \theta_{A(2)}, \theta_W)$. Given these parameters, optimal weights $\hat{\boldsymbol{V}}, \hat{\boldsymbol{u}}$ can be learned by minimizing the empirical stage 1 loss $\hat{\mathcal{L}}^{\mathrm{cond}}$ and empirical stage 2 loss $\hat{\mathcal{L}}^{\mathrm{2SR}}$, respectively. These minimizations can be solved analytically, where the solutions are

$$\hat{\boldsymbol{V}}(\boldsymbol{\theta}) = \Psi_1^\top \Phi_1 (\Phi_1^\top \Phi_1 + m\lambda_1 I)^{-1}, \qquad \hat{\boldsymbol{u}}(\boldsymbol{\theta}) = \left(\Phi_2^\top \Phi_2 + n\lambda_2 I\right)^{-1} \Phi_2^\top \boldsymbol{y}_2, \tag{7}$$

where we denote $\boldsymbol{\theta} = (\theta_{A(1)}, \theta_Y, \theta_{A(2)}, \theta_W)$ and define matrices as follows:

$$\begin{aligned}
&\Psi_1 = \left[\boldsymbol{\psi}_{\theta_W}(w_1), \ldots, \boldsymbol{\psi}_{\theta_W}(w_m)\right]^\top, && \Phi_1 = \left[\boldsymbol{v}_1(a_1, y_1), \ldots, \boldsymbol{v}_1(a_m, y_m)\right]^\top,\\
&\boldsymbol{y}_2 = \left[\dot{y}_1, \ldots, \dot{y}_n\right]^\top, && \Phi_2 = \left[\boldsymbol{v}_2(\dot{a}_1, \dot{y}_1), \ldots, \boldsymbol{v}_2(\dot{a}_n, \dot{y}_n)\right]^\top,\\
&\boldsymbol{v}_1(a,y) = \boldsymbol{\phi}_{\theta_{A(1)}}(a) \otimes \boldsymbol{\phi}_{\theta_Y}(y), && \boldsymbol{v}_2(a,y) = \boldsymbol{\psi}_{\theta_{A(2)}}(a) \otimes \left(\hat{\boldsymbol{V}}(\boldsymbol{\theta})\left(\boldsymbol{\phi}_{\theta_{A(1)}}(a) \otimes \boldsymbol{\phi}_{\theta_Y}(y)\right)\right).
\end{aligned}$$

Given these weights $\hat{\boldsymbol{u}}(\boldsymbol{\theta}), \hat{\boldsymbol{V}}(\boldsymbol{\theta})$, we can update feature parameters by a gradient descent method with respect to the residuals of the loss of each stage, while regrading $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{u}}$ as functions of parameters. Specifically, we take the gradient of the losses

$$\begin{aligned}
\hat{\mathcal{L}}^{\mathrm{cond}}(\boldsymbol{\theta}) &= \frac{1}{m}\sum_{i=1}^{m}\left\|\boldsymbol{\psi}_{\theta_W}(w_i) - \hat{\boldsymbol{V}}(\boldsymbol{\theta})\left(\boldsymbol{\phi}_{\theta_{A(1)}}(a_i) \otimes \boldsymbol{\phi}_{\theta_Y}(y_i)\right)\right\|^2 + \lambda\|\hat{\boldsymbol{V}}(\boldsymbol{\theta})\|^2,\\
\hat{\mathcal{L}}^{\mathrm{2SR}}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\left(\dot{y}_i - \hat{\boldsymbol{u}}(\boldsymbol{\theta})^\top\left(\boldsymbol{\psi}_{\theta_{A(2)}}(\dot{a}_i) \otimes \hat{\boldsymbol{V}}(\boldsymbol{\theta})\left(\boldsymbol{\phi}_{\theta_{A(1)}}(\dot{a}_i) \otimes \boldsymbol{\phi}_{\theta_Y}(\dot{y}_i)\right)\right)\right)^2 + \eta\|\hat{\boldsymbol{u}}(\boldsymbol{\theta})\|^2,
\end{aligned}$$

where $\hat{\boldsymbol{V}}(\boldsymbol{\theta}), \hat{\boldsymbol{u}}(\boldsymbol{\theta})$ are given in (7). Given these losses, $(\theta_{A(1)}, \theta_Y)$ are minimized with respect to $\hat{\mathcal{L}}^{\mathrm{cond}}$, and $(\theta_{A(2)}, \theta_W)$ are minimized with respect to $\hat{\mathcal{L}}^{\mathrm{2SR}}$. Xu et al. (2021) reported that the learning procedure is stabilized by running several gradient descent steps on the stage 1 parameters $(\theta_{A(1)}, \theta_Y)$ before updating the stage 2 features $(\theta_{A(2)}, \theta_W)$.

Maximum Moment Restriction In maximum moment restriction, we minimize the following loss to obtain the bridge function

$$\hat{h} = \arg\min_h \max_g (\mathbb{E}\left[(Y - h(A, W))g(A, Y)\right])^2.$$

In SPMMR, we assume both h and g to be in RKHS. There are two approaches to extend this to the neural network model. One is to model h as a neural network while still assuming g is in RKHS, which is considered in Kompa et al. (2022). The other is to model both h, g as neural network functions as in Kallus et al. (2021).

If we assume g to be the RKHS function, we can learn the bridge function by minimizing the same loss $\mathcal{L}^{\mathrm{MMR}}$ to SPMMR.

$$\hat{h} = \arg\min_h \mathbb{E}\left[\Delta_{Y,A,W}(h)\Delta_{Y',A',W'}(h)k_{\mathcal{A}}(A, A')k_{\mathcal{Y}}(Y, Y')\right], \tag{8}$$

where $\Delta_{Y,A,W}(h) = (Y - h(A, W))$ and (W', A', Y') are independent copies of (W, A, Y), respectively. Empirically, we may add regularizer Ω and minimize the following loss.

$$\hat{h} = \arg\min_h \frac{1}{n^2} \sum_{i,j=1}^{n} \Delta_i(h)\Delta_j(h)k_{\mathcal{A}}(a_i, a_j)k_{\mathcal{Y}}(y_i, y_j) + \eta\Omega(h)$$

Kompa et al. (2022) suggested to use the L^2-penalty on the parameter in h as the regularizer.

When we model g as a neural net function, we may still obtain the bridge function h by solving (8). However, Kallus et al. (2021) proposed the following loss, which is more computationally efficient.

$$\hat{h} = \arg\min_h \max_g \frac{1}{n} \sum_{i=1}^{n} (y_i - h(a_i, w_i))g(a_i, y_i) - \frac{\lambda}{n} \sum_{i=1}^{n} g^2(a_i, y_i)$$

Note that the objective of minimax problem can be computed in $O(n)$ time, while the original loss (8) requires $O(n^2)$. The second term is called stabilizer (Kallus et al., 2021). Kallus et al. (2021) showed that stabilizers are different from regularizers, and generally should not let it vanish when the sample size grows.

E Details of Identifiability

In this appendix, we prove the propositions given in the main text.

E.1 Existence of bridge function

Here, we present the proof of Theorem 3.1. We introduce the following operator Let us consider the following operators:

$$E_a : L^2(P_{W|A=a}) \to L^2(P_{Y|A=a}),\ E_a f \coloneqq \mathbb{E}\left[f(W)|A = a, Y = \cdot\right],$$
$$F_a : L^2(P_{Y|A=a}) \to L^2(P_{W|A=a}),\ F_a g \coloneqq \mathbb{E}\left[g(Y)|A = a, W = \cdot\right],$$

Our goal is to show that $I(y) = y$ is in the range of E_a, i.e., we seek a solution to the inverse problem defined by

$$E_a h = I(y). \tag{9}$$

This suffices to prove the existence of the function h_a for if there exists a function h_a^* for each $a \in \mathcal{A}$ such that

$$\mathbb{E}\left[h_a^*|A = a, Y = \cdot\right] = y,$$

we can define $h_0(a, w) \coloneqq h_a^*(w)$. We apply the discussion similar to Xu et al. (2021, Appendix B), which ensures the existence using the following theorem.

Proposition E.1 (Kress, 1999, Theorem 15.18)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Hilbert spaces. Let $E : \mathcal{X} \to \mathcal{Y}$ be a compact linear operator with singular system $\{(\mu_n, \varphi_n, g_n)\}_{n=1}^{\infty}$. The equation of the first kind*

$$E\varphi = f$$

is solvable if and only if $f \in N(E^*)^{\perp}$ and

$$\sum_{n=1}^{\infty} \frac{1}{\mu_n^2} |\langle f, g_n \rangle|^2 < \infty.$$

Here, $N(E^*)$ denotes the null space of the operator E^*. Then a solution is given by

$$\phi = \sum_{n=1}^{\infty} \frac{1}{\mu_n} \langle f, g_n \rangle \varphi_n.$$

To apply Proposition E.1, we make the following additional assumptions.

Assumption E.2. For each $a \in \mathcal{A}$, the operator E_a is compact with singular system $\{(\mu_{a,n}, \varphi_{a,n}, g_{a,n})\}_{n=1}^{\infty}$.

Assumption E.3. The identity map $I(y)$ satisfies

$$\sum_{n=1}^{\infty} \frac{1}{\mu_{a,n}^2} |\langle I, g_{a,n} \rangle_{L^2(P_{Y|A=a})}|^2 < \infty,$$

for a singular system $\{(\mu_{a,n}, \phi_{a,n}, g_{a,n})\}_{n=1}^{\infty}$ given in Assumption E.2.

Please refer to Xu et al. (2021, Remark 1) for the discussion on these assumptions. It is easy to see that Assumptions E.2 and E.3 are required for using Proposition E.1. The remaining condition to show is that $I(y)$ is in $N(E_a^*)^{\perp}$. We show that the structural assumption (Assumption 2.1), completeness assumption (Assumption 2.2), and deterministic confounding (Assumption 2.3) imply the required condition.

Lemma E.4. *Under Assumptions 2.1 to 2.3, the identity $I(y)$ is in the orthogonal complement of the null space $N(E_a^*)$.*

Proof. We first show that the adjoint of E_a is given by F_a. For the operator E_a, any $f \in L_2(P_{W|A=a})$ and $g \in L_2(P_{Y|A=a})$, we have

$$\begin{aligned}
\langle E_a f, g \rangle_{L_2(P_{Y|A=a})} &= \mathbb{E}_{Y|A=a} \left[\mathbb{E} \left[f(W) | A = a, Y \right] g(Y) \right] \\
&= \mathbb{E}_{Y|A=a} \left[\mathbb{E}_{U|A=a,Y} \left[\mathbb{E} \left[f(W) | A = a, Y, U \right] \right] g(Y) \right] \\
&\overset{(a)}{=} \mathbb{E}_{Y|A=a} \left[\mathbb{E}_{U|A=a,Y} \left[\mathbb{E} \left[f(W) | A = a, U \right] \right] g(Y) \right] \\
&= \mathbb{E}_{U,Y|A=a} \left[\mathbb{E} \left[f(W) | A = a, U \right] g(Y) \right] \\
&= \mathbb{E}_{U|A=a} \left[\mathbb{E} \left[f(W) | A = a, U \right] \mathbb{E} \left[g(Y) | A = a, U \right] \right]
\end{aligned}$$

where (a) follows from $W \perp\!\!\!\perp (A, Y) | U$, which is from Assumption 2.1. Similarly,

$$\begin{aligned}
\langle f, F_a g \rangle_{L^2(P_{W|A=a})} &= \mathbb{E}_{W|A=a} \left[f(W) \mathbb{E} \left[g(Y) | A = a, W \right] \right] \\
&= \mathbb{E}_{W|A=a} \left[f(W) \mathbb{E}_{U|A=a,W} \left[\mathbb{E} \left[g(Y) | A = a, W, U \right] \right] \right] \\
&\overset{(b)}{=} \mathbb{E}_{W|A=a} \left[f(W) \mathbb{E}_{U|A=a,W} \left[\mathbb{E} \left[g(Y) | A = a, U \right] \right] \right] \\
&= \mathbb{E}_{W,U|A=a} \left[f(W) \mathbb{E} \left[g(Y) | A = a, U \right] \right] \\
&= \mathbb{E}_{U|A=a} \left[\mathbb{E} \left[f(W) | A = a, U \right] \mathbb{E} \left[g(Y) | A = a, U \right] \right] \\
&= \langle E_a f, g \rangle_{L_2(P_{Y|A=a})}.
\end{aligned}$$

Again, (b) is given by $W \perp\!\!\!\perp A, Y | U$ from Assumption 2.1. For any $f^* \in N(E_a^*) = N(F_a)$, by iterated expectations, we have

$$\begin{aligned}
0 &= \mathbb{E} \left[f^*(Y) | A = a, W = \cdot \right] \\
&= \mathbb{E}_U \left[\mathbb{E} \left[f^*(Y) | A, U, W \right] | A = a, W = \cdot \right] \\
&= \mathbb{E}_U \left[f^*(\gamma_0(a, U)) | A = a, W = \cdot \right]. \quad (\because \text{Assumption 2.3}) \qquad (10)
\end{aligned}$$

From Assumption 2.2,

$$\mathbb{E}\left[l(U) \mid A=a, W=w\right]=0 \quad \forall w \in \mathcal{W} \quad \Leftrightarrow \quad l(u)=0 \quad P_U\text{-a.e.}$$

for all functions $l \in L^2(P_{U|A=a})$. Hence, (10) and Assumption 2.2 implies

$$f^*(\gamma_0(a,U))=0 \quad P_U\text{-a.s.}$$

Then, the inner product between f^* and $I(y)$ is given as follows:

$$\begin{aligned}\langle f^*, I\rangle_{L^2(P_{Y|A=a})} &= \mathbb{E}_{Y|A=a}\left[f^*(Y)Y\right] \\ &= \mathbb{E}_{U|A=a}\left[\mathbb{E}_{Y|U,A=a}\left[f^*(Y)Y\right]\right] \\ &\overset{(c)}{=} \mathbb{E}_{U|A=a}\left[f^*(\gamma_0(a,U))\gamma_0(a,U)\right] \\ &= 0.\end{aligned}$$

Here, (c) uses the fact that Y is deterministic given (a,u). Hence, we have

$$I \in N(E_a^*)^{\perp}.$$

□

Now, we are able to apply Proposition E.1, leading to the Theorem 3.1.

Proof of Theorem 3.1. By Lemma E.4, the identity $I(y)$ is in $N(E_a^*)^{\perp}$. Therefore, by Proposition E.1, under the given assumptions, there exists a solution to (9). Letting the solution be h_a^* completes the proof. □

E.2 Identifiability in deterministic confounding

Here, we show that the bridge function h_0 can be used to compute the structural function.

Proof of Theorem 3.3. First, we show that under Assumption 2.3, we have

$$\mathbb{E}\left[\gamma_0(\tilde{a},U)|A=\tilde{a},Y=y\right]=y. \tag{11}$$

Let $B(\tilde{a},y)=\{u\in\mathcal{U}|\gamma(\tilde{a},u)=y\}$. Then, we have

$$\begin{aligned}\mathbb{E}\left[\gamma_0(\tilde{a},U)|A=\tilde{a},Y=y\right] &= \int \gamma_0(\tilde{a},U)\mathrm{d}P(U|A=\tilde{a},Y=y) \\ &= \int_{\neg B(\tilde{a},y)} \gamma_0(\tilde{a},U)\mathrm{d}P(U|A=\tilde{a},Y=y) + \int_{B(\tilde{a},y)} \gamma_0(a,U)\mathrm{d}P(U|A=\tilde{a},Y=y) \\ &= \int_{\neg B(\tilde{a},y)} \gamma_0(\tilde{a},U)\mathrm{d}P(U|A=\tilde{a},Y=y) + y\int_{B(\tilde{a},y)} \mathrm{d}P(U|A=\tilde{a},Y=y)\end{aligned}$$

Now, from Bayes' theorem,

$$P(U|A=\tilde{a},Y=y) \propto P(Y=y|U,A=\tilde{a})P(U|A=\tilde{a}).$$

Since Y is generated deterministically, the support of $P(Y=y|U,A=\tilde{a})$ is $B(\tilde{a},y)$. Therefore,

$$\int_{\neg B(\tilde{a},y)} \gamma_0(\tilde{a},U)\mathrm{d}P(U|A=\tilde{a},Y=y)=0, \quad \int_{B(\tilde{a},y)} \mathrm{d}P(U|A=\tilde{a},Y=y)=1.$$

Hence, we have (11). Using this, we have

$$\begin{aligned}&\mathbb{E}\left[h_0(\tilde{a},W)|A=\tilde{a},Y=y\right]-y=0 \\ \Leftrightarrow\quad &\mathbb{E}_{U|A=\tilde{a},Y=y}\left[\mathbb{E}\left[h_0(\tilde{a},W)|U\right]\right]-\mathbb{E}_{U|A=\tilde{a},Y=y}\left[\gamma_0(\tilde{a},U)\right]=0 \quad (\because (11), \text{Assumption 2.1})\end{aligned}$$

If we assume informative outcome assumption (3), this means

$$\mathbb{E}\left[h_0(\tilde{a}, W)|U\right] = \gamma_0(\tilde{a}, U),$$

and thus

$$f_{\text{struct}}(\tilde{a}) = \mathbb{E}_U\left[\gamma_0(\tilde{a}, U)\right] = \mathbb{E}_U\left[\mathbb{E}\left[h_0(\tilde{a}, W)|U\right]\right] = \mathbb{E}\left[h_0(\tilde{a}, W)\right].$$

We can draw the same conclusion if we instead assume $A \perp\!\!\!\perp W|Y^{(\tilde{a})}$ as

$$\begin{aligned}
\mathbb{E}\left[h(\tilde{a}, W)\right] &\overset{(a)}{=} \mathbb{E}_{Y^{(\tilde{a})}}\left[\mathbb{E}\left[h(\tilde{a}, W)|Y^{(\tilde{a})}, A = \tilde{a}\right]\right] \\
&\overset{(b)}{=} \int \mathbb{E}\left[h(\tilde{a}, W)|Y = y, A = \tilde{a}\right] P(Y^{(\tilde{a})} = y)\mathrm{d}y \\
&\overset{(c)}{=} \mathbb{E}\left[Y^{(\tilde{a})}\right],
\end{aligned}$$

where we used $A \perp\!\!\!\perp W|Y^{(\tilde{a})}$ in (a), "consisitency" $\mathbb{E}\left[h(\tilde{a}, W)|Y^{(\tilde{a})} = y, A = \tilde{a}\right] = \mathbb{E}\left[h(\tilde{a}, W)|Y = y, A = \tilde{a}\right]$ in (b), and $\mathbb{E}\left[h(\tilde{a}, W)|A = \tilde{a}, Y = y\right] = y$ in (c). □

E.3 Identifiability in stochastic counfounding

In this section, we derive the error bound in the stochastic confounding.

Proof. The proof is similar to Theorem 3.3. First, we show that if additive noise is bounded $|\varepsilon| < M$, we have

$$\left|\mathbb{E}\left[\gamma_0(\tilde{a}, U)|A = \tilde{a}, Y = y\right] - y\right| < M \quad \forall y \in \mathcal{Y}. \tag{12}$$

Let $B(\tilde{a}, y, \varepsilon) = \{u \in \mathcal{U} \mid |\gamma(\tilde{a}, u) - y| \leq M\}$. Then, we have

$$\begin{aligned}
\mathbb{E}\left[\gamma_0(\tilde{a}, U)|A = \tilde{a}, Y = y\right] &= \int \gamma_0(\tilde{a}, U)\mathrm{d}P(U|A = \tilde{a}, Y = y) \\
&= \int_{\neg B(\tilde{a}, y)} \gamma_0(\tilde{a}, U)\mathrm{d}P(U|A = \tilde{a}, Y = y) + \int_{B(\tilde{a}, y)} \gamma_0(\tilde{a}, U)\mathrm{d}P(U|A = \tilde{a}, Y = y)
\end{aligned}$$

Now, from Bayes' theorem,

$$P(U|A = \tilde{a}, Y = y) \propto P(Y = y|U, A = \tilde{a})P(U|A = \tilde{a}).$$

Since $|\varepsilon| < M$, the support of $P(Y = y|U, A = \tilde{a})$ is $B(\tilde{a}, y)$. Therefore,

$$\int_{\neg B(\tilde{a}, y)} \gamma_0(\tilde{a}, U)\mathrm{d}P(U|A = \tilde{a}, Y = y) = 0, \quad \int_{B(\tilde{a}, y)} \mathrm{d}P(U|A = \tilde{a}, Y = y) = 1$$

Furthermore, we have

$$\begin{aligned}
\int_{B(\tilde{a}, y)} \gamma_0(\tilde{a}, U)\mathrm{d}P(U|A = \tilde{a}, Y = y) &\geq \int_{B(\tilde{a}, y)} (y - M)\mathrm{d}P(U|A = \tilde{a}, Y = y) \\
&= y - M.
\end{aligned}$$

Similarly, we can also show

$$\int_{B(\tilde{a}, y)} \gamma_0(\tilde{a}, U)\mathrm{d}P(U|A = \tilde{a}, Y = y) \leq y + M.$$

Hence, we have (12). Using this, we have

$$\begin{aligned}
& \mathbb{E}\left[h_0(\tilde{a}, W)|A = \tilde{a}, Y = y\right] = y \\
\Leftrightarrow \quad & \left|\mathbb{E}_{U|A=\tilde{a}, Y=y}\left[\mathbb{E}\left[h_0(\tilde{a}, W)|U\right]\right] - \mathbb{E}_{U|A=\tilde{a}, Y=y}\left[\gamma_0(\tilde{a}, U)\right]\right| \leq M \quad (\because (12), \text{Assumption 2.1}) \\
\Leftrightarrow \quad & \left|\mathbb{E}\left[h_0(\tilde{a}, W)|U\right] - \gamma_0(\tilde{a}, U)\right| \leq M\Xi \quad (\because (4))
\end{aligned}$$

Hence, if we solve equation (1), we have

$$\begin{aligned} f_{\text{struct}}(\tilde{a}) &= \mathbb{E}_U\left[\gamma_0(\tilde{a}, U)\right] \\ &\leq \mathbb{E}_U\left[\mathbb{E}\left[h_0(\tilde{a}, W)|U\right] + M\Xi\right] \\ &= \mathbb{E}_W\left[h_0(\tilde{a}, W)\right] + M\Xi. \end{aligned}$$

Similarly, we have $f_{\text{struct}}(\tilde{a}) \geq \mathbb{E}_W\left[h_0(\tilde{a}, W)\right] - M\Xi$. □

F Derivation

Two-stage Regression Here, we present the proof of Theorem 4.1. First, we recall the empirical estimate of conditional mean embedding as follows.

Proposition F.1 ([Song et al., 2009, Theorem 5]). *Given stage 1 samples $\{w_i, a_i, y_i\}$, the solution of stage 1 regression is given as*

$$\hat{\mu}_{W|A,Y}(a, y) = \sum_{i=1}^{n} \beta_i(a, y)\phi(w_i), \quad \boldsymbol{\beta}(a, y) = (K_{AA} \otimes K_{YY} + n\lambda I)^{-1}(\boldsymbol{k}_A(a) \odot \boldsymbol{k}_Y(y))$$

where $\boldsymbol{k}_{\mathcal{A}}(a) = (k(a_i, a))_i \in \mathbb{R}^n, \boldsymbol{k}_{\mathcal{Y}}(y) = (k(y_i, y))_i$.

Given this, we consider stage 2 regression

$$\hat{h} = \underset{h \in \mathcal{H}_{\mathcal{A}\mathcal{W}}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left\langle h, \phi_{\mathcal{A}}(a_i) \otimes \hat{\mu}_{W|A,Y}(a_i, y_i)\right\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{W}}}\right)^2 + \eta \|h\|^2_{\mathcal{H}_{\mathcal{A}\mathcal{W}}}$$

can be seen as the kernel ridge regression with the kernel function $\tilde{k}$ is given as

$$\begin{aligned} \tilde{k}((a, y), (a', y')) &= \left\langle \phi_{\mathcal{A}}(a) \otimes \hat{\mu}_{W|A,Y}(a, y), \phi_{\mathcal{A}}(a') \otimes \hat{\mu}_{W|A,Y}(a', y')\right\rangle_{\mathcal{H}_{\mathcal{A}\mathcal{W}}} \\ &= k_{\mathcal{A}}(a, a')\left(\boldsymbol{\beta}^\top(a, y) K_{WW} \boldsymbol{\beta}(a', y')\right). \end{aligned}$$

Using the closed-form solution of kernel ridge regression (See (Murphy, 2013, Chapter 14.4.3)) yields Theorem 4.1.

Equivalence to Mastouri et al. (2021) Mastouri et al. (2021) considered a different way of deriving the closed-form of the two-stage regression.

Proposition F.2 (Mastouri et al. (2021, Proposition 2)). *Given stage 1 samples $\{w_i, a_i, y_i\}_{i=1}^n$ and stage 2 samples $\{\dot{a}_i, \dot{y}_i\}_{i=1}^m$, and regularizing parameter (λ, η), we have*

$$\hat{h}(a, w) = \boldsymbol{k}_W^\top(w) \Gamma \boldsymbol{k}_{\dot{A}}(a), \quad \Gamma \in \mathbb{R}^{n \times m}, \quad \text{vec}(\Gamma) = (B \bar{\otimes} I)(M + mn\eta I)^{-1} \dot{\boldsymbol{y}} \in \mathbb{R}^{nm}$$

where $\bar{\otimes}$ is the tensor product of associated columns of matrices with the same number of columns, and vec *is row-wise vectorization as* $\text{vec}(\begin{bmatrix} a & b \\ c & d \end{bmatrix}) = [a, b, c, d]^\top$. *The definitions of $\boldsymbol{k}_W, \boldsymbol{k}_{\dot{A}}, B, M$ can be found in Theorem 4.1.*

Although this learns mn parameters, we show this solution is equivalent to Theorem 4.1 restated as follows.

$$\hat{h}(a, w) = \boldsymbol{\alpha}^\top(\boldsymbol{k}_{\dot{A}}(a) \odot B^\top \boldsymbol{k}_W(w)), \quad \boldsymbol{\alpha} = (M + mn\eta I)^{-1} \dot{\boldsymbol{y}} \in \mathbb{R}^m$$

From the definition of $\bar{\otimes}$ and $\boldsymbol{\alpha}$, we have

$$\mathrm{vec}(\Gamma) = \begin{bmatrix} B_{11} & & & & \\ & B_{12} & & & \\ & & \ddots & & \\ & & & B_{1n} & \\ B_{21} & & & & \\ & B_{22} & & & \\ & & \ddots & & \\ & & & B_{2n} & \\ & & \vdots & & \\ B_{m1} & & & & \\ & B_{m2} & & & \\ & & \ddots & & \\ & & & B_{mn} & \end{bmatrix} \boldsymbol{\alpha}.$$

Hence, from definition of vec, we have

$$\Gamma = \begin{bmatrix} (\boldsymbol{B}_{1\cdot} \odot \boldsymbol{\alpha})^\top \\ (\boldsymbol{B}_{2\cdot} \odot \boldsymbol{\alpha})^\top \\ \vdots \\ (\boldsymbol{B}_{n\cdot} \odot \boldsymbol{\alpha})^\top \end{bmatrix} \in \mathbb{R}^{n \times m},$$

where $\boldsymbol{B}_{i\cdot}$ is the i-th row vector of matrix B. Therefore, Proposition F.2 can be written as

$$\hat{h}(a, w) = \sum_{i=1}^{m} k_W(w_i, w)(\boldsymbol{B}_{i\cdot} \odot \boldsymbol{\alpha})^\top \boldsymbol{k}_{\dot{A}}.$$

From linearity, we have

$$\begin{aligned} \hat{h}(a, w) &= \sum_{i=1}^{m} k_W(w_i, w)(\boldsymbol{B}_{i\cdot} \odot \boldsymbol{\alpha})^\top \boldsymbol{k}_{\dot{A}} \\ &= \left(\sum_{i=1}^{m} \boldsymbol{B}_{i\cdot} k_W(w_i, w) \odot \boldsymbol{\alpha} \right)^\top \boldsymbol{k}_{\dot{A}} \\ &= (B^\top \boldsymbol{k}_W(w) \odot \boldsymbol{\alpha})^\top \boldsymbol{k}_{\dot{A}} \\ &= (B^\top \boldsymbol{k}_W(w) \odot \boldsymbol{k}_{\dot{A}})^\top \boldsymbol{\alpha} \end{aligned}$$

The last equality holds for all vectors $(\boldsymbol{a} \odot \boldsymbol{b})^\top \boldsymbol{c} = (\boldsymbol{a} \odot \boldsymbol{c})^\top \boldsymbol{b} = \sum_i a_i b_i c_i$. Hence, we showed Proposition F.2 is equivalent to Theorem 4.1.

Maximum Moment Restriction Here, we present the proof of Theorem 4.2. Let $\boldsymbol{h} \in \mathbb{R}^n = (h(a_i, w_i))_i \in \mathbb{R}^n$. Then, we can rewrite empirical loss $\hat{\mathcal{L}}^{\mathrm{MMR}}$ as

$$\begin{aligned} \hat{h} &= \arg\min_h \frac{1}{n^2} \boldsymbol{h}^\top G \boldsymbol{h} - \frac{2}{n^2} \boldsymbol{y}^\top G \boldsymbol{h} + \frac{1}{n^2} \boldsymbol{y}^\top G \boldsymbol{y} + \eta \|h\|_{\mathcal{H}_{\mathcal{A}\mathcal{W}}} \\ &= \arg\min_h \frac{1}{n} (\sqrt{G/n}\boldsymbol{y} - \sqrt{G/n}\boldsymbol{h})^\top (\sqrt{G/n}\boldsymbol{y} - \sqrt{G/n}\boldsymbol{h}) + \eta \|h\|_{\mathcal{H}_{\mathcal{A}\mathcal{W}}} \\ &= \arg\min_h \frac{1}{n} \sum_{i=1}^{n} ((\sqrt{G/n}\boldsymbol{y})_i - (\sqrt{G/n}\boldsymbol{h})_i)^2 + \eta \|h\|_{\mathcal{H}_{\mathcal{A}\mathcal{W}}}. \end{aligned}$$

Here, we denote $(\boldsymbol{x})_i$ as the i-th element of vector $\boldsymbol{x}$. From reproducing characteristic, we have

$$(\sqrt{G/n}\boldsymbol{h})_i = \left\langle \tilde{\phi}_i, h \right\rangle, \quad \tilde{\phi}_i = \sum_{j=1}^{n} (\sqrt{G/n})_{ij} \phi_{\mathcal{A}\mathcal{W}}(a_j, w_j),$$

where $(\sqrt{G/n})_{ij}$ is the (i,j)-element of $\sqrt{G/n}$. Hence, we can regard this as kernel ridge regression, where the target is $(\sqrt{G/n}\boldsymbol{y})_i$ and the feature is $\tilde{\phi}_i$. Since

$$\left\langle \tilde{\phi}_i, \tilde{\phi}_j \right\rangle_{\mathcal{H}_{\mathcal{AW}}} = (\sqrt{G/n} L \sqrt{G/n})_{ij},$$

we have the closed-form solution as $\hat{h} = \boldsymbol{\beta}^\top \tilde{\boldsymbol{k}}(a,w)$

$$\begin{aligned}
\boldsymbol{\beta} &= (\sqrt{G/n} L \sqrt{G/n} + n\eta I)^{-1} \sqrt{G/n}\boldsymbol{y} \\
&= \sqrt{n}(\sqrt{G} L \sqrt{G} + n^2 \eta I)^{-1} \sqrt{G}\boldsymbol{y} \\
\tilde{\boldsymbol{k}}(a,w) &= \begin{bmatrix} \left\langle \tilde{\psi}_1, \phi_{\mathcal{AW}}(a,w) \right\rangle_{\mathcal{H}_{\mathcal{AW}}} \\ \left\langle \tilde{\psi}_2, \phi_{\mathcal{AW}}(a,w) \right\rangle_{\mathcal{H}_{\mathcal{AW}}} \\ \vdots \\ \left\langle \tilde{\psi}_n, \phi_{\mathcal{AW}}(a,w) \right\rangle_{\mathcal{H}_{\mathcal{AW}}} \end{bmatrix} \\
&= \sqrt{G/n}\boldsymbol{k}(a,w)
\end{aligned}$$

This yields the closed-form solution in Theorem 4.2.

G Consistency

We derive the consistency in this section. The proof is mostly adopted from Singh (2020) and Mastouri et al. (2021).

Two-stage Regression To derive the consistency of two-stage regression, we make the following assumptions.

Assumption G.1. Let operator C_μ be $C_\mu = \mathbb{E}\left[\phi_2(A,Y) \otimes \phi_2(A,Y)\right]$ for $\phi_2(A,Y) = \phi_{\mathcal{A}}(A) \otimes \mu_{W|A,Y}(A,Y) \in \mathcal{H}_{\mathcal{AW}}$ and its eigen-decomposition be $\{\eta_i, \phi_i\}$. We assume η_i, the i-th eigenvalue, can be bounded by $\eta_i \leq C i^{-b_0}$ for constant C, b_0. Furthermore, we assume that there exists $g \in \mathcal{H}_{\mathcal{AW}}$ such that $h_0 = C_\mu^{\frac{c_0-1}{2}} g$ for some $c_0 \in (1,2]$.

This assumption is made in the analysis of various kernel-based methods (Singh, 2020; Mastouri et al., 2021; Fischer and Steinwart, 2020). The constant b_0 is known as the effective dimension of $\mathcal{H}_{\mathcal{AW}}$, and $b_0 = \infty$ when RKHS $\mathcal{H}_{\mathcal{AW}}$ is of finite dimension. The constant c_0 measures the smoothness of the bridge function: a larger c_0 means a smoother bridge function. We assume a similar condition on operator $C_{W|A,Y}$.

Assumption G.2. Let the covariance operator of $\mathcal{H}_{\mathcal{AY}}$ be $C_{(A,Y),(A,Y)} = \mathbb{E}\left[\phi_{\mathcal{AY}}(A,Y) \otimes \phi_{\mathcal{AY}}(A,Y)\right]$. We assume η_i, the i-th eigenvalue of $C_{(A,Y),(A,Y)}$, can be bounded by $\eta_i \leq C i^{-b_1}$ for constant C, b_1. Furthermore, we assume that there exists $G \in \mathcal{H}_{\mathcal{W}(\mathcal{AY})}$ that $C_{W|A,Y} = G \circ C_{(A,Y),(A,Y)}^{\frac{c_1-1}{2}}$ for some $c_1 \in (1,2]$.

Here, we denote * as the adjoint of the operator. Given these assumptions, we obtain the following consistency result.

Proposition G.3. *Assume Assumptions G.1 and G.2. Given stage 1 samples $\{w_i, a_i, y_i\}_{i=1}^n$ and stage 2 samples $\{\dot{a}_i, \dot{y}_i\}_{i=1}^m$, and let $\lambda = n^{-\frac{1}{c_0+1/b_0}}$ and $n = m^{\xi \frac{c_0+1/b_0}{c_0-1}}$ for some $\xi > 0$. Then,*

$$\sup_{a,w} |h_0(a,w) - \hat{h}(a,w)| = O_P(m^{-\frac{\xi c_1 - 1}{2(c_1+1)}}),$$

with $\eta = m^{-\frac{\xi}{c_1+3}}$ when $\xi \leq \frac{c_1+3}{c_1+1/b_1}$. Otherwise, we have

$$\sup_{a,w} |h_0(a,w) - \hat{h}(a,w)| = O_P(m^{-\frac{c_1-1}{2(c_1+1/b_1)}}),$$

with $\eta = m^{-\frac{1}{c_1+1/b_1}}$. These upper bounds can be translated to the estimation error of structural function as

$$|f_{\text{struct}}(\tilde{a}) - \hat{f}_{\text{struct}}(\tilde{a})| \leq \sup_w |h_0(\tilde{a},w) - \hat{h}(\tilde{a},w)| + O_P(n^{-\frac{1}{2}}),$$

under the condition of Theorem 3.3.

The proof is a direct application of Theorem 3 in Singh (2020). From this, we can see that SKPV can consistently recover the bridge function and structural function. As in Singh (2020), we observe that the best data ratio is $n = m^{\frac{c_1+3}{c_1+1/b_1}\frac{c_0+1/b_0}{c_0-1}}$ from which $n \gg m$. The best rate for bridge function estimation under this data size ratio is $O_P(m^{-\frac{c_1-1}{2(c_1+1/b_1)}})$ which is the same as for kernel ridge regression (Fischer and Steinwart, 2020).

Maximum Moment Restriction For maximum moment restriction approach, we use the same assumption as Mastouri et al. (2021, Assumption 16) .

Assumption G.4. Let $\{\eta_i, \phi_i, \psi_i\}$ be a singular decomposition of operator $T = \mathbb{E}\left[\phi_{\mathcal{AW}}(A, W) \otimes \phi_{\mathcal{AY}}(A, Y)\right]$. We assume h_0 is in a γ-regularity space $\mathcal{H}^{\gamma}_{\mathcal{AW}}$ that is defined as

$$\mathcal{H}^{\gamma}_{\mathcal{AW}} = \left\{ f \in \mathcal{H}_{\mathcal{AW}} \;\middle|\; \sum_{i=1}^{\infty} \frac{\langle f, \phi_i \rangle^2_{\mathcal{H}_{\mathcal{AW}}}}{\eta_i^{2\gamma}} \right\}$$

This assumption is similar to Assumption G.1: a larger γ means a smoother bridge function. Given this, we can show the consistency of the SPMMR method.

Proposition G.5. *Given Assumption G.4 and samples $\{w_i, a_i, y_i\}$ size of n. Then, we have*

$$\|\hat{h} - h_0\|_{\mathcal{H}_{\mathcal{AW}}} \leq O\left(n^{-\frac{1}{2}\min\left(\frac{1}{2}, \frac{\gamma}{\gamma+2}\right)}\right)$$

with $\lambda = n^{-\frac{1}{2}\min\left(\frac{1}{2}, \frac{2}{\gamma+2}\right)}$.

The proof is the direct application of Mastouri et al. (2021, Theorem 3) . From this, we can see that the best rate $O(n^{-1/4})$ is achieved when $\gamma \geq 2$. Again, this best rate is the same as kernel ridge regression or SKPV. Furthermore, Mastouri et al. (2021) discussed the relation of the two-stage regression approach and maximum moment restriction approach in the asymptotic case; this relation also applies to our SKPV and SPMMR.

H Experiment Details

In this section, we provide the details of experiments with the regularization parameter tuning. All experiments are run on MacBook Air within a few minutes.

Regularization Parameter Tuning for SKPV and SPMMR We follow the procedure in Mastouri et al. (2021) to tune the regularization parameter (λ, η) for SKPV and λ in SPMMR. In SKPV, to select the best λ in stage 1 regression, we use leave-one-out error for $\hat{\mathcal{L}}^{\text{cond}}$. This can be computed as the closed-form solution

$$\hat{\mathcal{L}}^{\text{cond}-\text{loo}}(\lambda) = \text{tr}(\tilde{H}_1^{-1} H_1 K_{WW} H_1 \tilde{H}_1^{-1})$$

where $H_1 = I - (K_{AA} \odot K_{YY})(K_{AA} \odot K_{YY} + n\lambda I)^{-1}$, and $\tilde{H}_1$ is the diagonal matrix with the same diagonals H_1. Given the best parameter $\lambda = \arg\min \hat{\mathcal{L}}^{\text{cond}-\text{loo}}$, we can tune the second stage regularizer η by using the stage 1 data as "validation data" in stage 2 loss:

$$\eta = \arg\min \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left\langle \hat{h}, \phi_{\mathcal{A}}(a_i) \otimes \hat{C}_{W|A,Y} \phi_{\mathcal{AY}}(a_i, y_i) \right\rangle_{\mathcal{H}_{\mathcal{AW}}} \right)^2,$$

where $\hat{h}$ is the solution in Theorem 4.1. In SPMMR, we cannot use leave-one-out error to tune λ in SPMMR since $\hat{\mathcal{L}}^{\text{MMR}}$ is a V-statistic. We selected the best λ by minimizing the empirical loss $\hat{\mathcal{L}}^{\text{MMR}}$ over a held-out validation set.

Additive Noise Increases the Support In the sensitivity analysis on the deterministic assumption, we observe that adding Gaussian noise $\mathcal{N}(0, 1)$ can increase the performance. To explore this, we plot the data samples of treatment A and outcome Y in Appendix H. From it, we can see that the true structural function is completely away from the support when noise is zero. However, it is in the well-supported domain when we add the Gaussian noise.

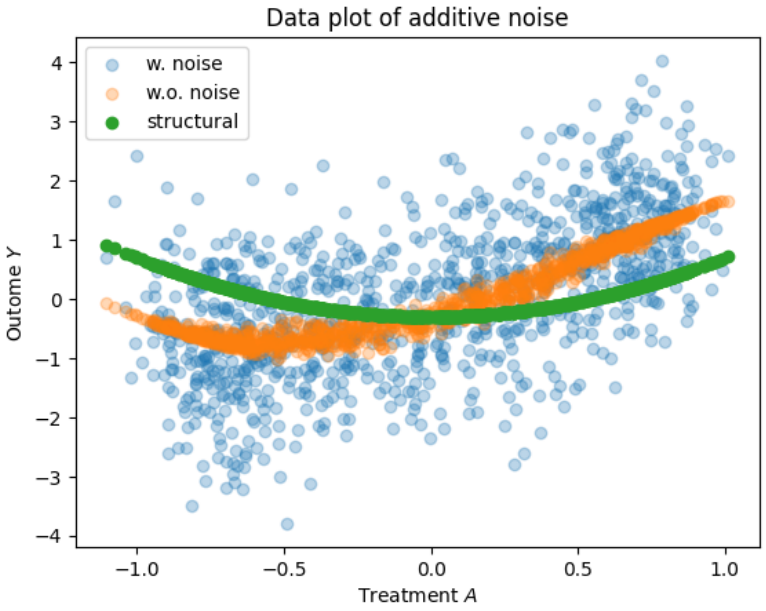


Figure 7: The plot of data samples with additive noise (blue) and without additive noise (orange). Green curve is true structural function $Y = f_{\text{struct}}(A)$