# Spectral Representation for Causal Estimation with Hidden Confounders

**Haotian Sun**[*]
Georgia Tech
haotian.sun@gatech.edu

**Antoine Moulin**[*†]
Universitat Pompeu Fabra
antoine.moulin@upf.edu

**Tongzheng Ren**[*†]
UT Austin
rtz19970824@gmail.com

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

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

[*] Equal contribution with random order [†] Part of the work was done while at Google DeepMind

## Abstract

We study the problem of causal effect estimation in the presence of unobserved confounders, focusing on two settings: instrumental variable (IV) regression with additional observed confounders, and proxy causal learning. Our approach uses a singular value decomposition of a conditional expectation operator combined with a saddle-point optimization method. In the IV regression setting, this can be viewed as a neural network generalization of the seminal approach due to Darolles et al. (2011). Saddle-point formulations have recently gained attention because they mitigate the double-sampling bias and are compatible with modern function approximation methods. We provide experimental validation across various settings and show that our approach outperforms existing methods on common benchmarks.

## 1 INTRODUCTION

Estimating causal effects is a fundamental problem across diverse fields, including economics, epidemiology, and social sciences. Unobserved confounders—variables influencing both the treatment and the outcome—present a major challenge to traditional estimation methods as they introduce spurious correlations leading to biased and inconsistent estimates (see Equation 6.1 and Chapter 12 in Stock and Watson, 2007). One approach to address unobserved confounding involves using additional variables to help identify the causal effect. Instrumental variable (IV) regression (Wright, 1928; Stock and Trebbi, 2003) and proxy causal learning (PCL) Kuroki and Pearl (2014) are two prominent examples of this approach.

IV regression aims to solve an ill-posed inverse problem of the form $Ef = r$, where $f$ lives in the same space as the causal effect to be identified, $r$ is the expected output conditioned on an *instrument*, and $E$ is the corresponding conditional expectation operator. An instrument is a variable that provides exogenous variation in the treatment—variation that is uncorrelated with the unobserved confounder—allowing us to isolate the treatment's causal effect on the outcome. The problem is ill-posed because solving for $f$ (assuming a solution exists) typically involves the inverse eigenvalues of $E$, which can be arbitrarily close to zero. A common approach to account for this is to assume the target function $f$ satisfies a *source condition* (Engl et al., 1996), which relates its smoothness to the operator $E$ and ensures that the target primarily depends on the largest eigenvalues of $E$. Existing methods for efficiently solving this inverse problem can be broadly 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).

Two-stage methods aim to minimize an expected mean-squared error (MSE) of the form $\|Ef - r\|^2$ and are designed to deal with the conditional expectation nested inside the square function. Stage 1 performs a regression to estimate the conditional means, and Stage 2 regresses the outcome on the estimates obtained in Stage 1. Prior works differ in the parametrization of

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

Stage 1. Hartford et al. (2017) introduce a deep mixture model to estimate the conditional means, while Darolles et al. (2011) estimate the conditional densities with kernel density estimators. Singh et al. (2019) instead learn a conditional mean embedding (Song et al., 2009; Grunewalder et al., 2012; Li et al., 2022) of features that map the input to a reproducing kernel Hilbert space (RKHS), providing a nonlinear generalization of two-stage least-squares. However, this approach uses fixed and pre-defined feature dictionaries.

Alternatively, features can be learned adaptively within the two-stage least-squares (2SLS) framework (Xu et al., 2020), using gradient-based methods to minimize the same MSE. However, a naive approach leads to biased estimates due to the presence of the conditional expectation, failing to minimize the loss of interest. This *double sampling bias* has been a particular challenge in causal inference and reinforcement learning (see Chapter 11.5 in Sutton and Barto, 2018, Antos et al., 2008; Bradtke and Barto, 1996). While samples from the conditional distribution would bypass the issue, we rarely have enough data to obtain multiple samples for each specific conditioning value. This motivates conditional moment methods which avoid minimizing the squared error directly. Instead, these methods consider a saddle-point optimization problem of the form $\min_{f\in\mathcal{F}} \max_{g\in\mathcal{G}} \mathcal{L}(f,g)$ for some function classes $\mathcal{F}$ and $\mathcal{G}$. This formulation can be derived for instance via the Lagrangian of a feasibility problem (Bennett et al., 2023b) or the Fenchel conjugate of the square function together with an interchangeability result (Dai et al., 2017; Dikkala et al., 2020; Liao et al., 2020).

These saddle-point formulations involve only expectations with respect to the *joint* data distribution. While facilitating first-order methods and function approximation, they require strong assumptions on the function classes used (Dikkala et al., 2020; Liao et al., 2020; Bennett et al., 2023a,c). A natural one is *realizability*; it requires the function classes to contain the objects to be estimated, *e.g.*, $f_0 \in \mathcal{F}$ where $f_0$ is a solution to the problem. Removing it at the cost of a constant error due to misspecification is possible. However, due to the double sampling issue, most algorithms cannot work under realizability alone, and an additional expressivity assumption is required. A common one is *closedness* and requires stability of the function classes under the conditional expectation operator $E$ and/or its adjoint $E^\star$, *e.g.*, $E\mathcal{F} \subset \mathcal{G}$ and/or vice-versa. While Bennett et al. (2023b) manage to avoid the closedness assumption, they use a specific source condition and require an additional realizability assumption on the dual function class $\mathcal{G}$. It is unclear whether their method can adapt to different degrees of smoothness. Let $\hat{f}_n$ be an estimator of $f_0$, and $n$ the number of samples at hand. Dikkala et al. (2020) show, under these assumptions, that the projected MSE, $\|E(\hat{f}_n - f_0)\|_2$, converges at a rate of $\mathcal{O}\left(n^{-1/2}\right)$. Liao et al. (2020) achieve a stronger $L_2$ guarantee, $\|\hat{f}_n - f_0\|_2 = \mathcal{O}\left(n^{-1/6}\right)$, with an additional source condition and uniqueness of the solution. Their analysis crucially hinges on the latter to convert a projected MSE guarantee to an $L_2$ guarantee by controlling the *measure of ill-posedness* $\sup_{f\in\mathcal{F}} \frac{\|f-f_0\|_2^2}{\|E(f-f_0)\|_2^2}$, which can be unbounded if several solutions exist. On the other hand, Bennett et al. (2023b) obtain $\|\hat{f}_n - f_0\|_2 = \mathcal{O}\left(n^{-1/4}\right)$ with only a source condition and realizable classes, and Bennett et al. (2023c) obtain guarantees on both the projected MSE and the $L_2$ norm with two additional closedness assumptions.

The assumptions required for IV regression can be restrictive. Prior work extends IV regression to accommodate observable confounders Horowitz (2011); Xu et al. (2020). When a valid instrument is unavailable, an alternative is to use *proxy variables*, which contain relevant information on the unobserved confounder. Kuroki and Pearl (2014) provide necessary conditions on the proxy variables for identifying the true causal effect; Miao et al. (2018) generalize these conditions. Recent methods estimate causal effects in the proxy causal learning (PCL) setting, including (Deaner, 2018; Mastouri et al., 2021) for fixed feature dictionaries, and (Xu et al., 2021; Kompa et al., 2022) for adaptive feature dictionaries. However, the theoretical understanding of IV regression with observed confounders and PCL is less mature than that of IV regression.

While conditional moment methods offer strong statistical guarantees for IV regression, choosing the function classes $\mathcal{F}$ and $\mathcal{G}$ remains unclear. We propose a method based on a low-rank assumption on the conditional operator $E$, drawing inspiration from reinforcement learning (Jin et al., 2020). This assumption yields function classes satisfying the crucial realizability and closedness conditions, enabling optimization over finite-dimensional variables. We leverage this to derive efficient algorithms for IV regression with observed confounders (IV-OC) and PCL. Section 2 formalizes the settings and derives the saddle-point formulations. We introduce the low-rank assumption and show how to leverage it in Section 3. Importantly, unlike previous works focusing on standard IV regression, our method also learns an adaptive basis for IV with observed confounder and PCL. One component of our method is a representation learning algorithm inspired from Wang et al. (2022), discussed in Section 4. Finally, we provide experimental validation in Section 5 and show our approach outperforms existing methods on IV and PCL benchmarks.

## 2 PRELIMINARIES

We formalize the three settings of interest and derive their equivalent saddle-point formulations.

**Notation**. For a random variable $A$ taking values in $\mathcal{A}$, let $\mathbb{P}_A$ denote its probability distribution, and $L_2(\mathbb{P}_A)$ denote the space of $\mathbb{P}_A$-square-integrable functions $f : \mathcal{A} \to \mathbb{R}$. Given another random variable $B$, we denote $\mathbb{P}_{A|B}$ a regular version of the conditional distribution of $A$ given $B$. Given $n$ samples, $\mathbb{E}_n$ denotes the empirical expectation. denote the Euclidean inner product as $\langle \cdot, \cdot \rangle$, and the range of an operator $E$ as $\mathcal{R}(E)$.

### 2.1 Causal Estimation with Hidden Confounders

Given random variables $X$, $S$ and $Y$, in IV regression with and without observed confounders, we aim to find a function $f$ satisfying

$$\mathbb{E}[Y - f(X, S) \mid S] = 0\,. \tag{1}$$

Here, $Y$ is the outcome, $X$ is the treatment, and $S$ contains the accessible side information. IV regression and PCL differ in their assumptions and applicability.

IV regression is used when the unobserved confounder affects the outcome linearly. As shown in Figure 1a, it involves an *instrument* $Z$ that (i) is independent of the outcome $Y$ conditional on the treatment $X$ and the confounder $\varepsilon$, $Z \perp\!\!\!\perp Y \mid (X, \varepsilon)$, and (ii) satisfies $\mathbb{E}[\varepsilon \mid Z] = 0$. In this setting, the side information is given by the instrument, $S = Z$ in Equation (1). Since the instrument $Z$ does not affect the outcome $Y$, we only consider functions of the treatment $X$ alone

$$\mathbb{E}[Y - f(X) \mid Z] = 0\,. \tag{2}$$

With an additional observed confounder $O$, the available information becomes $S = (Z, O)$ (Figure 1b), and we are interested in solving the equation

$$\mathbb{E}[Y - f(X, O) \mid Z, O] = 0\,, \tag{3}$$

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

PCL (Miao et al., 2018; Deaner, 2018) uses two *proxies* correlated with the unobserved confounder $\varepsilon$. One proxy, $Z$, correlates with the treatment $X$; the other, $W$, correlates with the outcome $Y$ (Figure 1c). The proxies $Z$ and $W$ satisfy the independence properties $Z \perp\!\!\!\perp Y \mid (X, \varepsilon)$ and $W \perp\!\!\!\perp (Z, X) \mid \varepsilon$. Here, $S = (Z, W)$, and we consider

$$\mathbb{E}[Y - f(X, W) \mid X, Z] = 0\,. \tag{4}$$

These equations are ill-posed inverse problems we are interested in. Next, we derive their equivalent saddle-point formulations.

### 2.2 Primal-Dual Framework for Causal Estimation

We now derive a saddle-point formulation from Equation (1). Similar derivations apply to PCL. Prior work minimizes the mean-squared error plus a regularizer, $\Omega$, to address ill-posedness

$$\min_{f \in \mathcal{F}} \mathbb{E}\left[\mathbb{E}[Y - f(X, S) \mid S]^2\right] + \lambda \Omega(f) \triangleq \mathcal{E}(f)\,, \tag{5}$$

where $\mathcal{F} \subset L_2(\mathbb{P}_{XS})$ is convex, $\lambda > 0$, and $\Omega$ is a strongly convex regularizer (*e.g.*, a $L_2$ regularizer Darolles et al. (2011) or an RKHS norm Singh et al. (2019); Zhang et al. (2023); Wang et al. (2022)). Following, *e.g.*, Dai et al. (2017), we start with the Fenchel conjugate of the square function to write

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}[Y - f(X, S) \mid S]^2\right] \\
&= 2\,\mathbb{E}\left[\max_{g \in \mathbb{R}} \left\{ g\,\mathbb{E}[Y - f(X, S) \mid S] - \frac{1}{2} g^2 \right\}\right] \\
&= 2 \max_{g \in L_2(\mathbb{P}_S)} \mathbb{E}\left[ g(S)\,\mathbb{E}[Y - f(X, S) \mid S] - \frac{1}{2} g(S)^2 \right] \\
&= 2 \max_{g \in L_2(\mathbb{P}_S)} \mathbb{E}\left[ g(S)\,(Y - f(X, S)) - \frac{1}{2} g(S)^2 \right]\,,
\end{aligned}$$

where the second equality follows from (Dai et al., 2017, Lemma 1), and the last equality uses the tower rule. Thus, given a convex class $\mathcal{G} \subset L_2(\mathbb{P}_S)$, we consider

$$\min_{f \in \mathcal{F}} \max_{g \in \mathcal{G}} \underbrace{\mathbb{E}\left[ g(S)\,(Y - f(X, S)) - \frac{1}{2} g(S)^2 \right]}_{\triangleq \mathcal{L}(f, g)} + \lambda \Omega(f)\,. \tag{6}$$

$\mathcal{L}$ is strongly-convex in $f$ and strongly-concave in $g$. 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 $\mathbb{E}[\mathcal{L}_n(f, g)] = \mathcal{L}(f, g)$.

For IV regression, we consider $S = Z$ and $\mathcal{F} \subset L_2(\mathbb{P}_X)$ in Problem (6), or $S = (Z, O)$ and $\mathcal{F} \subset L_2(\mathbb{P}_{XO})$ with an observable confounder. For PCL, the min-max problem becomes

$$\begin{aligned}
\min_{f \in \mathcal{F}} \max_{g \in \mathcal{G}} \mathbb{E}&\left[ g(X, Z) \cdot (Y - f(X, W)) - \frac{1}{2} g(X, Z)^2 \right] \\
&+ \lambda \Omega(f)\,.
\end{aligned} \tag{7}$$

For consistency of the estimator

$$\widehat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \max_{g \in \mathcal{G}} \mathcal{L}_n(f, g)\,,$$

the classes $\mathcal{F}$ and $\mathcal{G}$ must ensure that the minimizer $f_0 = \operatorname{argmin}_{f \in \mathcal{F}} \max_{g \in \mathcal{G}} \mathcal{L}(f, g)$ solves the original inverse problem. A key challenge is choosing the function

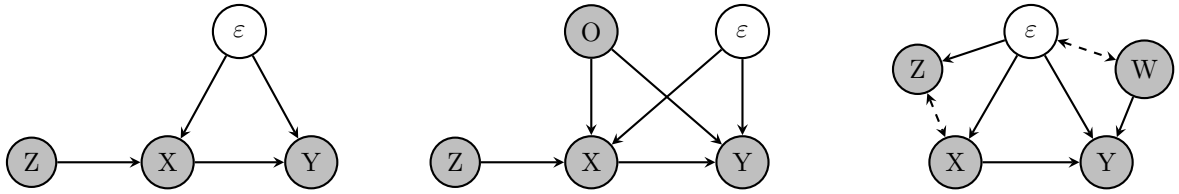


Figure 1: Causal graphs for the three settings: (a) IV regression, (b) IV regression with observed confounder ($O$), and (c) proxy causal inference. Gray nodes represent observed variables; white nodes are unobserved.

classes $\mathcal{F}$ and $\mathcal{G}$ to be small enough for optimization to be tractable, yet large enough for realizability to hold.

## 3 CHARACTERIZING THE FUNCTION CLASSES

We first characterize the function classes F and G for IV regression under a low-rank assumption, then generalize this to the IV-OC and PCL settings. For clarity, this section assumes the feature maps introduced in the following assumptions are known; Section 4 discusses learning these mappings.

### 3.1 Instrument Variable Regression

Let $E : L_2(\mathbb{P}_X) \to L_2(\mathbb{P}_Z)$ be the conditional expectation operator defined as $Ef = \mathbb{E}[f(X) \mid Z]$ for any $f \in L_2(\mathbb{P}_X)$, and assume Equation (2) has a solution.

**Assumption 1.** *There exists $f_0 \in L_2(\mathbb{P}_X)$ such that $Ef_0 = \mathbb{E}[Y \mid Z]$.*

Our goal is to find subspaces of $L_2(\mathbb{P}_X)$ and $L_2(\mathbb{P}_Z)$ such that the saddle-point of Problem (6) yields a solution to Equation (2). We now introduce the key low-rank assumption on the conditional and marginal distributions of $X$ and $Z$.

**Assumption 2.** *The distributions $\mathbb{P}_X$ and $\mathbb{P}_{X|Z=z}$, for any $z \in \mathcal{Z}$, have densities with respect to the Lebesgue measure, denoted $p_X$ and $p_{X|Z}(\cdot \mid z)$, 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 $x, z$,*

$$p_{X|Z}(x \mid z) = p_X(x) \langle \phi(x), \psi(z) \rangle . \tag{8}$$

We denote $\Psi : \mathbb{R}^d \to L_2(\mathbb{P}_Z)$ and $\Phi : \mathbb{R}^d \to L_2(\mathbb{P}_X)$ the operators defined for any vector $u$ as $\Psi u = \langle \psi(\cdot), u \rangle$, and $\Phi u = \langle \phi(\cdot), u \rangle$.

**Remark on Assumption 2.** The existence of densities is similar to Assumption A.1 in Darolles et al. (2011), but is stated on the conditional and marginal distributions rather than the joint distribution. Equation (8) is akin to assuming that the conditional expectation operator $E$ admits a finite singular-value decomposition (SVD). Compactness implies the existence of a countable SVD; if the spectrum of $E$ decays sufficiently fast, a finite-dimensional approximation is meaningful (Ren et al., 2022a). This is also equivalent to the low-rank assumption used in reinforcement learning (Jin et al., 2020).

A consequence of Assumption 2 is that for any function $f \in L_2(\mathbb{P}_X)$, the conditional expectation $Ef$ is linear in the features $\psi$.

**Proposition 1.** *If Assumption 2 holds, then for any function $f \in L_2(\mathbb{P}_X)$, there exists a vector $v_f \in \mathbb{R}^d$ such that $Ef = \langle \psi(Z), v_f \rangle$.*

By Assumption 1, there exists a vector $v_0 \in \mathbb{R}^d$ such that $\mathbb{E}[Y \mid Z] = \langle \psi(Z), v_0 \rangle$. Thus, for a given function $f \in L_2(\mathbb{P}_X)$, the maximizer (over $L_2(\mathbb{P}_Z)$) in Equation (6) is $g_f^\star = \langle \psi(\cdot), v_0 - v_f \rangle$. This suggests the following the dual function class.

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

$$\mathcal{G} \triangleq \left\{ z \mapsto \langle \psi(z), v \rangle, v \in \mathbb{R}^d \right\} , \tag{9}$$

*that is, we have*

$$\min_{f \in L_2(\mathbb{P}_X)} \max_{g \in L_2(\mathbb{P}_Z)} \mathcal{L}(f, g) = \min_{f \in L_2(\mathbb{P}_X)} \max_{g \in \mathcal{R}(\Psi)} \mathcal{L}(f, g) .$$

Next, we define the class $\mathcal{F}$ as follows.

**Proposition 3** (Primal space for IV)**.** *Under Assumptions 1, 2, we have*

$$\min_{f \in L_2(\mathbb{P}_X)} \max_{g \in \mathcal{R}(\Psi)} \mathcal{L}(f, g) = \min_{f \in \mathcal{R}(\Phi)} \max_{g \in \mathcal{R}(\Psi)} \mathcal{L}(f, g) ,$$

*and the class $\mathcal{F}$ defined as follows contains a solution to Problem 2,*

$$\mathcal{F} \triangleq \left\{ x \mapsto \langle \phi(x), u \rangle, u \in \mathbb{R}^d \right\} . \tag{10}$$

Compared to parametrizations of $f$ using RKHS functions or deep neural networks (Dai et al., 2017; Muandet et al., 2020; Liao et al., 2020), where realizability is assumed to hold, we consider a structural assumption that allows us to characterize explicitly realizable function classes.

**Remark on the parametrizations.** The characterization of the primal function class $\mathcal{F}$ through the spectral decomposition of $E$ has been exploited in IV regression (Darolles et al., 2011; Wang et al., 2022) and reinforcement learning (Jin et al., 2020; Yang and Wang, 2020; Ren et al., 2022b). Specifically, the feature map $\phi$ is used as in Proposition 3. Critically, we also exploit $\psi$ to characterize the dual function class $\mathcal{G}$ in Proposition 2 and use it within the saddle-point formulation.

Furthermore, Darolles et al., 2011 approximate the conditional operator by estimating unconditional densities with kernel density estimators, then taking their ratio. Instead, we propose a representation learning algorithm (Section 4) to directly learn the feature maps $\phi$ and $\psi$ under Assumption 2, which provides an estimate of the density ratio. We next generalize these results to the IV-OC and PCL settings, returning to the optimization problem and algorithm in Section 4.

### 3.2 Instrument Variable Regression with Observable Confounding, Proxy Causal Learning

In this section, the operator $E : L_2(\mathbb{P}_{XO}) \to L_2(\mathbb{P}_{ZO})$ is defined for any $f$ as $Ef = \mathbb{E}[f(X,O) \mid Z,O]$, and we assume Problem 3 has a solution.

**Assumption 3.** *There exists $f_0 \in L_2(\mathbb{P}_{XO})$ such that $Ef_0 = \mathbb{E}[Y \mid Z,O]$.*

Due to the additional dependence on the observed confounder $O$, we introduce a different low-rank assumption.

**Assumption 4.** *The distributions $\mathbb{P}_X$ and $\mathbb{P}_{X|Z=z,O=o}$, for any $z,o$, have densities with respect to the Lebesgue measure, denoted $p_X$ and $p_{X|Z,O}(\cdot \mid z,o)$, respectively. Furthermore, there exist feature maps $\phi : \mathcal{X} \to \mathbb{R}^{d_x}$, $\psi : \mathcal{Z} \to \mathbb{R}^{d_z}$, and $V : \mathcal{O} \to \mathbb{R}^{d_x \times d_z}$ such that for any $x,z,o$,*

$$p_{X|Z,O}(x \mid z,o) = p_X(x) \langle \phi(x), V(o)\psi(z) \rangle . \quad (11)$$

Under Assumption 4, the operator $E$ "linearizes" any input function $f \in L_2(\mathbb{P}_{XO})$ in the sense that $Ef$ is linear in $\psi$ and $V$ (separately, but not jointly).

**Proposition 4.** *If Assumption 4 holds, then for any function $f \in L_2(\mathbb{P}_{XO})$, there exists a function $v_f : \mathcal{O} \to \mathbb{R}^{d_z}$ such that $Ef = \langle \psi(Z), v_f(O) \rangle$.*

Fixing $O = o$, we can apply the same argument as in the proof of Proposition 3 to the integral $\int_{\mathcal{X}} \phi(x) f(x,o) \mathbb{P}_X(\mathrm{d}x)$. This shows that for any $o$, it suffices to consider $f(\cdot,o)$ in the span of $\Phi$. Therefore, there exists a function $u : \mathcal{O} \to \mathbb{R}^{d_x}$ such that for any $x,o$, we have $f(x,o) = \langle \phi(x), u(o) \rangle$. However, it is still unclear what the function $u$ looks like.

In general, we cannot determine the representation for $u$ solely through $E$, as we did in the IV regression setting. To see this, consider the special case where the function $f$ does not depend on the treatment, *i.e.* for any $x,o$, $f(x,o) = h(o)$ for some function $h \in L_2(\mathbb{P}_O)$. Then, we have that $\mathbb{E}[h(O) \mid Z,O] = h(O)$, and the operator $E$ provides no information about $h$ (and thus, no information about $u$).

Given a solution to Equation (3), we have $\mathbb{E}[f(X,O) \mid Z,O] = \mathbb{E}[Y \mid Z,O]$. This implies that (i) the space for $u(o)$ should also lie in the space of $o$ in RHS, and (ii) the space for $z$ on both sides should be the same. Thus, we make the following assumption.

**Assumption 5.** *The distributions $\mathbb{P}_Y$ and $\mathbb{P}_{Y|Z=z,O=o}$, for any $z,o$, have densities with respect to the Lebesgue measure, denoted $p_Y$ and $p_{Y|Z,O}(\cdot \mid z,o)$, respectively. Furthermore, there exist feature maps $\nu : \mathcal{Y} \to \mathbb{R}^{d_y}$, and $W : \mathcal{O} \to \mathbb{R}^{d_y \times d_z}$ such that for any $y,z,o$,*

$$p_{Y|Z,O}(y \mid z,o) = p_Y(y) \langle \nu(y), W(o)\psi(z) \rangle , \quad (12)$$

*where $\psi$ is the feature map from Assumption 4.*

With this additional assumption, let us denote $Q(o) \triangleq \left(V(o)^\intercal\right)^+ W(o)^\intercal$ for any $o$. We then obtain the following class.

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

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

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

**Proposition 6** (Dual space for IV-OC)**.** *Under Assumptions 3, 4, and 5, the dual function of IV-OC is realizable in*

$$\mathcal{G} \triangleq \left\{ (z,o) \mapsto \left\langle \psi(z), V(o)^\intercal Q(o)\gamma \right\rangle, \gamma \in \mathbb{R}^{d_y} \right\} .$$

**Remark (Representation for PCL):** IV-OC and PCL share the same conditioning structure. Thus, the representation characterization for IV-OC also applies to PCL, yielding the following primal and dual function classes

$$\mathcal{F} \triangleq \Big\{ (x,w) \mapsto \phi(w)^\intercal BQ(x)\beta, \quad B \in \mathbb{R}^{d_w \times d_w}, \beta \in \mathbb{R}^{d_x} \Big\},$$
$$\mathcal{G} \triangleq \left\{ (x,z) \mapsto \psi(z)^\intercal V(x)^\intercal Q(x)\gamma, \gamma \in \mathbb{R}^{d_x} \right\} , \quad (13)$$

where the spectral representations arise from similar factorizations

$$p_{W|X,Z}(w \mid x,z) = p_W(w)\langle \phi(w), V(x)\psi(z)\rangle,$$
$$p_{Y|X,Z}(y \mid x,z) = p_Y(y)\langle \nu(y), Q(x)^\intercal V(x)\psi(z)\rangle. \tag{14}$$

**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 $f(x,w) = \zeta^\intercal(\theta(x) \otimes \phi(w))$, where $\theta$ and $\phi$ either fixed feature dictionaries or learned neural net feature dictionaries. This parametrization shares some similarity to Equation (13), if we rewrite it as

$$f(x,w) = \phi(w)^\intercal BQ(x)\beta = \langle \beta \mathrm{vec}(B), Q(x)^\intercal \otimes \phi(w)^\intercal\rangle.$$

However, we obtain the features $Q(x)$ and $\phi(w)$ from a spectral perspective, differing from prior work.

**Remark (Comparison to Wang et al. (2022)):** Wang et al. (2022) also exploit the spectral structure for causal inference, with a representation for IV regression that follows Darolles et al. (2011). However, their spectral representation for IV-OC and PCL differs. They simply reduce the IV-OC and PCL to standard IV by augmenting the treatments and instrumental variables, while we explicitly characterize the function classes for the IV-OC and PCL min-max problems and provide practical algorithms. As noted in Xu et al. (2020), reducing to IV via augmentation ignores the problem structure, potentially leading to unnecessary complexity.

# 4 CAUSAL ESTIMATION WITH SPECTRAL REPRESENTATION

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

## 4.1 Contrastive Representation Learning

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

Consider the factorization of $p_{X|Z}$ in Assumption 2 as an example. If we want to learn mappings $\phi$ and $\psi$ such that

$$p_{X|Z}(x \mid z) = p_X(x)\langle \phi(x), \psi(z)\rangle,$$

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

$$\max_{(\tilde\phi,\tilde\psi)\in\mathcal{S}} \mathcal{L}_{\mathrm{rep}}(\tilde\phi,\tilde\psi) \triangleq \frac{2}{n}\sum_{i=1}^n \tilde\phi(x_i)^\intercal \tilde\psi(z_i) - \frac{1}{n(n-1)}\sum_{i\neq j}\left(\tilde\phi(x_i)^\intercal\tilde\psi(z_j)\right)^2 - 1, \tag{15}$$

which has been used in Wang et al. (2022). Another choice is to minimize the following objective

$$\min_{(\tilde\phi,\tilde\psi)\in\mathcal{S}} \mathcal{L}_{\mathrm{rep}}(\tilde\phi,\tilde\psi) \triangleq -\frac{1}{n}\sum_{i=1}^n \log \tilde\phi(x_i)^\intercal\tilde\psi(z_i) + \frac{1}{n}\sum_{i=1}^n \log\left(\sum_{i\neq j}\tilde\phi(x_i)^\intercal\tilde\psi(z_j)\right), \tag{16}$$

which has been used in Zhang et al. (2022); Qiu et al. (2022). Under mild assumptions such as realizability, both methods provide a consistent estimation of the mapping $(x,z)\mapsto \phi(x)^\intercal\psi(z)$, but with different theoretical guarantees. This method learns the ratio $p_{X|Z}(x \mid z)/p_X(x)$ 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 thus be scaled up to large datasets.

Similarly, we can construct corresponding contrastive losses for implementing the low-rank factorization of $p_{X|Z,O}$ and $p_{Y|Z,O}$ for IV-OC as (11) and (12), and $p_{W|X,Z}$ and $p_{Y|X,Z}$ as (14) for PCL, respectively. For instance, the contrastive loss $\mathcal{L}_{\mathrm{rep}}$ in (16) for $p_{X|Z,O}$ in IV-OC takes the form:

$$\mathcal{L}_{\mathrm{rep}}(\tilde\phi,\tilde\psi,\tilde V) \triangleq -\frac{1}{n}\sum_{i=1}^n \log\left(\tilde\phi(x_i)^\intercal\left(\tilde V(o_i)\tilde\psi(z_i)\right)\right) + \frac{1}{n}\sum_{i=1}^n \log\left(\sum_{i\neq j}\tilde\phi(x_i)^\intercal\left(\tilde V(o_j)\tilde\psi(z_j)\right)\right).$$

**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). Their deep features $\psi(x)$ and $\phi(z)$ are obtained to fulfill the condition

$$\mathbb{E}[\psi(X) \mid Z = z] = A\phi(z), \tag{17}$$

where $A \in \mathbb{R}^{d\times p}$ is some constant matrix independent of $x$ and $z$. Naturally, the feature maps $\phi$ and $\psi$ from the low-rank decomposition of $p_{X|Z}$ (Assumption 2) provide one solution to (17) with $A = I_{d\times p}$ by definition of the SVD. The algorithm proposed by Xu et al. (2020) employs an additional $d\times d$ matrix besides the mappings $(\phi,\psi)$ through a bi-level optimization.

However, it requires propagating gradients through a Cholesky decomposition, increasing the computational cost. Likewise, the representations $\phi$ of $W$, $\theta$ of $X$, and $\psi$ of $Z$ in DFPV (Xu et al., 2021) are learned through bi-level optimization, with the same computational drawbacks.

### 4.2 Estimation of 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 in 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 for two reasons: (i) the representations are *learned from the data* and (ii) the user may choose a representation dimensionality $d'$ that differs from the true (potentially infinite) dimensionality $d$.

The first issue induces a bias because the learned representations are evaluated on the same data used for their training. This bias can be mitigated by splitting the data into separate sets for representation learning phase and saddle-point optimization. Alternatively, if data splitting is not feasible, a uniform covering argument can be used to show the additional error introduced by using the same data scales with the complexity of the function classes used in representation learning (*e.g.* the number of parameters in the neural networks) divided by the square root of the number of samples. This resulting error is often reasonable, especially when compared to the initial double-sampling bias.

The second issue, where $d' \neq d$, leads to a misspecification error. If $d' < d$[1], this error depends on the magnitude of the "missing" eigenvalues of the conditional expectation operator. If the operator's spectrum decays sufficiently rapidly, this error will be small. A rigorous analysis of these errors is left for future work; however, similar issues have been addressed in prior work (*e.g.* Wang et al. (2022) in the context of representation learning, and Bennett et al. (2023b) regarding the misspecification error).

We now present our algorithms below.

**Instrument variable regression.** Given the estimated representation $\widehat{\phi}$ and $\widehat{\psi}$ obtained by factorizing $p_{X|Z}$ using Equation (15) or (16), we solve the following saddle-point optimization problem,

$$\min_{u\in\mathbb{R}^d}\max_{v\in\mathbb{R}^d}\mathbb{E}\left[\left(v^\intercal\widehat{\psi}(Z)\right)\left(Y-u^\intercal\widehat{\phi}(X)\right) - \frac{1}{2}\left(v^\intercal\widehat{\psi}(Z)\right)^2\right] + \lambda\Omega\left(u^\intercal\widehat{\phi}(\cdot)\right). \quad (18)$$

This yields parameters $\widehat{u}$ and $\widehat{v}$. We then define the estimated functions $f: x\mapsto \widehat{u}^\intercal\widehat{\phi}(x)$ and $g(z)=\widehat{v}^\intercal\widehat{\psi}(z)$.

**Instrument variable regression with observable confounding and PCL.** We denote $\widehat{\phi},\widehat{\psi},\widehat{V},\widehat{Q}$ the learned representations for IV-OC. Unlike the IV regression setting without observable confounding, we now have an additional feature of the observable $o$. The parametrization for $f$, $f:(x,o)\mapsto\phi(x)^\intercal BQ(o)\beta$, contains two parameters $B$ and $\beta$. Direct substitution of this parametrization into the saddle-point problem would induce non-convexity. To recover convexity, we consider the following reparametrization:

$$\begin{aligned}\phi(x)^\intercal BQ(o)\beta &= \operatorname{tr}\left(\phi(x)^\intercal BQ(o)\beta\right)\\ &= \operatorname{tr}\Big(\left(Q(o)^\intercal\otimes\phi(x)^\intercal\right)^\intercal \underbrace{\beta\operatorname{vec}(B)^\intercal}_{\triangleq G}\Big)\\ &= \left\langle G, Q(o)^\intercal\otimes\phi(x)^\intercal\right\rangle_{\mathrm{F}},\end{aligned}$$

where $\langle\cdot,\cdot\rangle_{\mathrm{F}}$ denotes the Frobenius inner product, and the second equality comes from a property of the Kronecker matrix-vector product

$$(C\otimes D)\operatorname{vec}(G)=\operatorname{vec}(CGD^\intercal).$$

We define $f_G:(x,o)\mapsto\left\langle G,\widehat{Q}(o)^\intercal\otimes\widehat{\phi}(x)^\intercal\right\rangle_{\mathrm{F}}$ for some $G\in\mathbb{R}^{d_y\times d_x^2}$ and $g_w:(z,o)\mapsto\widehat{\psi}(z)^\intercal\widehat{V}(o)^\intercal\widehat{Q}(o)w$ for a vector $w\in\mathbb{R}^{d_y}$. Based on this reparametrization, we consider the following convex-concave saddle-point optimization problem

$$\min_{G\in\mathbb{R}^{d_y\times d_x^2}}\max_{w\in\mathbb{R}^{d_y}}\widehat{\mathcal{L}}(G,w), \quad (19)$$

where we defined

$$\widehat{\mathcal{L}}(G,w)=\mathbb{E}\left[g_w(Z,O)\left(Y-f_G(X,O)\right) - \frac{1}{2}g_w(Z,O)^2\right]+\lambda\Omega(f_G).$$

Denoting $\widehat{G}$ and $\widehat{w}$ the parameters solving the saddle-point problem, we return the functions $f:(x,o)\mapsto f_{\widehat{G}}(x,o)$ and $g:(z,o)\mapsto g_{\widehat{w}}(z,o)$. We can perform similar derivations for the PCL setting. We provide the complete algorithms in Appendix B.

[1] Note that the case where $d' > d$ is typically handled by the regularizer, which would tend to select a solution with a small norm, effectively setting the components corresponding to the extra dimensions to zero.

## 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 [2]. 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 E.

**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) [3]. 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). We evaluate DFPV (Xu et al., 2021), KPV (Mastouri et al., 2023), PMMR (Mastouri et al., 2023), and CE-VAE (Louizos et al., 2017).

**Datasets.** We conduct experiments on the dSprites Dataset (Matthey et al., 2017) for both low- and high-dimensional IV and PCL settings, as well as the Demand Design dataset (Hartford et al., 2017) for IV with observable confounders. 1) **dSprites** comprises images determined by five latent parameters (`shape`, `scale`, `rotation`, `posX`, `posY`). Each image has dimensions of $64 \times 64 = 4096$ and serves as the treatment variable $X$. Following the setup in Xu et al. (2020), we keep `shape` fixed as `heart` and use `posY` as the hidden confounder. The remaining latent variables compose the instrument variable $Z$. In addition, we introduce a high-dimensional setting where the instruments are mapped to a high-dimensional variable in $\mathbb{R}^{2352}$, as proposed in Bennett et al. (2019). 2) **Demand Design** is a synthetic benchmark for nonlinear IV regression. Given the airplane ticket price $P$, the objective is to predict ticket demands, denoted as $Y$, in the presence of two observable confounders: price sensitivity $S \in \{1, \dots, 7\}$ and the year time $T \in [0, 10]$. Additionally, we introduce an unobservable confounder, represented as correlated noise in $P$ and $Y$. Furthermore, we set the fuel price $C$ as the instrument variable and $X$ as the treatment. We also employ the mapping function introduced in Bennett et al. (2019) to map both $P$ and $S$ to high-dimensional variables in $\mathbb{R}^{784}$. This represents a more challenging scenario, as the IV regression method must estimate the relevant variables from noisy, high-dimensional data. Details about the data generation are presented in Appendix F and H.

[2] Available at https://github.com/haotiansun14/SpecIV

[3] The algorithm Dual IV (Muandet et al., 2020) follows the same primal-dual framework of DE, but with closed-form solution, which involves explicit matrix inverse, while DE exploits the stochastic gradient algorithm that is more computationally efficient.

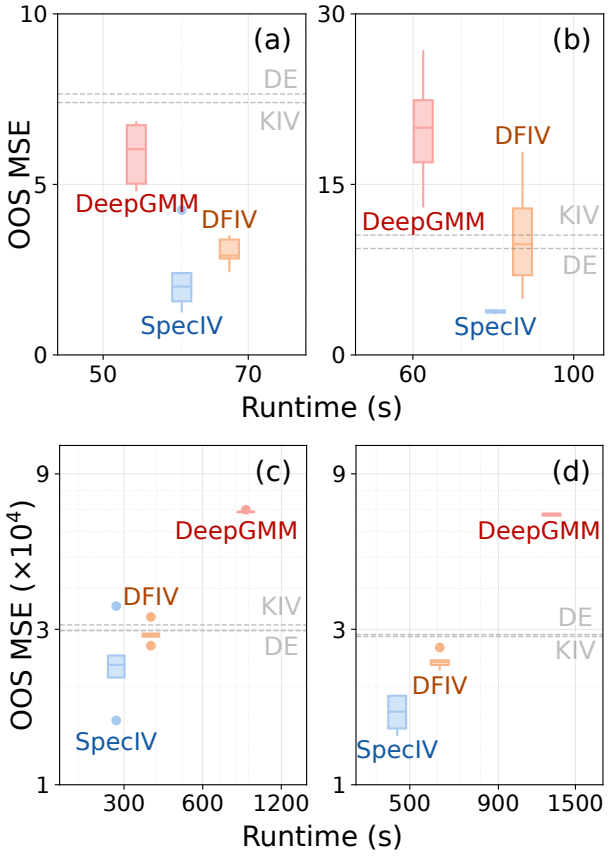


Figure 2: MSE and Runtime on dSprites Dataset with (a) Low-Dimensional Instruments (32) and (b) High-Dimensional Instruments (64), and Demand Design dataset with (c) 5,000 and (d) 10,000 training samples. Methods employing pre-specified features, *i.e.*, DE and KIV, are represented as dashed baselines.

**Instrument Variable Regression.** Figure 2a and 2b depict the performance and the execution time of various methods on the dSprites dataset. The optimal performance corresponds to proximity to the bottom-left corner, indicating low MSE and reduced runtime. 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. The cause might be their reliance on predetermined feature representations, which restricts adaptability. In low-dimensional scenarios, DeepGMM and DFIV manage to keep errors within a reasonable range, but DeepGMM's performance declines sharply with increased feature dimensions, while SpecIV maintains good estimation accuracy. Furthermore, both DeepGMM and DFIV exhibit increased variance in the high-dimensional setting,

possibly due to their 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 and the fuel price $C$ as the instrument, accompanied by the observables $(T, S)$. Each of these three components is presented using distinct neural networks. For other methods that don't model observables separately, we integrate the observables directly with the treatment and instrument. This means $(P, T, S)$ is used as the treatment and $(C, T, S)$ as the instrument. Figure 2c and Figure 2d represent the performance of these approaches given different amounts of training samples. In both scenarios, SpecIV consistently delivers the lowest error also achieving the best runtime. DeepGMM performs the least efficiently, with the highest error rates and runtimes. This may result from incorporating observables into both treatment and instrument. This overlooks the fact that we only need to consider the conditional expectation of $X$ given $(Z, O)$, thus making the problem suboptimal. KIV and DE, limited by their fixed feature representations, do not benefit from an increase in training sample size and remain less expressive. Overall, the SpecIV method outperforms other approaches in learning structured functions with the presence of observable confounders.

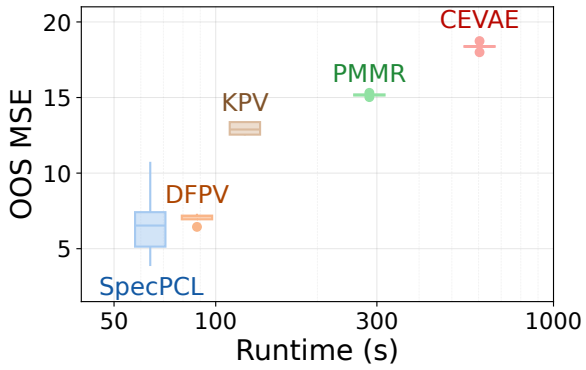


Figure 3: MSE and Runtime of structural function estimation on the dSprites dataset with 5,000 training samples.

**Proxy Causal Learning.** 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.

## 6 CONCLUSION

We introduced a novel spectral representation learning framework for causal estimation in the presence of hidden confounders. Our approach leverages a low-rank assumption on conditional densities to characterize function classes within a saddle-point optimization problem. This characterization enables the development of efficient algorithms for IV regression, both with and without observed confounders, and for proximal causal learning (PCL). Extensive experimental validation demonstrates that our method outperforms existing approaches. While our method benefits from existing theoretical guarantees for IV regression (*e.g.*, Bennett et al., 2023c; Wang et al., 2022), extending these guarantees to IV regression with observed confounders and to PCL remains a challenging open problem. Another interesting direction for future research is a formal comparison of the performance of two-stage methods and conditional moment methods.

### Acknowledgements

Arthur Gretton and Antoine Moulin wish to thank Dimitri Meunier for helpful discussions. This project has received funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 950180); from the Gatsby Charitable Foundation; from NSF ECCS2401391; from NSF IIS-2403240; and from Dolby support. Antoine Moulin acknowledges travel support from ELISE (GA no 951847) and the Gatsby Charitable Foundation.

### References

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

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.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

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.

S. Grunewalder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors. In *International Conference on Machine Learning*, 2012.

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.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. In *Advances in Neural Information Processing Systems*, 2022.

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.

Le Song, Jonathan Huang, Alexander Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *International Conference on Machine Learning*, pages 961 – 968, 2009.

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.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

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.

## Checklist

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. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

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. [Yes]
   (b) The license information of the assets, if applicable. [No]
   (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]

# 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 empirically. 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, understanding causal relationships in epidemiology 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 $\phi : \mathcal{X} \to \mathbb{R}^d$ and $\psi : \mathcal{Z} \to \mathbb{R}^d$ with samples $\{x_i, z_i\}$ using Equation (15) or Equation (16) to obtain the estimated representations $\widehat{\phi}$ and $\widehat{\psi}$.

Solve the min-max problem Equation (18) to obtain the parameters $v$ and $w$.

**output** $f : x \mapsto \left\langle v, \widehat{\phi}(x) \right\rangle$ and $g : z \mapsto \left\langle w, \widehat{\psi}(z) \right\rangle$.

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

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

Estimate $\psi : \mathcal{Z} \to \mathbb{R}^{d_z}, \phi : \mathcal{X} \to \mathbb{R}^{d_x}, V : \mathcal{O} \to \mathbb{R}^{d_x \times d_z}$ with samples $\{x_i, z_i, o_i\}$ to obtain the estimated representations $\widehat{\phi}$, $\widehat{V}$, and $\widehat{\psi}$.

Estimate $Q : \mathcal{O} \to \mathbb{R}^{d_x \times d_y}$, $\nu : \mathcal{Y} \to \mathbb{R}^{d_y}$ with samples $\{y_i, z_i, o_i\}$ and learned representations $\widehat{\psi}$ and $\widehat{V}$ to obtain the estimated representations $\widehat{Q}$, $\widehat{\nu}$.

Solve the min-max optimization Equation (19) and obtain the parameters $G$ and $w$.

**output** $f : (x, o) \mapsto \left\langle G, \widehat{V}(o)^\intercal \otimes \widehat{\phi}(x)^\intercal \right\rangle$ and $g : (z, o) \mapsto \left\langle w, \widehat{Q}(o)^\intercal \widehat{V}(o) \widehat{\psi}(z) \right\rangle$.

**Implementation of Representation-Learning Stage.** For IV tasks, we parameterize $\phi : \mathcal{X} \to \mathbb{R}^d$ and $\psi : \mathcal{Z} \to \mathbb{R}^d$ using two neural networks. In the IV-OC setting, we extend this by additionally parameterizing $\nu : \mathcal{Y} \to \mathbb{R}^{d_y}$ with a neural network and introducing another network $\xi : \mathcal{O} \to \mathbb{R}^{d_o}$ for the observable. To implement the low-rank factorization of $p_{X|Z,O}$ and $p_{Y|Z,O}$ for IV-OC, we introduce two learnable tensors, $P_V \in \mathbb{R}^{d_x \times d_z \times d_o}$ and $P_Q \in \mathbb{R}^{d_x \times d_y \times d_o}$. For any $o$, we then define $V(o) = P_V \otimes_3 \xi(o) \in \mathbb{R}^{d_x \times d_z}$, $Q(o) = P_Q \otimes_3 \xi(o) \in \mathbb{R}^{d_x \times d_y}$, and $W(o) = Q(o)^\intercal V(o)$, which are subsequently incorporated into (11) and (12) for representation learning. A similar strategy is adopted for the PCL setting.

# C IDENTIFIABILITY AND ROLE OF THE REGULARIZER

We now discuss how the choice of regularizer $\Omega$ influences the solution obtained by our method. While we focus on IV regression for clarity, similar considerations apply to the IV-OC and PCL settings.

Adding a regularizer $\Omega$ when the underlying integral equation has multiple solutions (*i.e.* is not identifiable) is crucial. In traditional IV regression, identifiability is established *before* considering estimation, typically through assumptions on the relationship between the instrument, treatment, and outcome (*e.g.*, relevance and exclusion restrictions). These assumptions ensure a *unique* solution to the integral equation. The estimation problem then focuses on finding this unique solution, often within a predefined function space (Liao et al., 2020).

In contrast, our approach combines representation learning with estimation. We first learn a spectral representation, and *then* solve a regularized optimization problem within the space spanned by the learned features. In general, *without regularization*, the underlying integral equation will have infinitely many solutions. The regularizer $\Omega$ then acts as a *selection mechanism*, choosing one particular solution from this infinite set and implicitly defining what we consider to be a "good" solution.

Our spectral representation, as described in Section 3, implicitly defines a norm depending on the underlying base measure. The eigenfunctions (and corresponding feature maps) obtained from the singular value decomposition of the conditional expectation operator are inherently tied to this base measure. In IV regression, we typically use the marginal distribution of the treatment, $\mathbb{P}_X$, and the euclidean norm on the space $L_2(\mathbb{P}_X)$. Because $f$ can be represented as $f = \langle \phi(\cdot), u \rangle$ (Proposition 3), an alternative norm can be defined directly on the representation space as $\|f\|_{\mathcal{R}(\Phi)} \triangleq \|u\|_2$. This norm is also tied to $\mathbb{P}_X$ through the learned representation $\phi$.

However, other base measures are possible. Consider, for example, an off-policy evaluation setting where we want to estimate the causal effect under a target policy $\pi$ different from the behavior policy that generated the data. In this case, it might be more appropriate to use, *e.g.*, a base measure weighted by the importance sampling ratio $\frac{\mathrm{d}\pi}{\mathrm{d}\mathbb{P}_X}$. This different base measure would result in different eigenfunctions and a different notion of what constitutes a "simple" or "smooth" solution. This highlights a crucial point: the choice of base measure, often implicit, significantly influences the learned representation and, consequently, the solution selected by the regularizer when the underlying problem is under-identified.

# D OMITTED PROOFS

We discuss the omitted proofs of Section 3.

## D.1 Proof of Proposition 1

**Proposition 1.** *If Assumption 2 holds, then for any function $f \in L_2(\mathbb{P}_X)$, there exists a vector $v_f \in \mathbb{R}^d$ such that $Ef = \langle \psi(Z), v_f \rangle$.*

*Proof.* Let $f \in L_2(\mathbb{P}_x)$. By definition of the conditional expectation operator $E$, we have

$$\begin{aligned} Ef &= \int_{\mathcal{X}} f(x)\, \mathbb{P}_{X|Z}(Z, \mathrm{d}x) \\ &= \int_{\mathcal{X}} f(x)\, \frac{\mathrm{d}\mathbb{P}_{X|Z}(Z, \cdot)}{\mathrm{d}\mathbb{P}_X}(x)\, \mathbb{P}_X(\mathrm{d}x) \\ &= \int_{\mathcal{X}} f(x)\, \langle \phi(x), \psi(Z) \rangle\, \mathbb{P}_X(\mathrm{d}x)\ , \end{aligned}$$

where the last two equalities follow from Assumption 2. The conclusion follows from the linearity of $\langle \cdot, \psi(Z) \rangle$

$$\begin{aligned} Ef &= \left\langle \int_{\mathcal{X}} \phi(x)\, f(x)\, \mathbb{P}_X(\mathrm{d}x), \psi(Z) \right\rangle \\ &= \langle \psi(Z), v_f \rangle\ , \end{aligned}$$

where we defined $v_f = \int_{\mathcal{X}} \phi(x)\, f(x)\, \mathbb{P}_X(\mathrm{d}x)$. □

### D.2 Proof of Proposition 2

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

$$\mathcal{G} \triangleq \left\{ z \mapsto \langle \psi(z), v \rangle, v \in \mathbb{R}^d \right\}, \tag{9}$$

*that is, we have*

$$\min_{f \in L_2(\mathbb{P}_X)} \max_{g \in L_2(\mathbb{P}_Z)} \mathcal{L}(f,g) = \min_{f \in L_2(\mathbb{P}_X)} \max_{g \in \mathcal{R}(\Psi)} \mathcal{L}(f,g).$$

*Proof.* For any $f \in L_2(\mathbb{P}_X)$ and $g \in L_2(\mathbb{P}_Z)$, the objective of the saddle-point problem (6) considered for IV regression is defined as

$$\mathcal{L}(f,g) = \mathbb{E}\left[ g(Z)(Y - f(X)) - \frac{1}{2} g(Z)^2 \right] + \lambda \Omega(f).$$

We aim to find a reasonable function class $\mathcal{G}$ that contains the maximizer $g_f^\star = \operatorname{argmax}_{g \in L_2(\mathbb{P}_Z)} \mathcal{L}(f,g)$ for any $f$. Note that by the tower rule, we can write

$$\begin{aligned}
\mathcal{L}(f,g) &= \mathbb{E}\left[ g(Z)\, \mathbb{E}[Y - f(X) \mid Z] - \frac{1}{2} g(Z)^2 \right] + \lambda \Omega(f) \\
&= \langle g, r_0 - Ef \rangle_{L_2(\mathbb{P}_Z)} - \frac{1}{2} \|g\|^2_{L_2(\mathbb{P}_Z)} + \lambda \Omega(f).
\end{aligned}$$

$\mathcal{L}(f,\cdot)$ is strongly concave. If a point $g_f^\star$ maximizes $\mathcal{L}(f,\cdot)$ then it must be a zero of its Frechet derivative, which we denote $D_g \mathcal{L}$,

$$\begin{aligned}
D_g \mathcal{L}(f, g_f^\star) = 0 \quad &\text{iff} \quad r_0 - Ef - g_f^\star = 0 \\
&\text{iff} \quad g_f^\star = r_0 - Ef.
\end{aligned}$$

By Assumptions 1, 2, there exist vectors $v_0, v_f \in \mathbb{R}^d$ such that $r_0 = \langle \psi(Z), v_0 \rangle$ and $Ef = \langle \psi(Z), v_f \rangle$. This implies $g_f^\star$ takes the form $g_f^\star = \langle \psi(Z), v_0 - v_f \rangle$, *i.e.* $g_f^\star \in \mathcal{R}(\Psi)$ and we can restrict the optimization problem to the class $\mathcal{G} = \mathcal{R}(\Psi)$. □

### D.3 Proof of Proposition 3

**Proposition 3** (Primal space for IV). *Under Assumptions 1, 2, we have*

$$\min_{f \in L_2(\mathbb{P}_X)} \max_{g \in \mathcal{R}(\Psi)} \mathcal{L}(f,g) = \min_{f \in \mathcal{R}(\Phi)} \max_{g \in \mathcal{R}(\Psi)} \mathcal{L}(f,g),$$

*and the class $\mathcal{F}$ defined as follows contains a solution to Problem 2,*

$$\mathcal{F} \triangleq \left\{ x \mapsto \langle \phi(x), u \rangle, u \in \mathbb{R}^d \right\}. \tag{10}$$

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

$$\begin{aligned}
Ef &= \int_{\mathcal{X}} \left( \langle u_f, \phi(x) \rangle + h^\perp(x) \right) \langle \phi(x), \psi(Z) \rangle \, \mathbb{P}_X(\mathrm{d}x) && \text{(Assumption 2)} \\
&= \left\langle u_f, \left( \int_{\mathcal{X}} \phi(x)\, \phi(x)^\intercal \, \mathbb{P}_X(\mathrm{d}x) \right) \psi(Z) \right\rangle_{\mathbb{R}^d} + \left\langle h^\perp, \Phi\psi(Z) \right\rangle_{L_2(\mathbb{P}_X)} \\
&= \left\langle u_f, \left( \int_{\mathcal{X}} \phi(x)\, \phi(x)^\intercal \, \mathbb{P}_X(\mathrm{d}x) \right) \psi(Z) \right\rangle_{\mathbb{R}^d} && \left( h^\perp \in \mathcal{R}(\Phi)^\perp \right).
\end{aligned}$$

Notice we also have

$$\begin{aligned}
\|f\|^2_{L_2(\mathbb{P}_X)} &= \mathbb{E}\left[ f(X)^2 \right] \\
&= \mathbb{E}\left[ \langle u_f, \phi(X) \rangle^2 \right] + \mathbb{E}\left[ h^\perp(X)^2 \right] && \left( h^\perp \perp \langle u_f, \phi(\cdot) \rangle \right) \\
&= \left\langle u_f, \mathbb{E}\left[ \phi(X)\, \phi(X)^\intercal \right] u_f \right\rangle + \mathbb{E}\left[ h^\perp(X)^2 \right].
\end{aligned}$$

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. □

### D.4 Proof of Proposition 4

**Proposition 4.** *If Assumption 4 holds, then for any function $f \in L_2(\mathbb{P}_{XO})$, there exists a function $v_f : \mathcal{O} \to \mathbb{R}^{d_z}$ such that $Ef = \langle \psi(Z), v_f(O) \rangle$.*

*Proof.* We follow the same reasoning as in the proof of Proposition 1. Let $f \in L_2(\mathbb{P}_{XO})$, and note we have

$$\begin{aligned} Ef &= \int_{\mathcal{X}} f(x,O)\, \mathbb{P}_{X|Z,O}((Z,O), \mathrm{d}x) \\ &= \int_{\mathcal{X}} f(x,O)\, \frac{\mathrm{d}\mathbb{P}_{X|Z,O}((Z,O),\cdot)}{\mathrm{d}\mathbb{P}_X}(x)\, \mathbb{P}_X(\mathrm{d}x) \\ &= \int_{\mathcal{X}} f(x,O)\, \langle \phi(x), V(O)\psi(Z)\rangle\, \mathbb{P}_X(\mathrm{d}x)\ , \end{aligned}$$

where the last two equalities follow from Assumption 4. The conclusion follows from the linearity of the inner product

$$\begin{aligned} Ef &= \left\langle \psi(Z), V(O)^{\intercal} \int_{\mathcal{X}} \phi(x) f(x,O)\, \mathbb{P}_X(\mathrm{d}x) \right\rangle \\ &= \langle \psi(Z), v_f(O) \rangle\ , \end{aligned}$$

where we defined $v_f : o \mapsto V(o)^{\intercal} \int_{\mathcal{X}} \phi(x) f(x,o)\, \mathbb{P}_X(\mathrm{d}x)$. □

### D.5 Proof of Proposition 5

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

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

*Proof.* By plugging Equation (12) into the RHS of Equation (3), we obtain

$$\left\langle \psi(z), V(o)^{\intercal} \left[\int_{\mathcal{X}} \phi(x)\phi(x)^{\intercal}\, \mathbb{P}_X(\mathrm{d}x)\right] v(o) \right\rangle = \left\langle \psi(z), W(o)^{\intercal} \int_{\mathcal{Y}} y\nu(y)\, \mathbb{P}_Y(\mathrm{d}y) \right\rangle\ ,$$

Using the fact that $\psi$ spans every direction, we get

$$V(o)^{\intercal} \underbrace{\left[\int \phi(x)\phi(x)^{\intercal}\, \mathbb{P}_X(\mathrm{d}x)\right]}_{\triangleq A} v(o) = W(o)^{\intercal} \underbrace{\int y\nu(y)\, \mathbb{P}_Y(\mathrm{d}y)}_{\triangleq \alpha}\ .$$

We can now write $v$ as

$$v(o) = A^{+} \left(V(o)^{\intercal}\right)^{+} W(o)^{\intercal} \alpha\,. \tag{20}$$

Finding the space $v$ belongs to requires taking several (pseudo-)inverses, which is computationally expensive. We can instead use an alternative parametrization in $Q(o) \triangleq \left(V(o)^{\intercal}\right)^{+} W(o)^{\intercal}$, which gives $W(o) = Q(o)^{\intercal} V(o)$ in Equation (12). This leads to the desired parametrization. □

### D.6 Proof of Proposition 6

**Proposition 6** (Dual space for IV-OC)**.** *Under Assumptions 3, 4, and 5, the dual function of IV-OC is realizable in*

$$\mathcal{G} \triangleq \left\{(z,o) \mapsto \left\langle \psi(z), V(o)^{\intercal} Q(o)\gamma \right\rangle, \gamma \in \mathbb{R}^{d_y}\right\}.$$

*Proof.* The argument from Proposition 2 also applies here. □

# E 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 | - |

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$(1568, 1024)$ |

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

$$X_{\text{high}} = \text{RandomImage}(\pi(X_{\text{low}})) \tag{21}$$

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

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 |

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 |

# G DATA GENERATION FOR DSPRITES

## G.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 (21) to each element of $Z$, we can transform these into a high-dimensional scenario, *i.e.*, $Z_{\text{high}} \in \mathbb{R}^{2352}$.

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 |

### G.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 $\eta \sim \mathcal{N}(0, 0.1I)$.

## H DATA GENERATION FOR DEMAND DESIGN

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

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

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

## I ADDITIONAL EXPERIMENTS

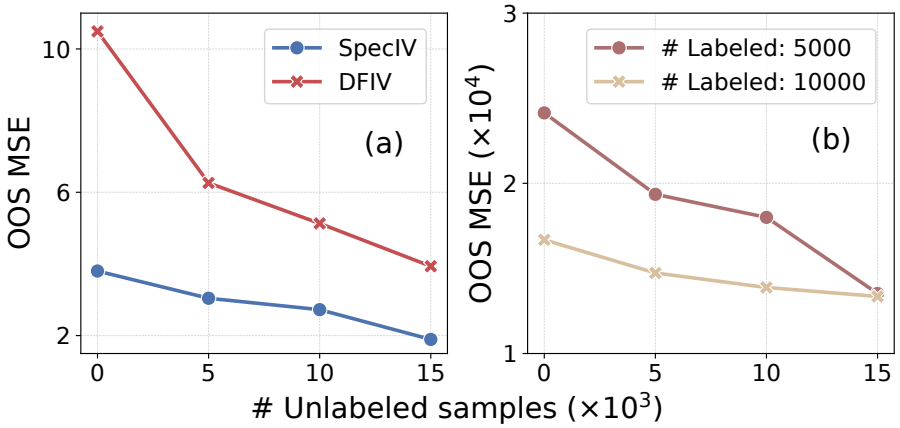


Figure 4: MSE comparison of (a) SpecIV and DFIV on dSprites dataset with 5,000 labeled training samples and (b) SpecIV on Demand Design Dataset with 5,000 and 10,000 labeled samples, respectively. Each method is trained with the specified amount of labeled data and an additional set of unlabeled samples.

**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 (11) while the original labeled data are employed for both (11) and (12). 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.

**Comparison with DeepIV (Hartford et al., 2017).** We compare SpecIV with DeepIV (Hartford et al., 2017), another important baseline, and observe that it exhibits inferior performance for both low- and high-dimensional treatment variables. Specifically, for uni-dimensional treatments, DeepIV is consistently outperformed by DFIV, as reported in Figure 4, Section 4.2, and Appendix E.2 by Xu et al. (2020). When the treatment variable is high-dimensional, DeepIV struggles to provide meaningful predictions due to the inherent challenges of performing conditional density estimation in high-dimensional spaces. To illustrate this, we report the MSE results on the Demand Design dataset, which consists of 10,000 training samples with a treatment dimension of 1,569. Table 6 validate that with a high-dimensional treatment variable, DeepIV performs significantly worse than SpecIV and other baselines.

| Method | SpecIV | DFIV | DeepIV | KIV | DE |
|---|---|---|---|---|---|
| OOS MSE ($\times 10^4$) | 1.67 | 2.40 | 4.31 | 2.85 | 2.89 |

Table 6: MSE comparison across different methods on Demand Design dataset with 10,000 training samples.