Spectral Representation for Causal Estimation with Hidden Confounders

Tongzheng Ren
UT Austin
haotian.sun@gatech.edu

Haotian Sun
Georgia Tech
haotian.sun@gatech.edu

Antoine Moulin
Universitat Pompeu Fabra
antoine.moulin@upf.edu

Arthur Gretton
University College London & Google DeepMind
arthur.gretton@gmail.com

Bo Dai
Georgia Tech & Google DeepMind
bodai@cc.gatech.edu

Abstract

We address the problem of causal effect estimation where hidden confounders are present, with a focus on two settings: instrumental variable regression with additional observed confounders, and proxy causal learning. Our approach uses a singular value decomposition of a conditional expectation operator, followed by a saddle-point optimization problem, which, in the context of IV regression, can be thought of as a neural net generalization of the seminal approach due to Darolles et al. [2011]. Saddle-point formulations have gathered considerable attention recently, as they can avoid double sampling bias and are amenable to modern function approximation methods. We provide experimental validation in various settings, and show that our approach outperforms existing methods on common benchmarks.

1 Introduction

We consider the problem of estimating causal effects that arise in various disciplines, including economics, epidemiology, and social sciences. The presence of unobserved confounders, *i.e.*, variables that affect both the cause and the effect, poses a significant challenge to traditional estimation methods as they introduce spurious correlations that lead to biased and inconsistent estimates [Stock and Watson, 2007]). A workaround for dealing with hidden confounders involves having access to additional variables that can help identify the object of interest. Two instances of this idea, *instrumental variable (IV) regression* [Wright, 1928, Stock and Trebbi, 2003] and *proxy causal learning (PCL)* Kuroki and Pearl [2014], have emerged as powerful tools.

IV regression involves solving an ill-posed inverse problem, $Ef = r$, where f is the causal effect, r is the expected output conditioned on an instrument, and E is the associated conditional expectation operator. It is ill-posed as solving this equation in f (assuming a solution exists) would typically involve the inverse eigenvalues of E, which can be arbitrarily close to zero. To ensure a stable solution, the target is assumed to satisfy a source condition [Engl et al., 1996], which relates to its smoothness relative to E. Existing methods for solving this inverse problem can be categorized into two-stage estimation methods [Newey and Powell, 2003, Darolles et al., 2011, Chen and Christensen, 2018, Singh et al., 2019] and conditional moment methods [Dai et al., 2017, Dikkala et al., 2020, Liao et al., 2020, Bennett et al., 2019, 2023b,c].

These methods require strong assumptions on the function classes, such as realizability ($f_0 \in \mathcal{F}$) and closedness (stability under the operator E and/or its adjoint $E^\star$) [Dikkala et al., 2020, Liao et al., 2020, Bennett et al., 2023a,c]. Bennett et al. [2023b] avoid the closedness assumption but require a specific source condition and an additional realizability assumption on the dual function class $\mathcal{G}$. The assumptions required for IV can be restrictive. Previous works have considered an extension

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

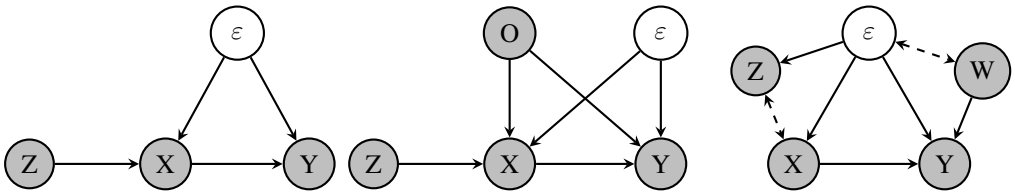
of IV regression to accommodate observable confounders [Horowitz, 2011, Xu et al., 2020]. In the absence of a valid instrument, proxy causal learning (PCL) assumes access to proxy variables containing relevant information on the confounder. Kuroki and Pearl [2014] and Miao et al. [2018] provided necessary conditions on the proxy variables for identifying the true causal effect. Recent methods have proposed estimation techniques for PCL, using fixed feature dictionaries [Deaner, 2018, Mastouri et al., 2021] and adaptive feature dictionaries [Xu et al., 2021, Kompa et al., 2022].

We propose an approach based on a low-rank assumption on conditional densities, inspired by reinforcement learning [Jin et al., 2020]. This assumption yields function classes satisfying realizability and closedness conditions, enabling optimization over finite-dimensional variables. We leverage these to derive efficient algorithms for IV regression with observed confounders and PCL. Section 2 formalizes the settings and explains the derivation of saddle-point problems. Section 3 introduces the low-rank assumption and its application. Our method can learn an adaptive basis for IV with observed confounders and PCL, unlike previous works focusing solely on IV. Section 3 discusses a representation learning algorithm inspired by Wang et al. [2022]. Section 5 provides experimental validation, demonstrating that our approach outperforms existing methods on IV and PCL benchmarks.

2 Preliminaries

We formalize the three settings of interest and provide the equivalent saddle-point problems.

Notation. For a random variable x, we let P_x be its probability distribution and $L_2(P_x)$ be the associated L_2 space. Given n samples, we denote $\mathbb{E}_n$ the expectation with respect to the empirical measure. We denote the Euclidean inner product as $\langle \cdot, \cdot \rangle$ and the range of an operator E as $\mathcal{R}(E)$.



(a) Instrumental Variable (b) IV with Observable Confounder (c) Proxy Causal Inference

Figure 1: Causal graphs considered, with grey node denoting the observable variable.

2.1 Causal Estimation with Hidden Confounders

Given random variables x, s and y, in IV with and without observed confounders, we aim to find a function f that satisfies an equation of the form

$$\mathbb{E}[y - f(x, s) \,|\, s] = 0, \quad \mathbb{P}(s) - \text{a.e.} \tag{1}$$

Here, y denotes the outcome, x is the treatment, and s contains the side information one can access. IV regression and PCL differ in their assumptions and their applicability, as we now discuss.

IV regression is used when the confounder linearly affects the output. As shown in Figure 1a, it involves using an *instrument* z that (i) is independent of the output y conditional on the input x and the confounder ε, and (ii) is such that $\mathbb{E}[\varepsilon \,|\, z] = 0$. Here, $s = z$ in (1), and since the instrument does not affect the outcome, we only consider functions of x. We aim to find a function f such that

$$\mathbb{E}[y - f(x) \,|\, z] = 0, \quad \mathbb{P}(z) - \text{a.e.} \tag{2}$$

In the presence of an additional observed confounder o in IV, the information available becomes $s = (z, o)$ (Figure 1b) and we are interested in solving the equation

$$\mathbb{E}[y - f(x, o) \,|\, z, o] = 0, \quad \mathbb{P}(z, o) - \text{a.e.}, \tag{3}$$

where the input space of f includes the observed confounder because of its effect on the outcome.

PCL [Miao et al., 2018, Deaner, 2018] uses two *proxies* that are correlated with the unobserved confounder ε. One proxy z correlates with the treatment x, while the other proxy w correlates with the outcome y (Figure 1c). The proxies are assumed to satisfy the independence properties $z \perp\!\!\!\perp y \,|\, (x, \varepsilon)$ and $w \perp\!\!\!\perp (z, x) \,|\, \varepsilon$. Then, $s = (z, w)$ and we consider a slightly different equation in f

$$\mathbb{E}[y - f(x, w) \,|\, x, z] = 0, \quad \mathbb{P}(x, z) - \text{a.e.} \tag{4}$$

With those inverse problems at hand, we must derive the optimization problems to solve.

2.2 Primal-Dual Framework for Causal Estimation

We now derive a saddle-point formulation from Equation (1). Similar derivations apply to PCL. Previous works considered minimizing the mean-squared error, plus a regularizer Ω to account for the ill-posedness of the problem

$$\min_{f\in\mathcal{F}} \mathbb{E}\left[\mathbb{E}\left[y - f\left(x,s\right)|\, s\right]^2\right] + \lambda\Omega\left(f\right) := \mathcal{E}\left(f\right), \tag{5}$$

where $\mathcal{F} \subset L_2\left(P_{xs}\right)$ is convex, and Ω is a strongly convex regularizer, which can be a Tikhonov regularization [Darolles et al., 2011] or an RKHS norm [Singh et al., 2019, Zhang et al., 2023, Wang et al., 2022]. To derive the saddle-point problem, and following e.g. [Dai et al., 2017], we start with the Fenchel conjugate of the square function to write

$$\begin{aligned}
\mathcal{E}\left(f\right) &= \mathbb{E}\left[\max_{g\in\mathbb{R}}\left\{g\,\mathbb{E}\left[y - f\left(x,s\right)|\, s\right] - \frac{1}{2}g^2\right\}\right] + \lambda\Omega\left(f\right) \\
&= \max_{g\in L_2(P_s)} \mathbb{E}\left[g\left(s\right)\mathbb{E}\left[y - f\left(x,s\right)|\, s\right] - \frac{1}{2}g\left(s\right)^2\right] + \lambda\Omega\left(f\right) \qquad \text{(Dai et al. [2017])} \\
&= \max_{g\in L_2(P_s)} \mathbb{E}\left[g\left(s\right)\left(y - f\left(x,s\right)\right) - \frac{1}{2}g\left(s\right)^2\right] + \lambda\Omega\left(f\right). \qquad \text{(Tower rule)}
\end{aligned}$$

Therefore, given a convex class $\mathcal{G} \subset L_2\left(P_s\right)$, one can consider

$$\min_{f\in\mathcal{F}}\max_{g\in\mathcal{G}} \mathbb{E}\left[g\left(s\right)\left(y - f\left(x,s\right)\right) - \frac{1}{2}g\left(s\right)^2\right] + \lambda\Omega\left(f\right) := \mathcal{L}\left(f,g\right). \tag{6}$$

Note we obtain a strongly-convex-strongly-concave objective. Unlike $\mathcal{E}$, the absence of conditional expectations makes it straightforward to derive unbiased estimators of $\mathcal{L}$ or its derivatives and avoid the *double sampling bias* mentioned earlier. Denoting $\mathcal{L}_n$ as the empirical counterpart of $\mathcal{L}$, for any functions f, g, we have $\mathbb{E}\left[\mathcal{L}_n\left(f,g\right)\right] = \mathcal{L}\left(f,g\right)$.

In IV regression, the min-max problem is obtained by considering $s = z$ and $\mathcal{F} \subset L_2\left(P_x\right)$ in Problem (6), or $s = (z,o)$ and $\mathcal{F} \subset L_2\left(P_{xo}\right)$ when there is an observable confounder. For PCL, the min-max problem becomes

$$\min_{f\in\mathcal{F}}\max_{g\in\mathcal{G}} \mathbb{E}\left[g\left(x,z\right)\cdot\left(y - f\left(x,w\right)\right) - \frac{1}{2}g\left(x,z\right)^2\right] + \lambda\Omega\left(f\right). \tag{7}$$

For the estimate $\widehat{f} = \operatorname{argmin}_{f\in\mathcal{F}}\max_{g\in\mathcal{G}}\mathcal{L}_n\left(f,g\right)$ to be consistent, the classes $\mathcal{F}$ and $\mathcal{G}$ must be such that $f_0 = \operatorname{argmin}_{f\in\mathcal{F}}\max_{g\in\mathcal{G}}\mathcal{L}\left(f,g\right)$ is indeed a solution to the original inverse problem.

3 Characterizing the Function Classes

We first characterize the function classes $\mathcal{F}$ and $\mathcal{G}$ in IV regression under a low-rank assumption, and then discuss how to generalize this observation for the IV-OC and the PCL settings.

3.1 Instrument Variable Regression

We define the conditional expectation operator $E : L_2\left(P_x\right) \to L_2\left(P_z\right)$ as $Ef = \mathbb{E}\left[f\left(x\right)|\, z\right]$ for any $f \in L_2\left(P_x\right)$, and we assume Equation (2) has at least a solution.

Assumption 1. *We have $\mathbb{E}\left[y|\, z\right] \in \mathcal{R}\left(E\right)$, i.e., there exists $f_0 \in L_2\left(P_x\right)$ such that $Ef_0 = \mathbb{E}\left[y|\, z\right]$.*

Our goal is to find subspaces of $L_2\left(P_x\right)$ and $L_2\left(P_z\right)$ such that the saddle-point of Problem (6) provides a solution to Problem (2). We make the following low-rank assumption on the conditional distribution of x given z and the marginal of x.

Assumption 2. *The distributions P_x and $\left\{P_{x|z}\right\}_{z\in\mathcal{Z}}$ admit densities $P\left(\cdot\right)$ and $\left\{P\left(\cdot\,|\, z\right)\right\}_z$, respectively. Furthermore, there exist feature maps $\phi : \mathcal{X} \to \mathbb{R}^d$ and $\psi : \mathcal{Z} \to \mathbb{R}^d$ such that for any x, z,*

$$P\left(x|\, z\right) = P\left(x\right)\left\langle\phi\left(x\right),\psi\left(z\right)\right\rangle_{\mathbb{R}^d}. \tag{8}$$

Remark on Assumption 2. The first part is close to Assumption A.1 from Darolles et al. [2011] except that theirs is stated on the joint distribution instead. The second part of the assumption, *i.e.*, Equation (8), is akin to assuming the conditional expectation operator E admits a finite singular-value decomposition (SVD). It is known that compacity implies the existence of a countable SVD. If the spectrum of the operator E decays fast enough, it is meaningful to perform a finite-dimensional approximation [Ren et al., 2022a]. This assumption is also equivalent to the low-rank assumption made in reinforcement learning [Jin et al., 2020].

A consequence of Assumption 2 is that for any $f \in L_2(P_x)$, Equation 8 gives

$$Ef = \int f(x) P(x \mid z) \, \mathrm{d}x = \left\langle \psi(z), \int \phi(x) f(x) \, \mathrm{d}P_x(x) \right\rangle := \langle \psi(z), v_f \rangle . \tag{9}$$

By Assumption 1, there exists v_0 such that $\mathbb{E}[y \mid z] = \langle \psi(z), v_0 \rangle$. Thus, for a given f, the maximizer (over L_2) in Equation (6) can be written as $g_f^\star(z) = \langle \psi(z), v_0 - v_f \rangle$. This suggests the following.

Proposition 1 (Dual space for IV)**.** *Under Assumptions 1, 2, the dual embedding function for IV is realizable in*

$$\mathcal{G} := \left\{ z \mapsto \langle \psi(z), v \rangle, v \in \mathbb{R}^d \right\}. \tag{10}$$

Next, we introduce the operator $\Phi : u \mapsto \langle u, \phi(x) \rangle \in L_2(P_x)$ and define the class $\mathcal{F}$ as follows.

Proposition 2 (Primal space for IV)**.** *Under Assumptions 1, 2, we have* $\min_{f \in L_2(P_x)} \max_{g \in \mathcal{G}} \mathcal{L}(f, g) = \min_{f \in \mathcal{R}(\Phi)} \max_{g \in \mathcal{G}} \mathcal{L}(f, g)$ *and the following class contains a solution to Problem 2*

$$\mathcal{F} := \left\{ x \mapsto \langle \phi(x), u \rangle, u \in \mathbb{R}^d \right\}. \tag{11}$$

Compared to the parametrization of f with RKHS functions or deep neural networks [Dai et al., 2017, Muandet et al., 2020, Liao et al., 2020] for which realizability is assumed to hold, we consider a structural assumption with which we can explicitly characterize realizable function classes.

Remark on the parametrizations. The characterization of the representation for the target function through the spectral decomposition of $P(x \mid z)$ has been exploited in IV [Darolles et al., 2011, Wang et al., 2022] and reinforcement learning [Jin et al., 2020, Yang and Wang, 2020, Ren et al., 2022b]. In particular, the feature map $\phi(x)$ is exploited as in Proposition 2. A difference is that we also exploit $\psi(z)$ to characterize the dual space in Proposition 1 and plug it in the min-max problem.

We also note an important difference with Darolles et al., 2011, who approximate the conditional operator by estimating unconditional densities with kernel density estimators, and then using the ratio of the estimated densities. Instead, we propose a representation learning algorithm to directly learn the feature maps ϕ and ψ under Assumption 2. We next generalize the previous observations beyond IV, and return to the resulting optimization problem and algorithm in Section 4.

3.2 Instrument Variable Regression with Observable Confounding, Proxy Causal Learning

We begin with the assumption that Problem 3 has a solution.

Assumption 3. $\mathbb{E}[y \mid z, o] \in \mathcal{R}(E)$*, i.e., there exists* $f_0 \in L_2(P_{xo})$ *such that* $Ef_0 = \mathbb{E}[y \mid z, o]$.

Under this assumption, and generalizing from the original IV problem, we consider the operator $E : L_2(P_x \otimes P_o) \to L_2(P_{zo})$, defined for any f as $Ef = \mathbb{E}[f(x, o) \mid z, o]$. Hence, we cannot directly leverage the factorization in the previous section.

Assumption 4. *The distributions* P_x *and* $\left\{P_{x|z,o}\right\}_{z,o}$ *admit densities* $P(\cdot)$ *and* $\{P(\cdot \mid z, o)\}_{z,o}$*, respectively. Furthermore, there exist* $\phi : \mathcal{X} \to \mathbb{R}^{d_x}$*,* $\psi : \mathcal{Z} \to \mathbb{R}^{d_z}$*, and* $V : \mathcal{O} \to \mathbb{R}^{d_x \times d_z}$ *such that for any* x, z, o*,*

$$P(x \mid z, o) = P(x) \langle \phi(x), V(o) \psi(z) \rangle . \tag{12}$$

As before, notice that for any $f \in L_2(P_{xo})$,

$$Ef = \int P(x \mid z, o) f(x, o) \, \mathrm{d}x = \left\langle \psi(z), V(o)^\top \int \phi(x) f(x, o) \, \mathrm{d}P_x \right\rangle .$$

Given a fixed o, we apply the same argument as that of the proof of Proposition 2 for IV to $\int \phi(x) f(x, o) \, \mathrm{d}P_x$ to show that for any o, it is enough to consider $f(\cdot, o)$ in the span of Φ. That is, for any o, there exists $u(o)$ such that for any x, $f(x, o) = \phi(x)^\top u(o)$. However, it is still unclear what the function u looks like.

In general, we cannot find the representation for u solely through E as we did for IV. To see this, we turn to the special case that $f(x,o) = h(o)$ is independent of x, where we have that $\mathbb{E}[h(o) \mid z,o] = h(o)$. Clearly, the operator E has no information on $h(o)$, therefore $u(o)$.

Given a solution to Equation (3), we have $\mathbb{E}[f(x,o) \mid z,o] = \mathbb{E}[y \mid z,o]$, which implies that (i) the space for $u(o)$ should also lie in the space of o in RHS, and (ii) the space for z in both sides should be the same. Thus, we make the following assumption with a slight abuse of notation.

Assumption 5. *The distributions P_y and $\{P_{y|z,o}\}_{z,o}$ admit densities $P(\cdot)$ and $\{P(\cdot \mid z,o)\}_{z,o}$, respectively. Furthermore, there exist $\nu : \mathcal{Y} \to \mathbb{R}^{d_y}$, and $W : \mathcal{O} \to \mathbb{R}^{d_y \times d_z}$ such that for any y, z, o,*

$$P(y \mid z,o) = P(y) \langle \nu(y), W(o)\psi(z) \rangle, \tag{13}$$

where ψ is the feature map from Assumption 4.

With this additional assumption, we obtain the following class.

Proposition 3 (Primal space for IV-OC)**.** *Under Assumptions 3, 4, and 5, the following class contains solutions to Equation 3,*

$$\mathcal{F} := \left\{ (x,o) \mapsto \phi(x)^\top BQ(o)\beta, B \in \mathbb{R}^{d_x \times d_x}, \beta \in \mathbb{R}^{d_y} \right\}.$$

For any f, we still have a closed form for the maximizer, *i.e.*, $g_f^\star(z,o) = \mathbb{E}[y \mid z,o] - \mathbb{E}[f(x,o) \mid z,o]$, so we can follow the same argument than earlier.

Proposition 4 (Dual space for IV-OC)**.** *Under Assumptions 3, 4, and 5, the dual function of IV-OC is realizable in*

$$\mathcal{G} := \left\{ (z,o) \mapsto \psi(z)^\top V(o)^\top Q(o)\gamma, \gamma \in \mathbb{R}^{d_y} \right\}.$$

Remark (Representation for PCL): IV-OC and PCL share the same conditioning structure. Thus, the representation characterization for IV-OC can also be applied for PCL, which results in the primal and dual function space as

$$\begin{aligned} \mathcal{F} &:= \left\{ (x,w) \mapsto \phi(w)^\top BQ(x)\beta, B \in \mathbb{R}^{d_w \times d_w}, \beta \in \mathbb{R}^{d_x} \right\}, \\ \mathcal{G} &:= \left\{ (x,z) \mapsto \psi(z)^\top V(x)^\top Q(x)\gamma, \gamma \in \mathbb{R}^{d_x} \right\}, \end{aligned} \tag{14}$$

where the spectral representations come from similar factorizations

$$P(w \mid x,z) = P(w) \langle \phi(w), V(x)\psi(z) \rangle, \quad P(y \mid x,z) = P(y) \left\langle \nu(y), Q(x)^\top V(x)\psi(z) \right\rangle. \tag{15}$$

Remark (Connection to existing IV-OC and PCL parametrization): In [Deaner, 2018, Mastouri et al., 2021, Xu et al., 2021], the parametrization for f is $f(x,w) = \zeta^\top(\theta(x) \otimes \varphi(w))$, with $\theta(x)$ and $\varphi(w)$ either fixed feature dictionaries, or learned neural net feature dictionaries. This parametrization shares some similarity to (14), if we rewrite (14) as

$$f(x,w) = \phi(w)^\top BQ(x)\beta = \left\langle \beta \text{vec}(B), Q(x)^\top \otimes \phi(w)^\top \right\rangle.$$

As we illustrate, however, we obtain the features $Q(x)$ and $\phi(w)$ from a spectral viewpoint, which is different from the existing parametrization.

4 Causal Estimation with Spectral Representation

In this section, we introduce empirical algorithms based on our spectral representation.

4.1 Estimation of the primal and dual variables

With the *exact* spectral representation, we know which function classes to consider for both the primal and dual variables. By construction, they satisfy the assumptions required from previous works [Bennett et al., 2023b, Li et al., 2024] and thus our algorithm enjoys strong statistical guarantees, at least for IV regression.

In practice, this may not hold due to the fact that the representations are *learned from data* and possibly the wrong dimensionality, which induces an additional statistical error and misspecification error. While we leave a formal analysis for future work, it will be possible to control the former with the analysis from Wang et al. [2022], and deal with the misspecification as in Bennett et al. [2023b].

We now present the algorithm below, discuss identifiability in Appendix D.

For Instrument Variable Regression Given the representation $\widehat{\phi(x)}$ and $\widehat{\psi(z)}$ obtained by factorizing $P(x|z)$, we obtain $f(x)$ and $u(z)$ by solving the following min-max optimization problem,

$$\min_{v}\max_{w}\mathbb{E}_{x,y,z}\left[\left(w^{\top}\widehat{\psi}\left(z\right)\right)\left(y-v^{\top}\widehat{\phi}\left(x\right)\right)-\frac{1}{2}\left(w^{\top}\widehat{\psi}\left(z\right)\right)^{2}\right]+\lambda\Omega\left(v^{\top}\widehat{\phi}\left(x\right)\right), \tag{16}$$

which is convex w.r.t. v and concave w.r.t. w, therefore, the global optimum is achievable. Our solutions are $f(x)=v^{\top}\widehat{\phi}(x)$ and $u(z)=w^{\top}\widehat{\psi}(z)$.

For Instrument Variable Regression with Observable Confounding and Proxy Causal Learning Due to the essential equivalence between IV-OC and PCL, our derivation is based on IV-OC. We denote $\widehat{\phi}\left(x\right),\widehat{\psi}\left(z\right),\widehat{V}\left(o\right)\widehat{Q}\left(o\right)$ as the learned representations, however the derivation is also applicable to PCL. Unlike the case IV without observable confounding, we now have an additional feature of observable o. The parametrization for $f(x,o)=\phi\left(x\right)^{\top}BQ(o)\beta$ contains two parameters B and β, which will induces non-convexity, if we directly substitute into the min-max problem. To recover convexity, we consider the reparametrization

$$\phi\left(x\right)^{\top}BQ(o)\beta=\mathrm{tr}\left(\phi\left(x\right)^{\top}BQ(o)\beta\right)=\mathrm{tr}\left(\left(Q(o)^{\top}\otimes\phi\left(x\right)^{\top}\right)^{\top}\underbrace{\beta\mathrm{vec}\left(B\right)^{\top}}_{G}\right),$$

where the second equality comes from the property of Kronecker matrix-vector product

$$(C\otimes D)\mathrm{vec}(G)=\mathrm{vec}(CGD^{\top}).$$

Based on this reformulation, we restore the convexity and concavity in the min-max optimization problem:

$$\min_{G}\max_{w}\ \lambda\Omega\left(\left\langle G,\widehat{Q}(o)^{\top}\otimes\widehat{\phi}\left(x\right)^{\top}\right\rangle\right)+\mathbb{E}_{x,y,z}\left[\left(\widehat{\psi}\left(z\right)^{\top}\widehat{V}(o)^{\top}\widehat{Q}(o)w\right)\cdot\left(y-\left\langle G,\widehat{Q}(o)^{\top}\otimes\widehat{\phi}\left(x\right)^{\top}\right\rangle\right)\right] \tag{17}$$

With the solved V and w, our solutions are $f(x,o)=\left\langle G,\widehat{Q}(o)^{\top}\otimes\widehat{\phi}\left(x\right)^{\top}\right\rangle$ and $g(z,o)=\widehat{\psi}\left(z\right)^{\top}\widehat{V}(o)^{\top}\widehat{Q}(o)w$. We omit PCL, which has similar min-max to IV-OC. We provide the complete algorithm in Appendix B.

5 Experiments

We evaluate the empirical performance of the proposed SpecIV and SpecPCL and several modern methods for IV with and without observable confounders, as well as PCL. This evaluation is conducted on two datasets. Following Xu et al. [2020, 2021], we utilize the out-of-sample mean-square error (OOS MSE) as the metric for all test cases. Experiment details and setup can be found in Appendix F.

Baselines. For the IV regression, we contrast SpecIV with two methods with pre-specified features – the Kernel IV (KIV) Singh et al. [2019] and the Dual Embedding (DE) Dai et al. [2017]. Additionally, we evaluate several approaches that leverage deep neural networks for feature representation, namely, DFIV Xu et al. [2020], and DeepGMM Bennett et al. [2019].

Datasets. We conduct experiments on two datasets for both low- and high-dimensional IV and PCL settings: **1) dSprites** [Matthey et al., 2017]: This dataset comprises images determined by five latent parameters (`shape`, `scale`, `rotation`, `posX`, `posY`). Each 64×64 image serves as the treatment variable X. Following Xu et al. [2020], we keep `shape` fixed as `heart`, use `posY` as the hidden confounder, and the remaining latent variables as the instrument Z. We also introduce a high-dimensional setting where the instruments are mapped to a variable in $\mathbb{R}^{2352}$, as proposed in Bennett et al. [2019]. **2) Demand Design** [Hartford et al., 2017]: This synthetic benchmark for nonlinear IV regression aims to predict ticket demands Y given the ticket price P, in the presence of observable confounders: price sensitivity $S\in\{1,\ldots,7\}$ and year time $T\in[0,10]$. We introduce an unobservable confounder as correlated noise in P and Y, and set the fuel price C as the instrument. We map both P and S to high-dimensional variables in $\mathbb{R}^{784}$ using the function from Bennett et al. [2019], representing a challenging scenario where relevant variables must be estimated from noisy, high-dimensional data. Details about the data generation is presented in Appendix I and G.

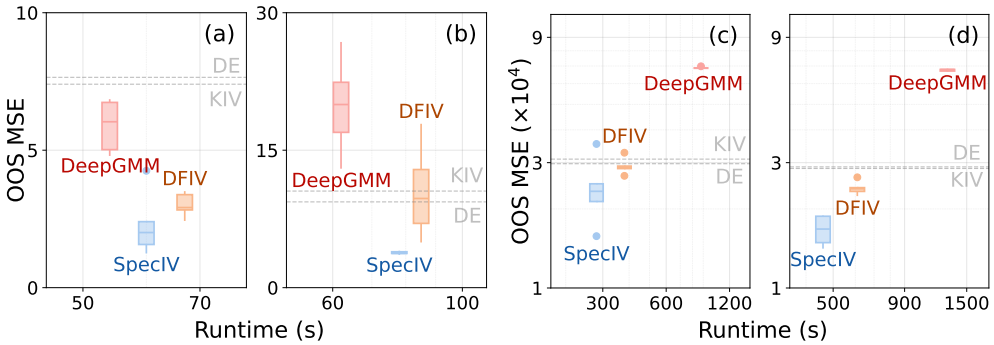


Figure 2: MSE and Runtime on dSprites Dataset with (a) Low-Dimensional Instruments (32) and (b) High-Dimensional Instruments (64), and Demand Design dataset with (c) 5,000 and (d) 10,000 training samples. Methods employing pre-specified features, *i.e.*, DE and KIV, are represented as dashed baselines.

Instrument Variable Regression. Figure 2a and 2b depict the performance and execution time of various methods on the dSprites dataset. Optimal performance corresponds to low MSE and reduced runtime (proximity to the bottom-left corner). SpecIV outperforms all other approaches in terms of MSE in both low- and high-dimensional contexts while maintaining a competitive runtime. Methods with fixed features, such as KIV and DE, yield high MSE due to their reliance on predetermined feature representations, which restricts adaptability. In low-dimensional scenarios, DeepGMM and DFIV manage to keep errors reasonable, but DeepGMM's performance declines sharply with increased feature dimensions, while SpecIV maintains good estimation accuracy. Both DeepGMM and DFIV exhibit increased variance in the high-dimensional setting, possibly due to ineffective representation learning for high-dimensional features. In summary, SpecIV demonstrates superior and consistent performance over the baseline methods in both settings for instrumental variable tasks.

Instrument Variable Regression with Observable Confounder. In the Demand Design dataset, we employ two settings for the observable confounders: For SpecIV and DFIV, we set the ticket price P as the treatment, the fuel price C as the instrument, and (T, S) as observables, each represented using distinct neural networks. For other methods that don't model observables separately, we integrate (T, S) directly with the treatment and instrument, using (P, T, S) as the treatment and (C, T, S) as the instrument. Figure 2c and Figure 2d show the performance of these approaches given different training sample sizes. In both scenarios, SpecIV consistently delivers the lowest error and best runtime. DeepGMM performs the least efficiently, possibly due to incorporating observables into both treatment and instrument, overlooking that we only need to consider the conditional expectation of X given (Z, O). KIV and DE, limited by their fixed feature representations, do not benefit from increased training samples and remain less expressive. Overall, SpecIV outperforms other approaches in learning structured functions with observable confounders.

Proxy Causal Learning We evaluate DFPV Xu et al. [2021], KPV Mastouri et al. [2023], PMMR Mastouri et al. [2023], and CEVAE Louizos et al. [2017]. Experimental results are shown in Figure 3. SpecPCL outperforms existing methods with superior estimation accuracy and computational efficiency, which demonstrates its strong capability in capturing complex structural functions. DFPV offers a reasonable balance between error and runtime but falls short of SpecIV's performance. KPV appears to yield a slightly lower error, possibly because KPV harnesses a greater number of parameters to express its solution. CEVAE, despite its flexibility through neural networks, underperforms all other methods. The reason could be that CEVAE does not leverage the relation between the proxies and the structural function.

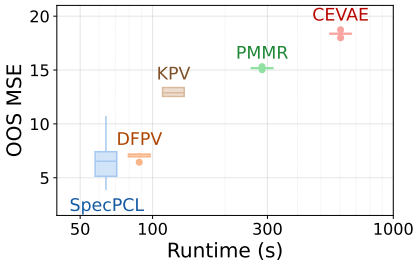


Figure 3: MSE and Runtime of structural function estimation on the dSprites dataset with 5,000 training samples.

6 Conclusion

We have introduced a novel spectral representation learning framework for causal estimation in the presence of hidden confounders. Our approach uses a low-rank assumption on conditional densities to characterize function classes within a saddle-point optimization problem, enabling efficient algorithms for IV regression (with and without observed confounders) and proximal causal learning (PCL). Extensive experiments validate that our method outperforms existing approaches.

References

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

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Inference on strongly identified functionals of weakly identified functions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2265–2265. PMLR, 2023a.

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Minimax instrumental variable regression and l_2 convergence guarantees without identification or closedness. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2291–2318. PMLR, 2023b.

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Source condition double robust inference on functionals of inverse problems. *arXiv preprint arXiv:2307.13793*, 2023c.

Xiaohong Chen and Timothy M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression: Nonlinear functionals of nonparametric iv. *Quantitative Economics*, 9(1):39–84, March 2018. ISSN 1759-7323. doi: 10.3982/qe722. URL http://dx.doi.org/10.3982/QE722.

Bo Dai, Niao He, Yunpeng Pan, Byron Boots, and Le Song. Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics*, pages 1458–1467. PMLR, 2017.

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

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

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

Joel L Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2): 347–394, 2011.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

Benjamin Kompa, David Bellamy, Tom Kolokotrones, Andrew Beam, et al. Deep learning methods for proximal inference via maximum moment restriction. *Advances in Neural Information Processing Systems*, 35:11189–11201, 2022.

Erwin Kreyszig. *Introductory functional analysis with applications*, volume 17. John Wiley & Sons, 1991.

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

Zihao Li, Hui Lan, Vasilis Syrgkanis, Mengdi Wang, and Masatoshi Uehara. Regularized deepiv with model selection. *arXiv preprint arXiv:2403.04236*, 2024.

Luofeng Liao, You-Lin Chen, Zhuoran Yang, Bo Dai, Mladen Kolar, and Zhaoran Wang. Provably efficient neural estimation of structural equation models: An adversarial approach. *Advances in Neural Information Processing Systems*, 33:8947–8958, 2020.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International conference on machine learning*, pages 7512–7523. PMLR, 2021.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt J. Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction, 2023.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

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

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. In *Advances in Neural Information Processing Systems*, volume 33, pages 2710–2721. Curran Associates, Inc., 2020.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Shuang Qiu, Lingxiao Wang, Chenjia Bai, Zhuoran Yang, and Zhaoran Wang. Contrastive ucb: Provably efficient contrastive self-supervised learning in online reinforcement learning. In *International Conference on Machine Learning*, pages 18168–18210. PMLR, 2022.

Tongzheng Ren, Chenjun Xiao, Tianjun Zhang, Na Li, Zhaoran Wang, Sujay Sanghavi, Dale Schuurmans, and Bo Dai. Latent variable representation for reinforcement learning. *arXiv preprint arXiv:2212.08765*, 2022a.

Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. *arXiv preprint arXiv:2208.09515*, 2022b.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

James Stock and Mark Watson. *Introduction to Econometrics 2nd edition*. Prentiss Hall, 2007.

James H Stock and Francesco Trebbi. Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194, 2003.

Ziyu Wang, Yucen Luo, Yueru Li, Jun Zhu, and Bernhard Schölkopf. Spectral representation learning for conditional moment models. *arXiv preprint arXiv:2210.16525*, 2022.

Philip Green Wright. *The tariff on animal and vegetable oils*. Number 26. Macmillan, 1928.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*, 2020.

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

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1), 2023.

Tianjun Zhang, Tongzheng Ren, Mengjiao Yang, Joseph Gonzalez, Dale Schuurmans, and Bo Dai. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pages 26447–26466. PMLR, 2022.

A Limitations and Broader Impacts

A.1 Limitations

In this paper, we introduced a novel spectral representation learning framework for causal estimation in the presence of hidden confounders. We demonstrated the superior performance of our method over existing approaches, showing reduced computational burden and increased accuracy. While our algorithm benefits from established theoretical guarantees for instrumental variables (IV) (e.g., Bennett et al., 2023c, Wang et al., 2022), our advantages for IV with observed confounders and principal component learning (PCL) are primarily justified empirical. Future work will extend existing theoretical proofs to these settings.

A.2 Broader Impacts

Potential Positive Societal Impacts. Accurately estimating causal effects has significant potential to benefit society across multiple disciplines. For example, in epidemiology, understanding causal relationships can enhance public health strategies, contributing to better disease prevention and control. This paper addresses the problem of estimating causal effects with hidden confounders. The proposed method can be effectively applied to various domains with different settings. This leads to more informed decision-making and improved outcomes in various fields.

Potential Negative Societal Impacts. While the proposed method for estimating causal effects offers multiple benefits, there are potential negative societal impacts to consider. Misapplication of the method without thorough validation could lead to incorrect conclusions and potentially harmful decisions, particularly in sensitive areas like healthcare or public policy.

B Complete Algorithms

To make our whole procedure clear, we provide the complete algorithm framework for causal estimation with spectral representation., with modifications for corresponding problems.

Algorithm 1 Spectral Representation Learning for Instrument Variable Regression

input Function class $\mathcal{F}$ of the representation.

Estimate $\phi(x)$ and $\psi(z)$ with $\{x_i, z_i\}$ using Equation (18) or Equation (18) and obtained the estimated representation $\widehat{\phi}(x)$ and $\widehat{\psi}(z)$.

Solve the min-max problem Equation (16) and obtain the v and w.

output $f(x) = v^\top \widehat{\phi}(x)$.

Algorithm 2 Spectral Representation Learning for Instrument Variable Regression with Observable Confounding

input Function class $\mathcal{F}_{xzo}$, $\mathcal{F}_{yzo}$.

Estimate $\psi(z), \phi(x), V(o)$ with $\{x_i, z_i, o_i\}$ from $P(x|z,o)$ and obtained the estimated representation $\widehat{\phi}(x)$, $\widehat{V}(o)$ and $\widehat{\psi}(z)$.

Estimate $Q(o)$, $\nu(y)$ with $P(y|z,o)$ and learned $\widehat{\psi}(z)$, $\widehat{V}(o)$ and obtain estimated $\widehat{Q}(o)$, $\widehat{\nu}(y)$.

Solve the min-max optimization Equation (17) and obtain the G and w.

output $f(x,o) = \left\langle G, \widehat{V}(o)^\top \otimes \widehat{\phi}(x)^\top \right\rangle$ and $u(z,o) = w^\top \widehat{Q}(o)^\top \widehat{V}(o) \widehat{\psi}(z)$.

C Contrastive Representation Learning

As we discussed in the previous section, we can obtain a representation of the covariates by decomposing a certain conditional expectation operator. We exploit the different contrastive learning objectives to implement the factorization, which is naturally compatible with neural network parameterized spectral representations and stochastic gradient descent, as we illustrate below.

Taking the factorization of $P(x \mid z)$ as an example, if we want to learn ϕ and ψ such that

$$P(x \mid z) = P(x) \langle \phi(x), \psi(z) \rangle,$$

one method is to consider a set of possible representations $\mathcal{S}$ and maximize the following objective

$$\max_{(\phi,\psi)\in\mathcal{S}} \mathcal{L}_{\text{rep}}(\phi,\psi) := \frac{2}{n} \sum_{i\in[n]} \phi(x_i)^\top \psi(z_i) - \frac{1}{n(n-1)} \sum_{i\neq j} \left(\phi(x_i)^\top \psi(z_j)\right)^2 - 1,$$

that has been used in Wang et al. [2022]. Another choice is to minimize the following objective

$$\min_{(\phi,\psi)\in\mathcal{S}} \mathcal{L}_{\text{rep}}(\phi,\psi) = -\frac{1}{n} \sum_{i\in[n]} \log \phi(x_i)^\top \psi(z_i) + \frac{1}{n} \sum_{i\in[n]} \log \left(\sum_{i\neq j} \phi(x_i)^\top \psi(z_j) \right), \tag{18}$$

that has been used in Zhang et al. [2022], Qiu et al. [2022]. Under mild assumptions like realizability, both methods provide a consistent estimation of $\phi(x)^\top \psi(z)$, but with different theoretical guarantees. The key idea is to separate the sample of x from $P(x|z)$ and $P(x)$, which eventually learns the ratio of $P(x|z)/P(x)$ in the form of $\langle \phi(x), \psi(z) \rangle$. Another benefit of the contrastive loss is that it is naturally compatible with stochastic gradient descent (SGD) and can, therefore, be easily scaled up for huge datasets.

Similarly, we can construct corresponding contrastive losses for implementing the conditional operator factorization of $P(x|o,z)$ and $P(y|o,z)$ for IV-OC as (12) and (13) and $P(w|x,z)$ and $P(y|x,z)$ as (15) for PCL, respectively.

Remark (Connection to DFIV [Xu et al., 2020] and DFPV [Xu et al., 2021]): The spectral representation for IV is equivalent to the target deep feature in [Xu et al., 2020]. The deep features $\psi(x)$ and $\varphi(z)$ are obtained to fulfill the condition

$$\mathbb{E}_{x|z}[\psi(x)] = A\varphi(z), \tag{19}$$

where $A \in \mathbb{R}^{d\times p}$ is some constant matrix independent w.r.t. x and z. Obviously, $\phi(x)$ and $\psi(z)$ as the spectral decomposition of $P(x|z)$ provide one solution to (19) with $A = I_{d\times p}$ by the definition of SVD. The algorithm proposed by Xu et al. [2020] employs an additional $d \times d$ matrix besides $(\phi(x), \psi(z))$ through a bi-level optimization. However, it requires propagating gradients through a Cholesky decomposition, increasing the computational cost. Likewise, the representations $\phi(w)$, $\theta(x)$, and $\psi(z)$ in DFPV [Xu et al., 2021] are learned through bi-level optimization, with the same computational drawbacks.

D Identifiability

We reveal the effect of the choice of regularizer in terms of identifibility with overcomplete spectral representation in population version. We mainly focus on the IV problem in our discussion, and similar conclusion also holds for IV-OC and PCL.

Assumption 6 (Overcompleted representation)**.** *With larger dimension $d' \geqslant d$ for spectral representation in primal* (10) *and dual space* (11)*, we obtain $\mathcal{F}'$ and $\mathcal{U}'$ correspondingly, with $\mathcal{F} \subseteq \mathcal{F}'$ and $\mathcal{U} \subseteq \mathcal{U}'$.*

Proposition 5 (Least norm solution)**.** *Under Assumption 1 and Assumption 6, with appropriate λ in* (5)*, the estimator provides unique solution f with least $\Omega(f)$.*

This can be verified by reformulating the estimator (5) into a constraint optimization, *i.e.*,

$$\min_{f\in\mathcal{F},\varepsilon} \lambda\Omega(f) + \mathbb{E}\left[\varepsilon(s)^2\right] \quad \text{s.t., } \mathbb{E}[y - f(x,s) \mid s] = \varepsilon(s), \quad \mathbb{P}(s) - \text{a.e.}. \tag{20}$$

Based on the Assumption 1 and Assumption 6, there exists one $\tilde{f}$ achieving

$$\mathbb{E}\left[y - \tilde{f}(x,s) \mid s\right] = 0, \quad \mathbb{P}(s) - \text{a.e.}.$$

with the corresponding $\tilde{\varepsilon(s)} = 0$. By setting an appropriate λ, the optimal solution of (20), *i.e.*, , our estimator (5), will be $\tilde{f}$, which is the least norm solution.

Remark: The choice of $\Omega(f)$ reveals an interesting but largely ignored topic on the definition of *idenfiability with representation learning*. Specifically, in the classic IV setting [Liao et al., 2020, Bennett et al., 2023b], the function space with norm is given and assumed to be realizable, *i.e.*, the function space contains the groundtruth target function. Therefore, the least norm identifiability [Bennett et al., 2023b, Li et al., 2024] is naturally based on the norm of the given function space. However, we are aiming for identifying both the function space and the target function by exploiting the representation learning, where the function space is learned, therefore, no norm is predefined.

Our representation characterization in Section 3 clearly provides measure-dependent norms. Specifically, the representation is intimately coupled with the inner product w.r.t. a base measure, which naturally induces norms. We take IV as an example, we exploit the inner product w.r.t. $P(x)$, which induces the norm $\|\cdot\|_{L_2(P_x)}$. Meanwhile, since $f = v^\top \phi(x) \in \mathcal{R}(\phi)$ as shown in Proposition 2, we can also simply set $\|f\|^2_{\mathcal{R}(\phi)} := v^\top v$. We emphasize our characterization reveals the **importance of base measure** in representation learning stage and least norm estimation stage. For different $P(x)$ and the induced space $L_2(P_x)$, our method can have a different set of the learned representation of $\phi(x)$, and potentially leads to different norm definitions, and thus, different least norm solution when the Fredholm equation (1) is unidentifiable. Intuitively, the inner product with the measure plays as an essential prior implies the region we should focus in causal estimation.

E Proofs of Section 3

We discuss the omitted proofs of Section 3.

E.1 Proof of Proposition 1

The proof is in the main text.

E.2 Proof of Proposition 2

Notice that the subspace $\mathcal{R}(\Phi)$ of $L_2(P_x)$ is closed by continuity of the inner product in $\mathbb{R}^d$. Therefore, it is in direct sum with its orthogonal, *i.e.*, $L_2(P_x) = \mathcal{R}(\Phi) \oplus \mathcal{R}(\Phi)^\perp$ (Theorem 3.3.4. in Kreyszig, 1991). Following this observation, we write $f \in L_2(P_x)$ as $f = \langle \mathbf{u}_f, \varphi(\cdot)\rangle + h^\perp$, with $\mathbf{u}_f \in \mathbb{R}^d$, $h^\perp \in \mathcal{R}(\Phi)^\perp$, and we have

$$
\begin{aligned}
\mathcal{T} f(z) &= \int_{\mathcal{X}} \left(\langle \mathbf{u}_f, \varphi(x)\rangle + h^\perp(x)\right) \langle \phi(x), \psi(z)\rangle \, \mathrm{d}P_x && \text{(Assumption 2)} \\
&= \left\langle \mathbf{u}_f, \left(\int_{\mathcal{X}} \varphi(x)\varphi(x)^\intercal \, \mathrm{d}P_x\right)\psi(z)\right\rangle_{\mathbb{R}^d} + \left\langle h^\perp, \Phi\psi(z)\right\rangle_{L_2(P_x)} \\
&= \left\langle \mathbf{u}_f, \left(\int_{\mathcal{X}} \varphi(x)\varphi(x)^\intercal \mu(dx)\right)\psi(z)\right\rangle_{\mathbb{R}^d} && \left(h^\perp \in \mathcal{R}(\Phi)^\perp\right).
\end{aligned}
$$

Notice we also have

$$
\begin{aligned}
\|f\|^2_{L_2(P_x)} &= \mathbb{E}\left[f(x)^2\right] \\
&= \mathbb{E}\left[\langle \mathbf{u}_f, \phi(x)\rangle^2\right] + \mathbb{E}\left[h^\perp(x)^2\right] && \left(h^\perp \perp \langle \mathbf{u}_f, \varphi(\cdot)\rangle\right) \\
&= \left\langle \mathbf{u}_f, \mathbb{E}\left[\varphi(x)\varphi(x)^\intercal\right]\mathbf{u}_f\right\rangle + \mathbb{E}\left[h^\perp(x)^2\right].
\end{aligned}
$$

This shows $h^\perp$ influences the objective function only through $\mathbb{E}\left[h^\perp(x)^2\right]$. Since we are minimizing for f, we can set $h^\perp = 0$ and only consider $f \in \mathcal{R}(\Phi)$.

Furthermore, Assumption 1 states there exists a solution to Problem 2. Using the same decomposition, this provides a solution to Problem 2 in $\mathcal{R}(\Phi)$. This concludes the proof.

E.3 Proof of Proposition 3

Proof. By plugging Equation (13) into the RHS of Equation (3), we obtain

$$\left\langle \psi(z), V(o)^\top \left[\int \phi(x)\phi(x)^\top dP_x \right] v(o) \right\rangle = \left\langle \psi(z), W(o)^\top \int y\nu(y) dP_y \right\rangle,$$

Using the fact that ψ spans every direction,

$$V(o)^\top \underbrace{\left[\int \phi(x)\phi(x)^\top \mathrm{d}P_x \right]}_{:=A} v(o) = W(o)^\top \underbrace{\int y\nu(y)\,\mathrm{d}P_y}_{:=\alpha}.$$

We can now characterize the space of $v(o)$. Specifically, we have

$$v(o) = A^+ \left(V(o)^\top \right)^+ W(o)^\top \alpha. \tag{21}$$

Finding the space of $v(o)$ requires taking several (pseudo-)inverses, which is computationally expensive. We can instead use an alternative parametrization in $Q(o) := \left(V(o)^\top\right)^+ W(o)^\top$, which gives $W(o) = Q(o)^\top V(o)$ in Equation (13). This leads to the desired parametrization. □

E.4 Proof of Proposition 4

The argument from Proposition 1 also applies here.

F Experiment Details

Setup. For the IV tasks, each method is assessed using a test set comprising 2,000 samples, while for the PCL task, the evaluation is based on a test set of 500 samples. The test samples are randomly generated using the same data generation setting as the training set. All experiments are conducted on a system equipped with an Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz and a Quadro RTX 8000 GPU. Note that for IV with observable confounders, we simplify the implementation by omitting the decomposition of $P(x|z)$. This helps reduce the number of parameters that need to be optimized.

Hyperparameters. For the IV regression experiments, we employ the same hyper-parameter setting for DFIV, KIV, and DeepGMM used in Xu et al. [2020]. For DE, we use the same Gaussian kernel with the bandwidth determined by the median trick Singh et al. [2019]. The network structures for SpecIV on different datasets are provided in Table 3 and 4.

Dataset Dimensions. The feature dimensions for these configurations are presented in Table 1.

Table 1: Dimensions of treatment, instrument, and observable confounder Variables in two datasets with varying settings. In the dSprites dataset, the 'low' and 'high' dimensions pertain to the instrument feature dimension. In the Demand Design dataset, 'SC' represents separated observable confounders. SpecIV and DFIV are assessed with separate representations for the observables. The remaining baseline methods are evaluated without separate observables, *i.e.*, , via incorporating these confounders into both treatment and instrumental variables.

Dataset	**dSprites**		**Demand Design**	
	Low-dim	High-dim	With SC	No SC
Treatment	4,096	4,096	784	1,569
Instrument	3	2,352	1	786
Observable	-	-	785	-

G Dimension Mapping

We utilize the mapping function in Bennett et al. [2019] to build the high-dimensional scenarios. Given a low-dimensional input $X_{\text{low}} \in \mathbb{R}$, we generate a corresponding high-dimensional variable

Table 2: Network Structure of the feature extractor for Demand Design dataset. Each 2D convolutional layer (Conv2d) is depicted with `(input dimension, output dimension, kernel size, stride, padding)`.

ImageFeature

Layer	Configuration
1	Input: 784
2	Conv2d(1, 16, 5, 1, 2), ReLU, MaxPool2d
3	Conv2d(16, 32, 5, 1, 2), ReLU, MaxPool2d
4	FC$(1568, 1024)$

Table 3: Network Structure of SpecIV for IV on the dSprites dataset. The tuple for each component represents the input/output dimensions within a particular component. FC refers to the fully-connected layers. BN represents batch normalization.

Treatment Feature Net (32)

Layer	Configuration
1	Input: 4096
2	FC$(4096, 1024)$, ReLU, BN
3	FC$(1024, 512)$, ReLU, BN
4	FC$(512, 128)$, ReLU, BN
5	FC$(128, 32)$, tanh

Instrument Feature Net (32)

Layer	Configuration
1	Input: 3
2	FC$(3, 256)$, ReLU, BN
3	FC$(256, 128)$, ReLU, BN
4	FC$(128, 128)$, ReLU, BN
5	FC$(128, 32)$, ReLU

Treatment Feature Net (64)

Layer	Configuration
1	Input: 4096
2	FC$(4096, 1024)$, ReLU, BN
3	FC$(1024, 256)$, ReLU, BN
4	FC$(256, 64)$, tanh

Instrument Feature Net (64)

Layer	Configuration
1	Input: 2352
2	FC$(2352, 1024)$, ReLU, BN
3	FC$(1024, 256)$, ReLU, BN
4	FC$(256, 64)$, ReLU

$X_{\text{high}} \in \mathbb{R}^{784}$ via:

$$X_{\text{high}} = \text{RandomImage}(\pi(X_{\text{low}})) \tag{22}$$

where $\pi(x) = \text{round}(\min(\max(1.5x + 5, 0), 9))$ transforms the input to an integer within the range 0 to 9; RandomImage(d) selects a random MNIST image corresponding to the input digit d.

Table 4: Network Structure of SpecIV for IV on the Demand Design dataset. The tuple for each component represents the input/output dimensions within a particular component. FC refers to the fully-connected layers. BN represents batch normalization. ImageFeature indicates the feature extractor detailed in Table 2.

Treatment Feature Net (32)

Layer	Configuration
1	Input: 784 (P)
2	ImageFeature(P)
3	FC$(1024, 512)$, ReLU, BN
4	FC$(512, 256)$, ReLU, BN
5	FC$(256, 32)$, tanh

Observable Feature Net (32)

Layer	Configuration
1	Input: 785 (T, S)
2	ImageFeature(S), T
3	FC$(1025, 512)$, ReLU, BN
4	FC$(512, 256)$, ReLU,BN
5	FC$(256, 32)$, tanh

Instrument Feature Net (32)

Layer	Configuration
1	Input: 1
2	FC$(1, 16)$, ReLU, BN
3	FC$(16, 4)$, ReLU

Outcome Feature Net (32)

Layer	Configuration
1	Input: 1
2	FC$(1, 16)$, ReLU, BN
3	FC$(16, 4)$, ReLU

Table 5: Network Structure of SpecPCL for PCL on the dSprites dataset. The tuple for each component represents the input/output dimensions within a particular component. FC refers to the fully-connected layers. BN represents batch normalization.

Treatment Feature Net

Layer	Configuration
1	Input: 4096
2	FC$(4096, 1024)$, ReLU, BN
3	FC$(1024, 512)$, ReLU, BN
4	FC$(512, 32)$, tanh

Observable Feature Net

Layer	Configuration
1	Input: 4096
2	FC$(4096, 1024)$, ReLU, BN
3	FC$(1024, 256)$, ReLU,BN
4	FC$(256, 32)$, tanh

Instrument Feature Net

Layer	Configuration
1	Input: 3
2	FC$(3, 16)$, ReLU, BN
3	FC$(16, 2)$, ReLU

Outcome Feature Net

Layer	Configuration
1	Input: 1
2	FC$(1, 16)$, ReLU, BN
3	FC$(16, 2)$, ReLU

H Data Generation for dSprites

H.1 IV Task

We generate data using the following relationships:

$$f(X) = \frac{\|AX\|_2^2 - 5000}{1000},$$
$$Y = f(X) + 32(\texttt{posX} - 0.5) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, 0.5).$$

where each entry of $A \in \mathbb{R}^{10\times 4096}$ is drawn from Uniform$(0, 1)$. We generate A once initially and maintain it fixed throughout the experiment.

The relationship above allows us to generate the treatment $X \in \mathbb{R}^{4096}$ and the instrument $Z \in \mathbb{R}^3$, denoted as the low-dimensional scenario. By applying equation (22) to each element of Z, we can transform these into a high-dimensional scenario, *i.e.*, , $Z_{\text{high}} \in \mathbb{R}^{2352}$.

H.2 PCL Task

For the PCL setting, we employ the same definition of $f(\cdot)$ and obtain Y through:

$$Y = \frac{1}{12}(\texttt{posX} - 0.5)f(X) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, 0.5).$$

We conduct the structural function estimation experiment on dSprites with treatment in $\mathbb{R}^{4096}$ and instrument in $\mathbb{R}^3$. Following Xu et al. [2021], we fix the `shape` as `heart` and use (`scale`, `rotation`, `posX`) as the treatment-inducing proxy. We then sample another image from the dSprites dataset with the same `posY` as the output-inducing proxy, *i.e.*, , Image(`scale`=0.8, `rotation`=0, `posX`=0.5, `posY`) $+ \eta$, with $\eta \sim \mathcal{N}(0, 0.1I)$.

I Data Generation for Demand Design

Following Singh et al. [2019], we generate data samples using the following relationships:

$$Y = f(P, T, S) + \varepsilon$$
$$f(P, T, S) = 100 + (10 + p)Sh(T) - 2P$$
$$h(t) = 2\left(\frac{(t-5)^4}{600} + \exp(-4(t-5)^2) + \frac{t}{10} - 2\right)$$

with

$$
\begin{aligned}
S &\sim \text{Uniform}\{1, \ldots, 7\} \\
T &\sim \text{Uniform}[0, 10] \\
C &\sim \mathcal{N}(0, 1) \\
V &\sim \mathcal{N}(0, 1) \\
\varepsilon &\sim \mathcal{N}(\rho V, 1 - \rho^2) \\
P &= 25 + (C + 3)h(T) + V
\end{aligned}
$$

The mapping function (22) is then used to generate high-dimensional versions of the variables P and S. This results in $P, S \in \mathbb{R}^{784}$ and $T, C \in \mathbb{R}$.

J Unlabeled Data Augmentation

In practical scenarios, it's often easier to collect a large amount of unlabeled data than labeled data. For instance, in the Demand Design task, the true ticket demand, which is the ground-truth label, is more challenging to acquire than other readily available information like ticket price and year time. Therefore, it's crucial to explore how much an IV method can improve by using extra unlabeled data during its training process. Note that for SpecIV with observable confounders, the unlabeled samples are used to optimize the decomposition in (12) while the original labeled data are employed for both (12) and (13). Figure 4 shows that as the number of unlabeled training samples increases (up to 3x labeled data), the MSE for both methods decreases. This trend confirms that adding unlabeled data contributes positively to the model's performance. Notably, Figure 4a shows that SpecIV consistently achieves a lower error rate compared to DFIV. Figure 4b illustrates that incorporating additional unlabeled data can effectively compensate for the lack of labeled data and enhance model performance. This demonstrates a promising characteristic for practical applications where unlabeled data is more accessible.

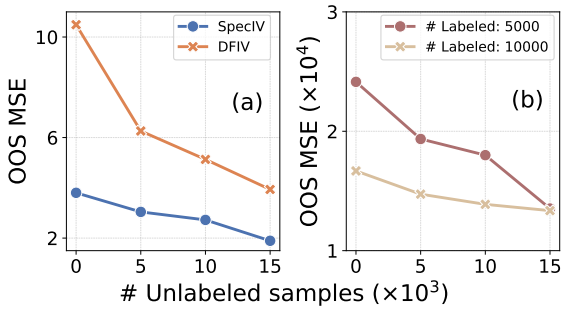


Figure 4: MSE comparison of (a) SpecIV and DFIV on dSprites dataset with 5,000 labeled training samples and (b) SpecIV on Demand Design Dataset with 5,000 and 10,000 labeled samples, respectively. Each method is trained with the specified amount of labeled data and an additional set of unlabeled samples.