# Instrumental Variable Regression with Confounder Balancing

**Anpeng Wu** [1] **Kun Kuang** [1] **Bo Li** [2] **Fei Wu** [1 3 4]

## Abstract

This paper considers the challenge of estimating treatment effects from observational data in the presence of unmeasured confounders. A popular way to address this challenge is to utilize an instrumental variable (IV) for two-stage regression, i.e., 2SLS and variants, but limited to the linear setting. Recently, many nonlinear IV regression variants were proposed to overcome it by regressing the treatment with IVs and observed confounders in stage 1, leading to the imbalance of the observed confounders in stage 2. In this paper, we propose a Confounder Balanced IV Regression (CB-IV) algorithm to jointly remove the bias from the unmeasured confounders and balance the observed confounders. To the best of our knowledge, this is the first work to combine confounder balancing in IV regression for treatment effect estimation. Theoretically, we re-define and solve the inverse problems for the response-outcome function. Experiments show that our algorithm outperforms the existing approaches.

## 1. Introduction

Treatment effect estimation is one fundamental problem in causal inference, and its key challenge is to remove the confounding bias induced by the confounders, which affect both treatment and outcome. Under the unconfoundedness assumption (i.e., no unmeasured confounders), many confounder balancing methods, such as (Rubin, 1973; Kuang et al., 2017; Shalit et al., 2017), have been proposed to break the dependence between the treatment and all confounders. In practice, however, the unconfoundedness assumption is hardly satisfied and there always exist unmeasured confounders. How to precisely estimate the treatment effect from observational data in the presence of unmeasured confounders is of vital importance for both academic research and real applications.

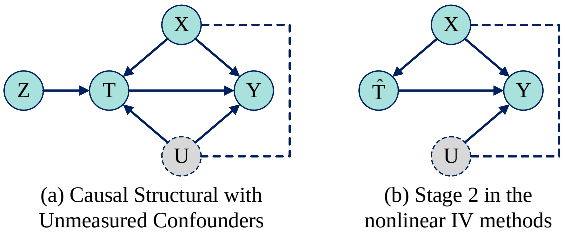
X
Z
T
Y
U
(a) Causal Structural with Unmeasured Confounders
X
$\hat{T}$
Y
U
(b) Stage 2 in the nonlinear IV methods

*Figure 1.* (a) Causal structural with unmeasured confounders. (b) Causal structure of outcome regression stage in the nonlinear IV methods. The observed confounders would affect both the estimated treatment $\hat{T}$ and outcome $Y$, leading to confounding bias.

A classical method to address the bias induced by unmeasured confounder is IV regression methods (Pearl et al., 2000; Heckman, 2008; Stock & Trebbi, 2003). As shown in Figure 1(a), let $T$ denote the treatment, $Y$ refer to the interest of outcome, $X$ and $U$ represent the observed and unobserved confounders, respectively, where $U$ might affect or be affected by $X$. $Z$ refers to the instrumental variables (IVs), which only influences $Y$ via $T$. Two-stage least squares (2SLS) regression and other variants (Imbens & Angrist, 1994; Angrist & Krueger, 2001; Carrasco et al., 2007; Buhlmann et al., 2014) can identify the treatment effect in the presence of unobserved confounders, but are limited to linear setting. For nonlinear scenarios (Example A.1), recently, many nonlinear IV regression variants (Hartford et al., 2017; Xu et al., 2021; Singh et al., 2019; Muandet et al., 2020) were proposed by two-stage regression. In stage 1, they learn a joint mapping from the instruments $Z$ and the observed confounders $X$ to the conditional distribution of the treatment $T$, i.e., $P(T|Z,X)$. Then, they estimate treatment $\hat{T}$ from $P(T|Z,X)$ obtained in stage 1 and perform nonlinear regression from the estimated treatment $\hat{T}$ and the observed confounders $X$ to the outcomes $Y$ in stage 2.

From the processes of these nonlinear IV methods, we know that the confounders $X$ would affect the estimated treatment $\hat{T}$ obtained from stage 1, and also influences the outcome $Y$ as shown in the figure 1(b). Thereby, the distribution of $X$ would become imbalanced between different arms of $\hat{T}$,

[1] Department of Computer Science and Technology, Zhejiang University, China [2] School of Economics and Managemen, Tsinghua University, China [3] Shanghai Institute for Advanced Study, Zhejiang University, China [4] Shanghai AI Laboratory, China. Correspondence to: Kun Kuang <kunkuang@zju.edu.cn>.

*Proceedings of the $39^{th}$ International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

hence bringing confounding bias[1] for outcome regression in stage 2 if their regression model is misspecified, leading to poor performance of these methods in practice. Therefore, explicitly removing confounding bias would improve treatment effect estimation performance and make extrapolation more credible (Abadie, 2003).

In this paper, we focus on treatment effect estimation with IV regression under homogeneity assumptions, and we propose a Confounder Balanced IV Regression (CB-IV[2]) algorithm to further remove the confounding bias from observed confounders by balancing in nonlinear scenarios. Specifically, our CB-IV algorithm contains the following three main components: (i) treatment regression: given $Z$ and $X$, identify conditional probability distribution of the treatment variable $T$ (i.e., $P(T|Z,X)$) for removing the confounding from unmeasured confounders; (ii) confounder balancing: learn a balanced representation of observed confounders $C = f_\theta(X)$, independent with the estimated treatment $\hat{T} \sim P(T|Z,X)$, to reduce the confounding from $X$ as shown in the figure 1(b); and (iii) outcome regression: regressing the outcome $Y$ on the estimated treatment $\hat{T}$ and confounders' representation $C$ for counterfactual inference. Theoretically, we re-define and solve two inverse problems under different sufficient identification assumptions separately, including: homogeneous treatment-outcome association and homogeneous instrument-treatment association. We also demonstrate the effectiveness of our proposed CB-IV with extensive empirical experiments.

The main contributions in this paper are as follows:

- We study the problem of treatment effect estimation with IV regression, and we find that previous IV regression methods would suffer from the confounding bias from the observed confounders, if the outcome model is misspecified and covariates are imbalanced.
- We propose a Confounder Balanced IV regression (CB-IV) method to jointly remove the bias from both the unobserved confounders with IV regression and the observed confounders by balancing. To the best of our knowledge, this is the first work to combine confounder balancing in IV regression.
- In two general settings satisfying homogeneous instrument-treatment association or homogeneous treatment-outcome association respectively, we give and solve the inverse relationship of response-outcome function of our algorithm. Empirical experiments also demonstrate the effectiveness of our algorithm.

## 2. Related Works

### 2.1. Instrumental Variable Methods

A popular way to estimate the causal effect from observational data in the presence of unmeasured confounders is to use an instrumental variable (IV). As a classical IV method, two-stage least squares (Pearl et al., 2000; Imbens & Angrist, 1994; Angrist & Krueger, 2001; Kuang et al., 2020b) performs linear regression to model the relationship between the treatments and outcomes conditional on the instruments. To relax linearity assumption, nonlinear IV regression variants learn a joint mapping from the instruments $Z$ and observed confounders $X$ to the treatments $T$ in stage 1. Sieve IV derives a finite dictionary of basis functions to replace the linear counterparts on the structural function and derives a lower bound. (Chen & Christensen, 2018; Newey & Powell, 2003). Kernel IV (Singh et al., 2019) and Dual IV (Muandet et al., 2020) implement 2-stage regression via mapping $X$ to a reproducing kernel Hilbert space and performing kernel ridge regression. DFIV (Xu et al., 2021) adopts deep neural nets to replace the kernel counterparts. DeepIV (Hartford et al., 2017) and OneSIV (Lin et al., 2019) estimate the conditional probability distribution of treatments $T$ using the instruments $Z$ and confounders $X$ in stage 1 and performs a joint mapping from resampled treatments $\hat{T} \sim P(T|Z,X)$ and confounders $X$ to the outcomes $Y$ in stage 2.

As shown in Figure 1(b), if the outcome model is misspecified, imbalanced variables $X$ will bring confounding bias for outcome regression in stage 2 in previous IV regression methods. To balance the overall sample, Abadie (2003); Singh & Sun (2019) uses regularized machine learning and achieves semiparametric efficiency with automatic kappa weights, which requires binary instrument, binary treatment and high dimensional covariates. In general settings, we propose a novel algorithm to combine confounder balance techniques with IV regression. To the best of our knowledge, this is the first provably efficient algorithm that combines the IV method with the confounder balance technique using deep representation learning.

### 2.2. Confounder Balance with Representation Learning

Nonexperimental studies are increasingly used to estimate treatment effect, and systematic differences between different treatment groups would introduce confounding bias. Inspired by traditional confounder balance works (Kuang et al., 2020a), such as propensity score methods(Rosenbaum & Rubin, 1983; Rosenbaum, 1987; Li et al., 2016; 2020), reweighting methods(Athey et al., 2018; He & Garcia, 2009), Doubly Robust (Funk et al., 2011) and backdoor criterion (Pearl, 2009a), CFR (Johansson et al., 2016; Shalit et al., 2017) formulates the problem of confounder balance as a covariate shift problem, and regard the treated group as the

[1] As introduced in Chapter 3.3 in Causality (Pearl, 2009b), the **confounding bias** between the treatment and outcome can be defined as the bias of treatment effect estimation when imbalanced confounders exist. More discussion on confounding bias is given in Section G in Appendix.

[2] The code is available at: https://github.com/anpwu/CB-IV

source domain and the control group as the target domain for domain adaptive balance under the unconfoundedness assumption. Johansson et al. (2016); Shalit et al. (2017) expect that representation $C = f_\theta(X)$, from all confounders $X$, discard information related to $T$, but retain as much information related to $Y$ as possible. SITE (Yao et al., 2018) preserves local similarity and balances the distributions of the representation $C$ simultaneously. DR-CFR (Hassanpour & Greiner, 2019a;b) and DeR-CFR (Wu et al., 2022) propose a disentanglement framework to identify the representation of confounders from all observed variables.

Deep representation learning has good performance and can capture complex relationships among treatments, observed confounders, and outcomes, but it requires the unconfoundedness assumption. Based on these confounder balance methods, we propose to use an instrumental variable to eliminate the bias from the unmeasured confounders.

## 3. Problem Setting and Assumptions

In this paper, we aim to estimate the average treatment effect by the structural function from observational data in the presence of unmeasured confounders. In the observational data $\mathbb{D} = \{z_i, x_i, t_i, y_i\}_{i=1}^n$, for each unit $i$, we observe a treatment variable $t_i \in T$ where $T \subset \mathbb{R}$, a outcome variable $y_i \in Y$ where $Y \subset \mathbb{R}$, instrumental variables $z_i \in Z$ where $Z \subset \mathbb{R}^{m_Z}$, and observed confounders $x_i \in X$ where $X \subset \mathbb{R}^{m_X}$. Besides, there are some unmeasured confounders $u_i \in U$ where $U \subset \mathbb{R}^{m_U}$ and might affect or be affected by $x_i$, but not recorded in the observational data. $m_X, m_Z$ and $m_U$ are the dimensions of the observed confounders $X$, instrumental variables $Z$ and unobserved confounders $U$. The causal relationship can be represented with the following model (as shown in Figure 1(a)):

$$\{Z, X, U\} \to T; \{T, X, U\} \to Y; Z \perp U, X; X \not\perp U \quad (1)$$

**Definition 3.1. The average treatment effect** ($ATE$):

$$ATE(t) = \mathbb{E}[Y \mid do(T = t), X] - \mathbb{E}[Y \mid do(T = 0), X] \quad (2)$$

**Definition 3.2. An Instrument Variable** $Z$ is an exogenous variable that affects the treatment $T$, but does not directly affect the outcome $Y$. Besides, an valid instrument variable satisfies the following three assumptions:

**Relevance:** $Z$ is a cause of $T$, i.e., $\mathbb{P}(T \mid Z) \neq \mathbb{P}(T)$.
**Exclusion:** $Z$ does not directly affect the outcome $Y$, i.e., $Z \perp Y \mid T, X, U$.
**Unconfounded:** $Z$ is independent of all confounders, including $X$ and $U$, i.e., $Z \perp X, U$

**Identification:** Even if the instrument satisfies these assumptions, at least one of the two homogeneity assumptions is required to identify the average treatment effect of $T$ on $Y$ (Imbens & Angrist, 1994; Angrist et al., 1996; Newey & Powell, 2003; Hernan & Robins, 2010; Wooldridge, 2010). The identifying assumptions in our paper basically follow the homogeneity assumptions (Heckman et al., 2006; Hernán & Robins, 2006; Hartwig et al., 2020), which is a more general version than Monotonicity Assumption (Imbens & Angrist, 1994; Angrist et al., 1996) and Additive Noise Assumption (Newey & Powell, 2003)) in the econometrics literature (Wooldridge, 2010; Hartwig et al., 2020; 2021). The two homogeneity assumptions are as follows:

**Assumption 3.3. Homogeneous Instrument-Treatment Association**: The association between the IV and the treatment is homogeneous in the different level of unmeasured confounders, i.e., $\mathbb{E}[T|Z = a, U] - \mathbb{E}[T|Z = b, U] = \mathbb{E}[T|Z = a] - \mathbb{E}[T|Z = b]$.

**Assumption 3.4. Homogeneous Treatment-Outcome Association**: The association between the treatment and the outcome is homogeneous in the different level of unmeasured confounders, i.e., $\mathbb{E}[Y|T = a, U] - \mathbb{E}[Y|T = b, U] = \mathbb{E}[Y|T = a] - \mathbb{E}[Y|T = b]$.

**Discussion about Confounder Imbalance:** To precisely estimate the treatment effect, The implementation of IV methods estimates a conditional treatment distribution $P(T \mid Z, X)$ using $\{Z, X\}$ in the treatment regression stage, then learns **the counterfactual prediction function** $h(T, X)$ from the re-sampled treatment $\hat{T} \sim P(T|Z, X)$ and the variables $X$ to $Y$ directly:

$$h(\hat{T}, X) = \mathbb{E}[Y \mid \hat{T}, X] \quad (3)$$

In outcome regression stage, the confounders $X$ would affect the resampled treatment $\hat{T}$ obtained from stage 1 as shown in the figure 1(b), leading to imbalance of $X$ between different resampled treatment options in stage 2 since the lack of randomization (Cook et al., 2002). If the outcome model is misspecified, such confounder imbalance would bring confounding bias for outcome regression in previous IV based methods, especially with high dimensional $X$, introducing bias and large variance in the estimation (Schroeder et al., 2016), i.e., $h(T, X) = \mathbb{E}[Y \mid T, X]$ holds only in the same distribution, and $h(t, X) \neq \mathbb{E}[Y \mid do(T = t), X]$ out of the distribution.

Based on this, we propose to reduce the confounding bias from observed confounders by confounder balancing (details in Section 4.1) in the outcome regression and re-build the inverse problem (details in Section 4.2) for relationship for the counterfactual prediction function $h(T, X)$.

## 4. Methodology

In this section, we first introduce the proposed algorithm (CB-IV) and achieve balanced confounder representation for eliminating confounding bias in Section 4.1; then, with representation obtained from CB-IV algorithm, we re-identify the inverse relationship for response-outcome function and avoid ill-posed identification problem under general settings in Section 4.2. The results in Section 4.2 can justify that the algorithm (proposed in Section 4.1) can achieve a accurate and robust estimation.

### 4.1. Algorithm and Optimization

IV regression is the classical method for addressing the unmeasured confounders, but recent nonlinear IV-based methods suffer the confounding bias from the observed confounders as shown in the figure 1(b), leading to poor performance on outcome regression and treatment effect estimation in practice.

To address these challenges, we propose a Confounder Balanced IV Regression (CB-IV) algorithm to achieve confounder balancing in IV regression. Specifically, confounder balancing for removing the bias from observed confounders and IV regression for eliminating the bias from unmeasured confounders. The proposed CB-IV algorithm consists of three main components: (i) treatment regression, (ii) confounder balancing, and (iii) outcome regression.

For simplicity, we introduce the proposed CB-IV algorithm with binary treatment, but it can also be applied for continuous treatment as verified in experiments (Section 5.2.3). The continuous version of CB-IV algorithm is elabrated in Section D in Appendix.

**Treatment Regression:** In this part, we propose to regress treatment $T$ with IVs $Z$ and observed confounders $X$ directly, as the treatment regression stage did in the previous nonlinear IV-based method. Specifically, we estimate the conditional probability distribution of the treatments $\hat{P}(T|Z,X)$ with a logistic regression network $\pi_\mu(z_i,x_i)$ with learnable parameter $\mu$ for each unit $i$, and optimize the following loss function for treatment regression:

$$\begin{aligned}\mathcal{L}_T \quad = \quad & -\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(t_i\log\left(\pi_\mu(z_i,x_i)\right) \\ + \quad & (1-t_i)\left(1-\log\left(\pi_\mu(z_i,x_i)\right)\right))\end{aligned} \tag{4}$$

**Confounder Balancing:** After treatment regression, we can obtain the causal graph as shown in the figure 1(b), where the observed variables $X$ would become the confounders for outcome regression. To address this problem, we propose to learn a representation of $X$ (i.e., $C = f_\theta(X)$) with a representation network $f_\theta(\cdot)$ with learnable parameter $\theta$, and minimize the discrepancy of distributions for different treatment arms to achieve $C \perp \hat{T}$ for confounder balancing:

$$\begin{aligned}\operatorname{disc}(\hat{T}, f_\theta(X)) = \mathrm{IPM}(&\{f_\theta(x_i)\hat{P}(t_i=0\mid z_i,x_i)\}_{i=1}^{n}, \\ &\{f_\theta(x_i)\hat{P}(t_i=1\mid z_i,x_i)\}_{i=1}^{n})\end{aligned} \tag{5}$$

where $\{f_\theta(x_i)\hat{P}(t_i=k\mid z_i,x_i)\}_{i=1}^{n}, k\in\{0,1\}$ denotes the distribution of representation $C=f_\theta(x_i)$ in the group $T=k$ given the $\hat{P}(t_i\mid z_i,x_i)$. The constraint term has a another choice that force $f_\theta(X)$ and original $T$ to be independent directly, $f_\theta(X)\perp T$.

Although many integral probability metrics (IPMs) can be used to measure the discrepancy of distributions, there is no known way or a simple method for some function families to compute IPM or its gradients efficiently. As a distance measure widely used in deep learning, Wass distance have consistent estimators which can be efficiently computed in the finite sample case (Shalit et al., 2017; Sriperumbudur et al., 2012) and achieves many breakthroughs (Arjovsky et al., 2017; Cuturi & Doucet, 2014). In CFR (Shalit et al., 2017) and DR-CFR (Hassanpour & Greiner, 2019b), practitioners adopt Wasserstein distance (Wass) to calculate the dissimilarity of distributions from different treatment arms and fit a balanced representation by minimizing the discrepancy. For the sake of fairness, in binary (or multi-valued) treatment $T$ cases, we uniformly use Wass distance (Cuturi & Doucet, 2014) as the discrepancy metrics. More discussion on Wass distance is given in Section G.2 in Appendix.

Besides, for continuous treatment $T$, we learn a ”balanced” representation (i.e., $C$) of the observed confounders $X$ as $C = f_\theta(X)$ via mutual information (MI) minimization constraints (Cheng et al., 2020): firstly, we use variational distribution $Q_\psi(\hat{T}\mid C)=\mathcal{N}(\mu_\psi(C),\sigma_\psi(C))$ parameterized by neural networks $\{\mu_\psi,\sigma_\psi\}$ to approximate the true conditional distribution $P(\hat{T}\mid C)$; then, we minimize the log-likelihood loss function of variational approximation $Q_\psi(\hat{T}\mid C)$ with $n$ samples to estimate MI:

$$\operatorname{disc}(\hat{T},C)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\log Q_\psi\left(\hat{t}_i\mid c_i\right)-\log Q_\psi\left(\hat{t}_j\mid c_i\right)\right].$$

where, $C=f_\theta(X)$. We adopt an alternating training strategy to iteratively optimize $Q_\psi(\hat{T}\mid C)$ and the network $C=f_{\theta(X)}$ to implement balanced representation in the Confounder Balancing. The continuous version of CB-IV algorithm is elabrated in Section D in Appendix.

**Outcome Regression:** Finally, we propose to regress the outcome with the estimated treatment $\hat{T}\sim P(T|Z,X)$ obtained in treatment regression module and the representation of confounders $C=f_\theta(X)$ obtained in confounder balancing module. With considering that high dimensional representation $f_\theta(X)$ would induce the loss of treatment information in outcome regression function $h_\xi(\hat{T},f_\theta(X))$ (Shalit et al., 2017) with single neural network $h_\xi(\cdot)$. We propose to regress the potential outcomes (i.e., $Y(do(T=1),X)$ and $Y(do(T=0),X)$) by optimizing $h_{\xi^0}(f_\theta(X))$ and $h_{\xi^1}(f_\theta(X))$ as two different regression network with learnable parameters $\xi^0$ and $\xi^1$, respectively:

$$\mathcal{L}_Y=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\sum_{t_i\in\{0,1\}}h_{\xi^{t_i}}(f_\theta(x_i))\hat{P}(t_i\mid z_i,x_i)\right)^2 \tag{6}$$

where $\hat{P}(t_i = 1 \mid z_i, x_i) = \pi_\mu(z_i, x_i)$ and $f_\theta(x_i)$ are derived from treatment regression module and confounder balancing module, respectively.

**Optimization:** Like the optimization of the previous IV regression method, we also set two-stage optimization for our algorithm. In the first stage, we optimize the treatment regression module $\pi_\mu$ by minimizing the loss $\mathcal{L}_T$ as defined in Eq. (4). In the second stage, then, we simultaneously optimize the confounder balancing and outcome regression modules by setting the balanced learning representations $C$ as a kind of regularization on the outcome regression model, with the following loss:

$$\min_{\theta, \xi^0, \xi^1} \mathcal{L}_Y + \alpha \, \text{disc}(\hat{T}, f_\theta(X)) \tag{7}$$

where $\alpha$ is a trade-off hyper-parameter.

Then, the average treatment effect can be estimated by

$$\widehat{ATE} = \mathbb{E}[h_{\xi^1}(f_\theta(X)) - h_{\xi^0}(f_\theta(X))]. \tag{8}$$

The details of pseudo-code (Algorithm 1) and the network structures (Table 5) of CB-IV are provided in Section E.1 in Appendix. Besides, the discussion of hyper-parameters $\alpha$ (Figure 3) is detailed in Section E.2 in Appendix.

Like (Shalit et al., 2017), the expected treatment effect estimation error $\epsilon(h, \theta)$ can be bounded by the standard generalization-error and the distance between the treated and control distributions induced by the representation:

$$\begin{aligned} \epsilon(h, \theta) \quad \leq & \quad 2 \left( \epsilon_F^{t=0}(h, \theta) + \epsilon_F^{t=1}(h, \theta) \right. \\ + & \quad \left. B_\theta IPM_{\mathrm{G}} \left( p_\theta^{t=1}, p_\theta^{t=0} \right) - 2\sigma_Y^2 \right) \end{aligned} \tag{9}$$

where $\epsilon_F^{T=t}(h, \theta) = \int_{\mathcal{X}} \ell_2(y, h_{\xi^t}(f_\theta(x))) p^{T=t}(x) dx$ for $t \in \{0, 1\}$; $p^{T=t}(x)$ denotes the PDF of $x$ given $T = t$; $p_\theta^{T=t} = \{f_\theta(x_i)\}_{i:t_i=t}$; $B_\theta$ is a constant; $\sigma_Y^2$ is the expected variance of $Y$. More discussion on Error Bound is given in Section G.3 in Appendix.

### 4.2. Inverse Problem for Response-Outcome Function

Recent IV methods (Hartford et al., 2017; Newey & Powell, 2003; Lin et al., 2019) define an inverse problem for the counterfactual prediction function $h(T, X)$ with two observable functions $\mathbb{E}[Y \mid T, X]$ and $P(T \mid Z, X)$:

$$\mathbb{E}[Y \mid Z, X] = \int h(T, X) dP(T \mid Z, X) \tag{10}$$

The inverse relationship for $h(T, X)$ holds only under *the additive noise assumption* on response-outcome function:

$$Y = g(T, X) + U, \mathbb{E}[U \mid Z] = 0 \tag{11}$$

Nevertheless, in reality, the outcome functions are agnostic and cannot be artificially controlled and assumed. In contrast, the treatment made by human decision-making is always traceable. For example, consider a promotional activity that will affect the buying tendency of people homogeneously, but it is not easy to discuss the impact on the employment rate of these people in the future. Thus, we believe Homogeneous Instrument-Treatment Association (Hartwig et al., 2020; 2021) is a more common setting in the real-world. Based on the Homogeneous Instrument-Treatment Assumption, we model a more general causal relationship by relaxing the additive assumption to multiplicative assumption on response-outcome function as:

$$T = f_1(Z, X) + f_2(X, U) \tag{12}$$

$$Y = g_1(T, X) + g_2(T) g_3(U) + g_4(X, U), Z \perp U, X \tag{13}$$

where $f_i(\cdot), g_j(\cdot)$ are unknown and potentially non-linear continuous functions. $g_2(T) g_3(U)$ denotes the multiplicative terms of $U$ with $T$ (e.g., $U^2 T - UT + U$), and we define it as *the multiplicative assumption*. The completeness of $\mathbb{P}(T \mid Z, X)$ and $\mathbb{P}(Y \mid T, X)$ guarantees uniqueness of the solution (Newey & Powell, 2003). Binary treatment and outcome cases can be modeled similarly (Section C).

The Eqs. (12) & (13) are a general form of the homogeneity assumptions, and **the Counterfactual Prediction Function** can be re-defined as:

$$\begin{aligned} & \mathbb{E}[Y \mid \text{do}(T), C, X] = \mathbb{E}[h(T, C) \mid T, C, X] \\ = & \quad \mathbb{E}\left[ g_1^C(T, C) + g_2(T) \mathbb{E}[g_3(U) \mid C] \right. \\ + & \quad \left. \mathbb{E}\left[ g_4(X, U) \mid C \right] \mid T, C, X \right] \end{aligned} \tag{14}$$

where $h(T, C)$ is the conditional expectation of $Y$ given the observables $T$ and disentangled representation $C$. We transform $g_1(T, X)$ as $g_1^C(T, C)$ with the disentangled representation $C = f_\theta(X)$, satisfying $\mathbb{E}[g_1^C(T, C) \mid T, C, X] = \mathbb{E}[g_1(T, X) \mid T, C, X]$. $\mathbb{E}[g_3(U) \mid C]$ and $\mathbb{E}[g_4(X, U) \mid C]$ are constant for the specified $C$.

Under the causal relationship (12) & (13), we show the inverse problem for the response-outcome function, implying that the identification of the counterfactual prediction function can be identified, as follows:

**Theorem 4.1.** ***Inverse Relationship of Eqs. (12) & (13).*** *If the learned representation of observed confounders $C = f_\theta(X)$ is independent with the estimated treatment $\hat{T}$, then the counterfactual prediction function $h(T, C)$ can be identified with instrumental variables $Z$ and representation $C$. Then, we can establish an inverse relationship for $h(T, C)$ given $\mathbb{E}[Y \mid Z, C, X]$ and $P(T \mid Z, X)$, as follow:*

$$\mathbb{E}[Y \mid Z, C, X] = \int \left[ h(T, C) \right] dP(T \mid Z, X) \tag{15}$$

*where, $dP(T \mid Z, X)$ is the conditional treatment distribution. The proof is given in Section B in Appendix.*

Based on our proposed counterfactual prediction function $h(T, C)$ with the balanced representation $C$, similarly, we can also establish the inverse relationship of Eq. (11) as:

$$\mathbb{E}[Y \mid Z, C, X] = \int h(T, C) dP(T \mid Z, X) \tag{16}$$

where $\mathbb{E}[h(T, C) \mid Z, C, X] = \mathbb{E}[h(T, X) \mid Z, X]$, which is consistent with Eq. (10) under Assumption 3.4.

Combining confounder balancing in IV methods, CB-IV solves the inverse problem under Assumptions 3.3 or 3.4, and achieves a more accurate and robust estimation.

**Remark:** Sufficient assumptions for identification of average treatment effect (Imbens & Angrist, 1994; Newey & Powell, 2003; Hernan & Robins, 2010; Hartwig et al., 2020) with the instruments inculude: homogeneity in the causal effect of $T$ on $Y$ or homogeneity in the association of $Z$ with $T$. To avoid ill-posed identification problem (Kress et al., 1989; Newey & Powell, 2003), we follow these two identification assumptions, and focus on resolving the inverse problems. Under homogeneity assumptions, with the balanced representations $C = f_\theta(X)$, CB-IV eliminates the bias from observed confounders $X$ and guarantees uniqueness of the solution $h(T, C)$ from the inverse relationship.

## 5. Experiments

We evaluate our approach on both synthetic and real-world datasets. Although the proposed algorithm was introduced by setting the treatment as binary in section 4.1, we demonstrate the effectiveness of our approach with both binary and continuous treatment settings.

### 5.1. Baselines

We compare the proposed algorithm (CB-IV) with two groups of methods. One group is **IV based methods**: (1) *DeepIV-LOG* and *DeepIV-GMM* (Hartford et al., 2017): In the first stage, DeepIV models the treatment network with logistic regression network (LOG) or gaussian mixture models (GMM); (2) *KernelIV* (Singh et al., 2019) and *DualIV* (Muandet et al., 2020): they implement 2-stage regression with different dictionaries of basis functions from reproducing kernel Hibert spaces; (3) *OneSIV* (Lin et al., 2019): OneSIV merges the two stages to leverage the outcome to estimate the treatment distribution; (4) *DFIV* (Xu et al., 2021): DFIV uses neural networks to fit non-linear models to replace the linear counterparts in the conventional 2SLS approach. The other group is **confounder balancing methods**: (1) *DFL* (Xu et al., 2021): DFL, an ablation experiment of DFIV, performs the nonlinear outcome regression directly without using instrumental variables; (2) *DirectRep* and *CFR* (Johansson et al., 2016; Shalit et al., 2017): Both DirectRep and CFR learn the representation of the observed confounders, but the former does not make any constraints, and the latter requires the learned representation to be independent of the treatments; (3) *DRCFR* (Hassanpour & Greiner, 2019b): DRCFR identifies and balances the confounders from all observed variables.

Note that *OneSIV* can be seen as an ablation version of our *CB-IV* algorithm without confounder balancing, and *DirectRep* and *CFR* are the ablation versions of our *CB-IV* algorithm without IV regression.

### 5.2. Experiments on Synthetic Datasets

#### 5.2.1. Dataset.

In **binary treatment cases**, similar to (Hassanpour & Greiner, 2019b), we generate the synthetic datasets satisfying homogeneity assumption, as follows: the latent variables $\{Z, X, U\}$ drive from $Z_1, \cdots Z_{m_Z} \sim \mathcal{N}(0, \mathrm{I}_{m_Z}), X_1, \cdots X_{m_X}, U_1, \cdots U_{m_U} \sim \mathcal{N}(0, \Sigma_{m_X + m_U})$ where $m_Z$, $m_X$ and $m_U$ are the dimensions of instruments $Z$, observed confounders $X$ and unobserved confounders $U$, respectively. $\mathrm{I}_{m_Z}$ denotes $m_Z$ degree identity matrix, $\Sigma_{m_X + m_U} = \mathrm{I}_{m_X + m_U} * 0.95 + \mathbb{1}_{m_X + m_U} * 0.05$ means that all elements except diagonal are 0.05 in the covariance matrix, and $\mathbb{1}_{m_X + m_U}$ denotes $m_X + m_U$ degree all-ones matrix. The treatment variable $T$ and outcome variable $Y$ are generated as follows:

$$\begin{aligned} P(T \mid Z, X) &= \tfrac{1}{1 + \exp\left(-(\sum_{i=1}^{m_Z} Z_i X_i + \sum_{i=1}^{m_X} X_i + \sum_{i=1}^{m_U} U_i)\right)}, \\ T &\sim Bernoulli(P(T \mid Z, X)), m_X > m_Z \quad (17) \\ Y(T, X, U) &= \tfrac{T}{m_X + m_U}(\textstyle\sum_{i=1}^{m_X} X_i^2 + \sum_{i=1}^{m_U} U_i^2) \\ &+ \tfrac{1-T}{m_X + m_U}(\textstyle\sum_{i=1}^{m_X} X_i + \sum_{i=1}^{m_U} U_i) \quad (18) \end{aligned}$$

where $Bernoulli(P(T \mid Z, X))$ is the true logging policy of the treatments $T$. Eqs. (17) & (18) is a common setting used by Hassanpour & Greiner (2019a;b); Wu et al. (2022).

As for **continuous treatment cases**, demand Datasets satisfying homogeneity assumption (that applied in DeepIV (Hartford et al., 2017), KernelIV (Singh et al., 2019), DualIV (Muandet et al., 2020) and DFIV (Xu et al., 2021)) is a choice, and we report mean squared error (MSE) and its standard deviations over 10 trials: the outcome variabl is $Y = 100 + (10 + T)X_1\psi_{X_2} - 2T + E$; the treatment variable is $T = 25 + (Z + 3)\psi_{X_2} + U$; $\psi_{X_2} = 2\left((X_2 - 5)^4/600 + \exp\left[-4(X_2 - 5)^2\right] + X_2/10 - 2\right)$; where $X_1 \in \{1, \ldots, 7\}$, $X_2 \sim \mathrm{unif}(0, 10)$, $Z, U \sim \mathrm{N}(0, 1)$ and $E \sim \mathrm{N}(0.5U, 0.75)$. In this case, the instrument variable is $Z$, the treatment variable is $T$, the observed variables are $\{X_1, X_2\}$, the outcome variable is $Y$, the unmeasured confounder is $\{U, E\}$.

#### 5.2.2. Results in binary treatment cases.

The results of treatment effect estimation in binary treatment cases are reported in Table 1, where we use Syn-$m_Z$-$m_X$-$m_U$ to denote the synthetic dataset with $m_Z$ instruments, $m_X$ observed confounders and $m_U$ unobserved con-

*Table 1.* The results of ATE estimation, including bias (mean(std)), in binary treatment cases on Synthetic data with different settings (Syn-$m_Z$-$m_X$-$m_U$).

| Within-Sample | | | | |
|---|---|---|---|---|
| **Method** | **Syn-1-4-4** | **Syn-2-4-4** | **Syn-2-10-4** | **Syn-2-4-10** |
| **DeepIV-LOG** | 1.055(0.011) | 1.057(0.008) | 1.092(0.009) | 1.020(0.008) |
| **DeepIV-GMM** | 0.934(0.011) | 0.874(0.019) | 0.768(0.023) | 0.925(0.017) |
| **KernelIV** | 0.495(0.056) | 0.457(0.054) | 0.765(0.028) | 0.624(0.062) |
| **DualIV** | 1.469(0.072) | 1.423(0.076) | 1.719(0.076) | 1.534(0.073) |
| **OneSIV** | 0.823(0.075) | 0.661(0.096) | 0.689(0.054) | 0.850(0.073) |
| **DFIV** | 0.852(0.010) | 0.860(0.007) | 0.851(0.007) | 0.886(0.009) |
| **DFL** | 0.840(0.002) | 0.851(0.002) | 0.838(0.002) | 0.831(0.004) |
| **DirectRep** | 0.172(0.017) | 0.163(0.008) | 0.118(0.017) | 0.199(0.016) |
| **CFR** | 0.172(0.016) | 0.158(0.015) | 0.105(0.020) | 0.198(0.018) |
| **DRCFR** | 0.151(0.056) | 0.136(0.034) | **0.063(0.044)** | 0.154(0.032) |
| **CB-IV** | **0.038(0.071)** | **0.016(0.047)** | 0.077(0.041) | **0.009(0.065)** |
| **Out-of-Sample** | | | | |
| **Method** | **Syn-1-4-4** | **Syn-2-4-4** | **Syn-2-10-4** | **Syn-2-4-10** |
| **DeepIV-LOG** | 1.055(0.010) | 1.057(0.008) | 1.093(0.009) | 1.020(0.008) |
| **DeepIV-GMM** | 0.933(0.011) | 0.874(0.019) | 0.768(0.023) | 0.925(0.017) |
| **KernelIV** | 0.495(0.055) | 0.458(0.052) | 0.765(0.028) | 0.625(0.063) |
| **DualIV** | 1.472(0.079) | 1.467(0.076) | 1.732(0.072) | 1.513(0.066) |
| **OneSIV** | 0.822(0.076) | 0.661(0.095) | 0.690(0.053) | 0.851(0.073) |
| **DFIV** | 0.851(0.009) | 0.860(0.007) | 0.851(0.007) | 0.886(0.009) |
| **DFL** | 0.840(0.002) | 0.851(0.002) | 0.838(0.002) | 0.831(0.004) |
| **DirectRep** | 0.172(0.016) | 0.164(0.009) | 0.116(0.015) | 0.199(0.014) |
| **CFR** | 0.172(0.015) | 0.159(0.018) | 0.103(0.019) | 0.198(0.016) |
| **DRCFR** | 0.151(0.055) | 0.137(0.035) | **0.062(0.045)** | 0.154(0.032) |
| **CB-IV** | **0.037(0.075)** | **0.017(0.046)** | 0.075(0.040) | **0.010(0.064)** |

*Table 2.* The results of latent outcome estimation, including MSE (mean(std)), in continuous treatment cases on Demand datasets with different settings (Demand-$\gamma$-$\lambda$).

| Within-Sample | | | |
|---|---|---|---|
| **Method** | **Demand-0-1** | **Demand-0-5** | **Demand-5-1** |
| **DeepIV-LOG** | - | - | - |
| **DeepIV-GMM** | 1356(343.5) | 3102(744.4) | 1465(253.3) |
| **KernelIV** | 1526(141.7) | >5000 | 1428(227.3) |
| **DualIV** | >5000 | >5000 | >5000 |
| **OneSIV** | >5000 | >5000 | >5000 |
| **DFIV** | 195.2(9.342) | 1205(1740) | 197.2(16.80) |
| **DFL** | 195.9(11.13) | 1159(1902) | 200.3(8.916) |
| **DirectRep** | 191.2(5.514) | 888.6(1077) | 440.1(117.3) |
| **CFR** | 193.3(5.561) | 465.3(181.4) | 449.6(161.0) |
| **DRCFR** | 427.2(162.0) | 391.6(28.21) | 405.8(105.9) |
| **CB-IV** | **165.0(5.959)** | **234.1(30.06)** | **167.7(6.783)** |
| **Out-of-Sample** | | | |
| **Method** | **Demand-0-1** | **Demand-0-5** | **Demand-5-1** |
| **DeepIV-LOG** | - | - | - |
| **DeepIV-GMM** | 1006(313.7) | 2829(724.6) | 1151(284.1) |
| **KernelIV** | 994.9(146.2) | 5435(435.2) | 1004(216.7) |
| **DualIV** | >5000 | >5000 | >5000 |
| **OneSIV** | >5000 | >5000 | >5000 |
| **DFIV** | 190.5(8.977) | 668.3(566.7) | 196.2(16.66) |
| **DFL** | 182.9(11.52) | 597.6(622.1) | 189.7(7.422) |
| **DirectRep** | 193.9(7.380) | 689.6(692.1) | 489.9(121.1) |
| **CFR** | 192.0(8.932) | 417.3(123.5) | 469.7(140.7) |
| **DRCFR** | 532.4(199.5) | 497.3(26.37) | 470.5(143.4) |
| **CB-IV** | **172.9(5.340)** | **224.3(18.06)** | **165.8(7.142)** |

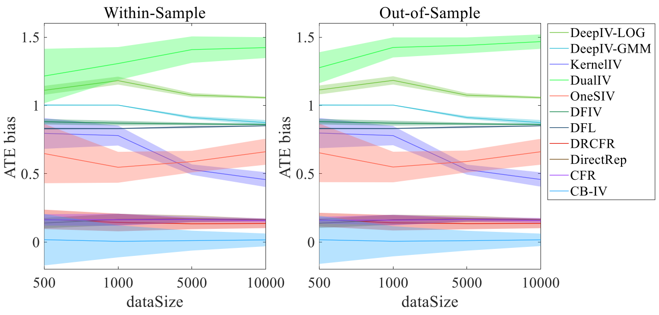

*Figure 2.* Results of CB-IV on Syn-2-4-4 by varying sample size.

founders. For each setting (such as *Syn-1-4-4*, *Syn-2-4-4*, *Syn-2-10-4*, *Syn-2-4-10*), we sample 10,000 units and perform 10 replications to report the mean and the standard deviation (std) of the bias of the average treatment effect (ATE) estimation, where *within-sample* error is computed over the training sets and *out-of-sample* error is over the test set. From the results in Table 1, we have the following observations: (1) For IV based methods, more valid IVs would bring more accuracy on treatment effect estimation by comparing with the results of setting *Syn-1-4-4* and setting *Syn-2-4-4*. (2) For confounder balancing methods, high dimension of unmeasured confounder would lead to poor performance by comparing with the results of setting *Syn-2-4-4* and setting *Syn-2-4-10*. (3) The existence of observed confounders would result in the poor performance of the IV-based methods, even worse than the confounder balancing-based methods because traditional IV-based methods ignored the bias of observed confounders in their second-stage regression. (4) Considering confounder balancing in IV regression, our CB-IV improves considerably over the traditional IV-based methods and achieves better performance than confounder balancing methods in most settings. When the observed confounders are high-dimensional, the low-dimensional instruments' information might get lost, and CB-IV would be equivalent to CFR.

As a data-driven representation learning method, CB-IV requires more training data to ensure performance. Hence we implement experiments with different data sizes (500, 1000, 5000, 10000) on *Syn-2-4-4* to study its impact on model performance. Figure (2) shows that the bias of the average treatment effect estimation of CB-IV is low in different data sizes, but the variance is huge above small data sets. As the number of data increases, the variance of CB-IV will decrease linearly. When the amount of data exceeds 5000, the upper bound of CB-IV's estimation will be lower than the lower bound of all baselines. In conclusion, our method relies more on a large amount of data. One possible solution is to perform each experiment many times (e.g., ten duplicates) and then take the average value to reduce the variance, but this is not the paper's focus. Due to limited space, more Experiments about different variables used in the different stages can be found in Section F in Appendix.

### 5.2.3. Results in continuous treatment cases.

We adjust the difficulty of the simulation and perform experiments to increase the importance of instrumental variables in the structure function of $T$ (e.g., adjust $\gamma$ and $\lambda$

*Table 3.* The results of latent outcome estimation, including MSE (mean(std)), in continuous treatment cases on Demand datasets.

| Within-Sample | | | | |
|---|---|---|---|---|
| **Method** | **500** | **1000** | **5000** | **10000** |
| **DeepIV-LOG** | - | - | - | - |
| **DeepIV-GMM** | >5000 | >5000 | 3163.3(266.4) | 1356(343.5) |
| **KernelIV** | 3078(647.2) | 2363(270.7) | 1692(72.6) | 1526(141.7) |
| **DualIV** | >5000 | >5000 | >5000 | >5000 |
| **OneSIV** | >5000 | >5000 | >5000 | >5000 |
| **DFIV** | 240.0(381.7) | 152.4(52.83) | 198.9(30.62) | 195.2(9.342) |
| **DFL** | 141.4(26.42) | 173.2(29.90) | 196.8(17.82) | 195.9(11.13) |
| **DirectRep** | 138.7(24.01) | 153.4(16.67) | 193.0(12.87) | 191.2(5.514) |
| **CFR** | 126.9(20.98) | 161.7(20.99) | 191.6(10.24) | 193.3(5.561) |
| **DRCFR** | 705.5(462.9) | 503.0(240.5) | 419.0(126.1) | 427.2(162.0) |
| **CB-IV** | **117.6(23.25)** | **142.0(16.11)** | **164.6(7.443)** | **165.0(5.959)** |
| **Out-of-Sample** | | | | |
| **Method** | **500** | **1000** | **5000** | **10000** |
| **DeepIV-LOG** | - | - | - | - |
| **DeepIV-GMM** | >5000 | >5000 | 3360(483.8) | 1006(313.7) |
| **KernelIV** | 2859(660.9) | 2280(547.9) | 1142(170.3) | 994.9(146.2) |
| **DualIV** | >5000 | >5000 | >5000 | >5000 |
| **OneSIV** | >5000 | >5000 | >5000 | >5000 |
| **DFIV** | 764.4(415.1) | 404.9(133.1) | 214.4(30.66) | 190.5(8.977) |
| **DFL** | 358.7(47.32) | 261.3(35.68) | 192.7(14.46) | 182.9(11.52) |
| **DirectRep** | 271.8(25.76) | **222.3(9.575)** | 199.8(5.453) | 193.9(7.380) |
| **CFR** | **266.2(28.45)** | 225.9(11.75) | 195.8(11.338) | 192.0(8.932) |
| **DRCFR** | 799.8(467.5) | 621.7(275.9) | 511.0(155.04) | 532.4(199.5) |
| **CB-IV** | 291.4(39.33) | 229.1(42.22) | **179.4(4.221)** | **172.9(5.340)** |

[*] The results of IV-based methods are consistent with those of the report in DeepIV (Hartford et al., 2017), KernelIV (Singh et al., 2019), DualIV (Muandet et al., 2020) and DFIV (Xu et al., 2021). The difference is that they scale the results by log10, but we don't.

*Table 4.* The results of ATE estimation, including bias (mean(std)), on real-world data with different settings (Data-$m_Z$-$m_X$-$m_U$).

| Within-Sample | | | | |
|---|---|---|---|---|
| **Method** | **IHDP-2-6-0** | **IHDP-2-4-2** | **Twins-5-8-0** | **Twins-5-5-3** |
| **DeepIV-LOG** | 2.874(0.058) | 2.623(0.065) | 0.013(0.021) | 0.024(0.011) |
| **DeepIV-GMM** | 3.776(0.032) | 3.740(0.040) | 0.019(0.005) | 0.022(0.004) |
| **KernelIV** | 3.061(0.305) | 2.994(0.463) | - | - |
| **DualIV** | 0.593(0.221) | 0.658(0.243) | - | - |
| **OneSIV** | 1.725(0.375) | 1.741(0.342) | 0.008(0.019) | 0.008(0.017) |
| **DFIV** | 3.554(0.089) | 3.622(0.104) | 0.027(0.001) | 0.026(0.000) |
| **DFL** | 3.202(0.050) | 3.199(0.037) | 0.062(0.059) | 0.085(0.005) |
| **DirectRep** | 0.068(0.056) | 0.460(0.071) | 0.017(0.017) | 0.019(0.025) |
| **CFR** | 0.085(0.058) | 0.483(0.064) | 0.011(0.017) | 0.022(0.018) |
| **DRCFR** | 0.055(0.064) | 0.434(0.069) | 0.011(0.022) | 0.012(0.017) |
| **CB-IV** | **0.012(0.388)** | **0.160(0.250)** | **0.007(0.027)** | **0.001(0.025)** |
| **Out-of-Sample** | | | | |
| **Method** | **IHDP-2-6-0** | **IHDP-2-4-2** | **Twins-5-8-0** | **Twins-5-5-3** |
| **DeepIV-LOG** | 2.876(0.055) | 2.623(0.069) | 0.014(0.021) | 0.024(0.011) |
| **DeepIV-GMM** | 3.777(0.035) | 3.739(0.042) | 0.019(0.005) | 0.022(0.004) |
| **KernelIV** | 3.070(0.306) | 3.023(0.440) | - | - |
| **DualIV** | 0.564(0.266) | 0.715(0.355) | - | - |
| **OneSIV** | 1.729(0.372) | 1.735(0.343) | 0.008(0.019) | 0.008(0.017) |
| **DFIV** | 3.554(0.090) | 3.623(0.106) | 0.027(0.001) | 0.026(0.000) |
| **DFL** | 3.204(0.050) | 3.199(0.038) | 0.062(0.058) | 0.085(0.005) |
| **DirectRep** | 0.061(0.082) | 0.457(0.076) | 0.016(0.018) | 0.019(0.025) |
| **CFR** | 0.079(0.081) | 0.480(0.069) | 0.011(0.016) | 0.022(0.018) |
| **DRCFR** | 0.045(0.095) | 0.432(0.067) | 0.011(0.022) | 0.012(0.017) |
| **CB-IV** | **0.015(0.393)** | **0.158(0.254)** | **0.006(0.027)** | **0.002(0.025)** |

[*] Most confounders are discrete variables and the outcome is binary variable in Twins data. The results of kernel-based IV methods in Twins are NaN. We use '-' to denote it.

in $T = 25 + \gamma Z + (\lambda Z + 3)\psi_{X_2} + U$), we name it as Demand-$\gamma$-$\lambda$. *Demand-0-1* is the original Demand data with $T = 25 + (Z + 3)\psi_{X_2} + U$. In *Demand-0-5* with $T = 25 + (5 * Z + 3)\psi_{X_2} + U$, we increase the information of the instrumental variable and amplify the confounding bias. As for *Demand-5-1* with $T = 25 + 5 * Z + (Z + 3)\psi_{X_2} + U$, we increase the information of the instrumental variable but keep the confounding bias unchanged.

The experimental results (reported in Table 2) shows that (i) if the information of instrumental variables and confounders increases, all methods will become worse, but the confounder balance based methods (e.g., CFR) still perform much better than the pure IV based methods (e.g., DeepIV). (ii) If we only increase the information of the instrumental variable, the results of the pure IV-based methods and our CB-IV are almost unchanged due to the same confounding bias. However, the balanced representation methods are basically worse, which is a very magical phenomenon. One conjecture is that the fluctuation of $T$ affects the change of $Y$. Perhaps we should regularize the treatment variables and outcome variables before regressing them. Anyway, the confounding bias from the treatment regression stage is a critical problem in IV-based methods.

Like the binary treatment studies in this paper, on this classical simulation data Demand-0-1 (Table 3) with different data sizes (500, 1000, 5000, 10000), the confounder balancing based methods (without using IV) still perform much better than the pure IV-based methods. Considering confounder balancing in IV regression, our CB-IV method improves considerably over the traditional IV-based methods and achieved better performance than confounder balancing-based methods in most settings. Nevertheless, our method still relies on large samples.

### 5.3. Experiments on Real-World Datasets

#### 5.3.1. Dataset.

We also check the performance of CB-IV method with experiments on two real-world datasets, which are adopted in Yao et al. (2018); Wu et al. (2022): IHDP tends to evaluate the effect of a specialist home visit on premature infants' cognitive test scores, and Twins aims to estimate the effect of the weight in twins on the infant's mortality.

**IHDP[3]:** The Infant Health and Development Program (IHDP) comprises 747 units (139 treated, 608 control). To develop the instrument variables, we generate 2-dimension random variables for each unit, i.e., $Z_1, \cdots Z_{m_Z} \sim \mathcal{N}(0, I_{m_Z}), m_Z = 2$. Then, we select 6 variables from the original data as the confounders, including $m_X$ variables as observed confounders $X$ and $m_U$ as unobserved $U$, where $m_X + m_U = 6$. The treatment assignment policy is $P(T \mid Z, X) = \frac{1}{1+\exp\left(-(\sum_{i=1}^{m_Z} Z_i X_i + \sum_{i=1}^{m_X} X_i) + \sum_{i=1}^{m_U} U_i)\right)}, T \sim$

[3] http://www.fredjo.com/

$Bernoulli(P(T \mid Z, X))$.

**Twins**[4]**:** Twins dataset is derived from all twins born in the USA between the years 1989 and 1991 (Almond et al., 2005). Similar to Yao et al. (2018), we select 5271 records from same-sex twins who weighed less than 2000 grams and had no missing characteristics. Then we generate 5-dimension random variables as the instrument variables and obtain $m_X$ variables as observed confounders $X$ and $m_U$ as unobserved $U$ to design the treatments $T$ according to the policy in Eq. (17).

### 5.3.2. Results.

We conduct our experiments over the 100 realizations of IHDP and 10 realizations of Twins with a 63/27/10 proportion of train/validation/test splits. In each realization, we shuffle the data and then redivide it into train/validation/test splits to simulate as many different data distributions as possible. **Data-**$m_Z$**-**$m_X$**-**$m_U$ means that there are $m_Z$ dimension instruments, $m_X$ observed confounders and $m_U$ unobserved confounders in the corresponding Data. We report the results in Table 4, including the mean and standard deviation (std) of the bias of average treatment effect estimation.

In the dataset without unmeasured confounders (*IHDP-2-6-0* and *Twins-5-8-0*), the performance of CB-IV is better than confounder balance methods (DRCFR, CFR), better than two-head methods (DirectRep), and the IV methods (DeepIV, KernelIV, DFIV) are the worst. DualIV and OneSIV have the best performance in the traditional IV methods on *IHDP* and *Twins*, respectively. When there are unmeasured confounders (*IHDP-2-4-2* and *Twins-5-5-3*), it is evident that the performance of the confounder balance methods decreased a lot. Still, the performance of CB-IV and IV methods are almost unaffected, which is in line with our expectations. CB-IV requires a larger amount of data to ensure the convergence of the variance. Because the training set of IHDP has only 471 samples, CB-IV has a small bias but a large variance. Despite this, in the presence of unobserved confounders, the upper bound of the error of CB-IV is much lower than these baselines. In general, CB-IV achieves the best performance among all baselines.

## 6. Conclusion

The majority of instrumental variable methods ignore the confounding bias in the outcome regression stage in nonlinear scenarios. A promising direction is to implement confounder balancing. Under sufficient identification assumption, we propose a Confounder Balanced IV Regression (CB-IV) algorithm to confirm this and solve two inverse problems under different Homogeneity Assumptions. Extensive experiments show that the proposed method achieves state-of-the-art performance in the average treatment effect estimation.

Like previous works on treatment effect estimation (Hartford et al., 2017; Xu et al., 2021; Shalit et al., 2017; Hassanpour & Greiner, 2019b), we also did not examine the statistical properties in inference (e.g., the convergence rates). It's generally challenging to analyze the statistical guarantees of inference after deep neural network training in multi-stage (Farrell et al., 2021). It is possible to use bootstrap to estimate the standard errors of the estimated treatment effects, and we leave this for future work.

## Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No. 62006207, No. 62037001), Key R & D Projects of the Ministry of Science and Technology (2020YFC0832500), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), Project by Shanghai AI Laboratory (P22KS00111), the Fundamental Research Funds for the Central Universities (226-2022-00142). Bo Li's research was supported by the National Natural Science Foundation of China (No.72171131); the Tsinghua University Initiative Scientific Research Grant (No. 2019THZWJC11); Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grants 2020AAA0108400 and 2020AAA0108403.

## References

Abadie, A. Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263, 2003.

Almond, D., Chay, K. Y., and Lee, D. S. The costs of low birth weight. *The Quarterly Journal of Economics*, 120 (3):1031–1083, 2005.

Angrist, J. D. and Krueger, A. B. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4):69–85, 2001.

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

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Athey, S., Imbens, G. W., and Wager, S. Approximate residual balancing: debiased inference of average treatment

[4] http://www.nber.org/data/

effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4): 597–623, 2018.

Buhlmann, P., Peters, J., and Ernest, J. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Carrasco, M., Florens, J.-P., and Renault, E. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751, 2007.

Chen, X. and Christensen, T. M. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1): 39–84, 2018.

Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., and Carin, L. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pp. 1779–1788. PMLR, 2020.

Cook, T. D., Campbell, D. T., and Shadish, W. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA, 2002.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pp. 685–693. PMLR, 2014.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Farrell, M. H., Liang, T., and Misra, S. Deep neural networks for estimation and inference. *Econometrica*, 89(1): 181–213, 2021.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

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

Hartwig, F. P., Bowden, J., Wang, L., Smith, G. D., and Davies, N. Average causal effect estimation via instrumental variables: the no simultaneous heterogeneity assumption. *arXiv preprint arXiv:2010.10017*, 2020.

Hartwig, F. P., Wang, L., Smith, G. D., and Davies, N. M. Homogeneity in the instrument-treatment association is not sufficient for the wald estimand to equal the average causal effect for a binary instrument and a continuous exposure. *arXiv preprint arXiv:2107.01070*, 2021.

Hassanpour, N. and Greiner, R. Counterfactual regression with importance sampling weights. In *IJCAI*, pp. 5880–5887, 2019a.

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019b.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Heckman, J. J. Econometric causality. *International statistical review*, 76(1):1–27, 2008.

Heckman, J. J., Urzua, S., and Vytlacil, E. Understanding instrumental variables in models with essential heterogeneity. *The review of economics and statistics*, 88(3): 389–432, 2006.

Hernán, M. A. and Robins, J. M. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, pp. 360–372, 2006.

Hernan, M. A. and Robins, J. M. Causal inference, 2010.

Imbens, G. W. and Angrist, J. D. Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, pp. 467–475, 1994.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kress, R., Maz'ya, V., and Kozlov, V. *Linear integral equations*, volume 82. Springer, 1989.

Kuang, K., Cui, P., Li, B., Jiang, M., and Yang, S. Estimating treatment effect in the wild via differentiated confounder balancing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 265–274, 2017.

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., and Jiang, Z. Causal inference. *Engineering*, 6(3):253–263, 2020a.

Kuang, Z., Sala, F., Sohoni, N., Wu, S., Córdova-Palomera, A., Dunnmon, J., Priest, J., and Re, C. Ivy: Instrumental variable synthesis for causal inference. In *AISTATS*, pp. 398–410. PMLR, 2020b.

Li, M., Wang, T., Zhang, H., Zhang, S., Zhao, Z., Miao, J., Zhang, W., Tan, W., Wang, J., Wang, P., Pu, S., and Wu, F. End-to-end modeling via information tree for one-shot natural language spatial video grounding. In *ACL*, 2022.

Li, S., Vlassis, N., Kawale, J., and Fu, Y. Matching via dimensionality reduction for estimation of treatment effects. In *IJCAI*, pp. 3768–3774, 2016.

Li, X.-H., Cao, C. C., Shi, Y., Bai, W., Gao, H., Qiu, L., Wang, C., Gao, Y., Zhang, S., Xue, X., et al. A survey of data-driven and knowledge-aware explainable ai. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

Lin, A., Lu, J., Xuan, J., Zhu, F., and Zhang, G. One-stage deep instrumental variable method for causal inference from observational data. In *IEEE International Conference on Data Mining (ICDM)*, pp. 419–428. IEEE, 2019.

Muandet, K., Mehrjou, A., Le Kai, S., and Raj, A. Dual instrumental variable regression. In *NeurIPS 2020*, 2020.

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

Pearl, J. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009a.

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

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

Pele, O. and Werman, M. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pp. 460–467. IEEE, 2009.

Rosenbaum, P. R. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D. B. Matching to remove bias in observational studies. *Biometrics*, pp. 159–183, 1973.

Schroeder, K., Jia, H., and Smaldone, A. Which propensity score method best reduces confounder imbalance? an example from a retrospective evaluation of a childhood obesity intervention. *Nursing research*, 65(6):465, 2016.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Singh, R. and Sun, L. Automatic kappa weighting for instrumental variable models of complier treatment effects. *arXiv preprint arXiv:1909.05244*, 2019.

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 4593–4605, 2019.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

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

Strehl, A., Langford, J., Kakade, S., and Li, L. Learning from logged implicit exploration data. *arXiv preprint arXiv:1003.0120*, 2010.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Wooldridge, J. M. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Wright, P. G. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

Wu, A., Yuan, J., Kuang, K., Li, B., Wu, R., Zhu, Q., Zhuang, Y. T., and Wu, F. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., and Gretton, A. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

Yuan, J., Wu, A., Kuang, K., Li, B., Wu, R., Wu, F., and Lin, L. Auto iv: Counterfactual prediction via automatic instrumental variable decomposition. *Transactions on Knowledge Discovery from Data*, 16(4):1–20, 2022.

Zhang, S., Jiang, T., Wang, T., Kuang, K., Zhao, Z., Zhu, J., Yu, J., Yang, H., and Wu, F. Devlbert: Learning deconfounded visio-linguistic representations. In *Multimedia*, pp. 4373–4382, 2020.

Zhang, S., Yao, D., Zhao, Z., Chua, T.-S., and Wu, F. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *SIGIR*, pp. 367–377, 2021.

## A. Nonlinear Case

**Example A.1.** *(Complicated nonlinear case).* $T = f(Z, X) + U = ZX + U, Y = g(T, X) + U = TX^2 + X + U$, *where* $Z \sim \mathcal{N}(0, 1), X, U \sim \mathcal{N}\left((0, 0), \begin{pmatrix} 1 & 0.05 \\ 0.05 & 1 \end{pmatrix}\right)$.

**Solution 1 (a)**. Stage 1, classical IV methods perform linear/nonlinear regression from $Z$ to $T$:

$$\mathbb{E}[T|Z] = \mathbb{E}[ZX + U|Z] = \mathbb{E}[ZX|Z] + \mathbb{E}[U|Z] = \mathbb{E}[X]Z = 0$$

Then, we get a wrong conclusion that $Z$ and $T$ are independent.

**Solution 1 (b)**. Stage 1, nonlinear IV regression variants perform linear/nonlinear regression from $\{Z, X\}$ to $T$:

$$\mathbb{E}[T|Z, X] = \mathbb{E}[ZX + U|Z, X] = \mathbb{E}[ZX|Z, X] + \mathbb{E}[U|Z, X] = XZ + \mathbb{E}[U|X]$$

where $\mathbb{E}[ZX|Z, X] = ZX$, because $Z$ and $X$ are independent. We define $\hat{T} = \mathbb{E}[T|Z, X] = XZ + \mathbb{E}[U|X]$ in the continuous case.

Stage 2, if we perform linear/nonlinear regression from $\{Z, X\}$ to $Y$:

$$\begin{aligned} \mathbb{E}[Y|Z, X] &= \mathbb{E}[TX^2 + X + U|Z, X] \\ &= \mathbb{E}[(ZX + U)X^2 + X + U|Z, X] \\ &= \mathbb{E}[(ZX^3 + X + U + UX^2|Z, X] \\ &= ZX^3 + X + \mathbb{E}[U|X](X^2 + 1) \\ &= (ZX + \mathbb{E}[U|X])X^2 + X + \mathbb{E}[U|X] \\ &= \hat{T}X^2 + X + \mathbb{E}[U|X] \\ &= g(\hat{T}, X) + \mathbb{E}[U|X] \end{aligned}$$

we will get the structure function ($g(\hat{T}, X) + \mathbb{E}[U|X]$) and an unbiased arverage treatment effect ($ATE_Z$) estimation of $Z$ on $Y$:

$$\begin{aligned} ATE_Z &= \mathbb{E}[Y|Z_1, X] - \mathbb{E}[Y|Z_0, X] \\ &= [\mathbb{E}[g(T'_1, X)] + \mathbb{E}[U|X]] - [\mathbb{E}[g(T'_0, X)] + \mathbb{E}[U|X]] \\ &= \mathbb{E}[g(T'_1, X)] - \mathbb{E}[g(T'_0, X)] \\ &= \mathbb{E}[(Z_1X + \mathbb{E}[U|X])X^2 + X) - (Z_0X + \mathbb{E}[U|X])X^2 + X)] \\ &= \mathbb{E}[Z_1X^3 - Z_0X^3] \\ &= \mathbb{E}[Z_1 - Z_0]\mathbb{E}[X^3] \\ &= 0 \end{aligned}$$

Nevertheless, we want to obtain the causal relationship ($ATE$) between the treatments $T$ and outcomes $Y$, instead of the average causal effect estimation ($ATE_Z$) of $Z$ on $Y$. $ATE$ and $ATE_Z$ are not equivalent. Therefore, We have to perform linear/nonlinear regression from $\{\hat{T}, X\}$ to $Y$ in stage 2, i.e., $\mathbb{E}[\hat{T}X^2 + X + U|\hat{T}, X], \hat{T} = \mathbb{E}[T|Z, X]$:

$$\mathbb{E}[Y|\hat{T}, X] = \mathbb{E}[TX^2 + X + U|\hat{T}, X] = \mathbb{E}[TX^2 + X|\hat{T}, X] + \mathbb{E}[U|X]$$

Obviously, $X$ would be a confounder ($\hat{T} = \mathbb{E}[T|Z, X]$ derives from $\{Z, X\}$, and $\{X, \hat{T}\}$ are the cause of $Y$) and these algorithms would get a biased causal effect between the $\hat{T}/T$ and $Y$ without prior knowledge of regression function. In other words, $T$ is related to $X$, so there may be multiple different solutions $\hat{g}$ of $\arg\min_{g'}\{\mathbb{E}[TX^2 + X|\hat{T}, X] - g'(T, X)\}$ and $\hat{g}$ may be different from true structural function $g$ without prior knowledge of regression function.

Fortunately, the unobserved confounders $U$ will no longer confound the causal relationship between $\hat{T}$ and $Y$ in stage 2 (figure 1(b)), and we only need to analyze and reduce the bias from the observed confounders $X$.

## B. Theorems

**Theorem B.1.** *(**Inverse Relationship of Eq. (13)**). If the learned representation of observed confounders $C = f_\theta(X)$ is independent with the estimated treatment $\hat{T}$, then the counterfactual prediction function $h(T, C)$ can be identified with instrumental variables $Z$ and representation $C$:*

$$h(T,C) = g_1^C(T,C) + g_2(T)\mathbb{E}[g_3(U)|C] + \mathbb{E}[g_4(X,U)|C] \tag{19}$$

*Then, we can establish an inverse relationship for $h(T,C)$ given $\mathbb{E}[Y \mid Z, C, X]$ and $P(T \mid Z, X)$, as follow:*

$$\mathbb{E}[Y \mid Z, C, X] = \int [h(T,C)]\, dP(T \mid Z, X) \tag{20}$$

*where, $dP(T \mid Z, X)$ is the conditional treatment distribution.*

*Proof.* In this paper, we model the causal relationship more general and relax the additive separability assumption to the multiplicative assumption (Eq. (12)(13)):

$$\begin{aligned} T &= f_1(Z,X) + f_2(X,U) \\ Y &= g_1(T,X) + g_2(T)g_3(U) + g_4(X,U), Z \perp U, X \end{aligned}$$

Then, we expect use the disentangled representation $C \perp T, C = f_\theta(X)$ to approximate the structural equation $Y = g_1^C(T,C) + g_2(T)g_3(U) + g_4(X,U)$.

Treatment Regression Stage, we perform nonlinear regression from $\{Z, X\}$ to $T$ using deep neural networks:

$$\begin{aligned} \mathbb{E}[T|Z,X] &= \mathbb{E}[f_1(Z,X) + f_2(X,U)|Z,X] \\ &= \mathbb{E}[f_1(Z,X)|Z,X] + \mathbb{E}[f_2(X,U)|Z,X] \\ &= f_1(Z,X) + \mathbb{E}[f_2(X,U)|X] \end{aligned}$$

where $\mathbb{E}[f_1(Z,X)|Z,X] = f_1(Z,X)$, because $Z$ and $X$ are independent. We define the conditional treatment distribution as $\hat{T} \sim P(T|Z,X)$.

Outcome Regression Stage, we perform linear/nonlinear regression from $\{Z, X\}$ to $Y = g_1^C(T,C) + g_2(T)g_3(U) + g_4(X,U)$ using deep neural networks:

$$\begin{aligned} \mathbb{E}[Y|Z,C,X] &= \mathbb{E}[g_1(T,X) + g_2(T)g_3(U) + g_4(X,U)|Z,C,X] \\ &= \mathbb{E}[g_1^C(T,f_\theta(X)) + g_2(T)g_3(U) + g_4(X,U)|Z,C,X] \\ &= \mathbb{E}[g_1^C(T,C) + g_2(T)g_3(U) + g_4(X,U)|Z,C,X] \\ &= \mathbb{E}[g_1^C(T,C)|Z,X] + \mathbb{E}[g_2(T)g_3(U)|Z,C,X] + \mathbb{E}[g_4(X,U)|C/X] \\ &= \int g_1^C(T,C)dP(T|Z,X) + \mathbb{E}[g_2(T)g_3(U)|Z,C,X] + \mathbb{E}[g_4(X,U)|C/X] \\ &= \int \left[g_1^C(T,C) + \mathbb{E}[g_4(X,U)|C]\right] dP(T|Z,X) + \mathbb{E}[g_2(T)g_3(U)|Z,C,X] \end{aligned} \tag{21}$$

where $P(T|Z,X)$ is the conditional treatment distribution, $\mathbb{E}[g_4^C(C,U)|X]$ is a constant for the specified $X/C$. Because $\mathbb{E}[g_1^C(T,C)|Z,X] = \int g_1^C(T,C)dP(T|Z,X)$: the completeness of $\mathbb{P}(T \mid Z, X)$ and $\mathbb{P}(Y \mid T, X)$ would guarantees uniqueness of the solution (Newey & Powell, 2003). The relationship in Equation (21) defines an inverse problem for $g_1$ in terms of two directly observable functions: $\mathbb{E}[Y|Z,X]$ and $P(T|Z,X)$. Eq. (5) in Hartford et al. (2017) and Eq. (6) in Lin et al. (2019) use same relationship to solve the inverse problem:

$$h(T,X) \equiv g(T,X) + \mathbb{E}[e \mid X]$$
$$\mathbb{E}[Y \mid X, Z] = \mathbb{E}[g(T,X) \mid X, Z] + \mathbb{E}[e \mid X] = \int \hbar(T,X) dF(T \mid X, Z)$$

where, again, $dF(T \mid X, Z)$ is the conditional treatment distribution in Hartford et al. (2017); Lin et al. (2019).

If $C \perp g_2(\hat{T})$, then $g_2(\hat{T}) \perp g_3(U) \mid Z, C$:

$$\begin{aligned} & \mathbb{E}[\mathbb{E}[g_2(\hat{T})g_3(U) \mid Z, C] \mid Z, X] \\ = \; & \mathbb{E}[\mathbb{E}[g_2(\hat{T}) \mid Z, C]\mathbb{E}[g_3(U) \mid Z, C] \mid Z, X] \\ = \; & \mathbb{E}[g_2(\hat{T})\mathbb{E}[g_3(U) \mid C] \mid Z, X] \\ = \; & \int g_2(T)\mathbb{E}[g_3(U) \mid C] dP(T|Z, X) \end{aligned}$$

Summarily, $\mathbb{E}[Y|Z,C,X] = \mathbb{E}[h(T,C)|Z,C,X] = \int \left[g_1^C(T,C) + g_2(T)\mathbb{E}[g_3(U)|C] + \mathbb{E}[g_4(X,U)|X]\right] dP(T|Z,X)$ . The counterfactual prediction function is $h(T,C) = g_1^C(T,X) + g_2(T)\mathbb{E}[g_3(U)|C] + \mathbb{E}[g_4(X,U)|X]$, and can be identified by IVs and balanced representation.

Then, we can establish an inverse relationship for $h(T,C)$ given $\mathbb{E}[Y \mid Z,C,X]$ and $P(T \mid Z,X)$, as follow:

$$\mathbb{E}[Y \mid Z,C,X] = \int \left[h(T,C)\right] dP(T \mid Z,X)$$

where, $dP(T \mid Z,X)$ is the conditional treatment distribution. □

## C. Binary Treatment and Binary Outcome Case

In this paper, we model a more general causal relationship and relax the additive separability assumption to the multiplicative assumption with Homogeneous Instrument-Treatment Association, as follows:

$$\begin{aligned} T &= f_1(Z,X) + f_2(X,U) & (22) \\ Y &= g_1(T,X) + g_2(T)g_3(U) + g_4(X,U), Z \perp U, X & (23) \end{aligned}$$

where $f_i(\cdot), g_j(\cdot)$ are unknown and potentially non-linear continuous functions. $g_2(T)g_3(U)$ denotes the multiplicative terms of $U$ with $T$ (e.g., $U^2T - UT + U$), and we define it as *the multiplicative assumption*. The completeness of $\mathbb{P}(T \mid Z,X)$ and $\mathbb{P}(Y \mid T,X)$ guarantees uniqueness of the solution (Newey & Powell, 2003). For binary treatment and binary outcome case, we can also model it similarly:

$$\begin{gathered} T \sim Bernoulli(P(T)), \text{where } P(T) = \tfrac{1}{1+\exp^{-(f_1(Z,X)+f_2(X,U))}}, \\ Y \sim Bernoulli(P(Y)), \text{where } P(Y) = \tfrac{1}{1+\exp^{-(g_1(T,X)+g_2(T)g_3(U)+g_4(X,U))}}, \\ \log\tfrac{P(T)}{1-P(T)} = f_1(Z,X) + f_2(X,U), \log\tfrac{P(Y)}{1-P(Y)} = g_1(T,X) + g_2(T)g_3(U) + g_4(X,U), Z \perp U, X \qquad (24) \end{gathered}$$

In this paper, all relevant theories and proofs can be transformed into binary cases. We can use the expectation of the samples to approximate the probability distribution of the data.

## D. Continuous Version of CB-IV

**Treatment Regression:** For continuous treatment $T$, we propose to regress treatment $T$ with IVs $Z$ and observed confounders $X$. Specifically, we estimate the conditional probability distribution[5] of the treatments $\hat{P}(T|Z,X) \sim \mathcal{N}(\mu_\omega(Z,X), \sigma_\omega(Z,X))$ with neural networks $\{\mu_\omega, \sigma_\omega\}$:

$$\phi(t \mid z_i, x_i) = \frac{1}{\sqrt{2\pi}\sigma_\omega(z_i,x_i)} \exp\left(-\frac{(t-\mu_\omega(z_i,x_i))^2}{2\sigma_\omega^2(z_i,x_i)}\right) \qquad (25)$$

where $\phi(\cdot)$ is probability density function for the conditional probability distribution $\hat{P}(T|Z,X) \sim \mathcal{N}(\mu_\omega(Z,X), \sigma_\omega(Z,X))$. Then, we optimize the following loss function for treatment regression:

$$\mathcal{L}_T = \int_{-\infty}^{+\infty} \left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}\left\{t_i \leq v\right\}\log(\Phi(v)) + \mathbf{1}\left\{t_i > v\right\}\log(1-\Phi(v))\right)\right) dv, \quad \Phi(v) = \int_{-\infty}^{v} \phi(t \mid z_i, x_i) dt \qquad (26)$$

[5] The conditional probability distribution is a mixture of gaussian distribution with multiple sub-networks $\{\mu_{\omega,k}, \sigma_{\omega,k}\}, k = 1, ..., K$, where $K$ denotes the number of latent gaussian distributions.

where $\Phi(\cdot)$ is cumulative distribution function, and $\mathbf{1}\{\cdot\}$ is an indicator function of the events $(t_i \leq v)$ or $(t_i > v)$.

**Besides**, to reduce computational complexity, we can set $\sigma_\psi = c$ as constant for low uncertainty models, and simplify the distribution estimation as a regression problem:

$$\mathcal{L}_T = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(t_i - \hat{t}_i^j\right)^2, \hat{t}_i^j \sim \hat{P}(t_i|z_i, x_i), \tag{27}$$

we sample $m$ (the larger the better) treatment $\{\hat{t}_i^j\}_{j=1,\ldots,m}$ for each unit $\{z_i, x_i\}$ to approximate the true treatment $t_i$. Empirically, the above objective (Eq. (27)) is sufficient to accurately estimate causal effects in continuous CB-IV framework.

**Confounder Balancing:** For continuous treatment $T$, we learn a ”balanced” representation (i.e., $C$) of the observed confounders $X$ as $C = f_\theta(X)$ via mutual information (MI) minimization constraints (Cheng et al., 2020): firstly, we use variational distribution $Q_\psi(\hat{T} \mid C) = \mathcal{N}(\mu_\psi(C), \sigma_\psi(C))$ parameterized by neural networks $\{\mu_\psi, \sigma_\psi\}$ to approximate the true conditional distribution $P(\hat{T} \mid C)$; then, we minimize the log-likelihood loss function of variational approximation $Q_\psi(\hat{T} \mid C)$ with $n$ samples to estimate MI:

$$\operatorname{disc}(\hat{T}, C) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\log Q_\psi\left(\hat{t}_i \mid c_i\right) - \log Q_\psi\left(\hat{t}_j \mid c_i\right)\right]. \tag{28}$$

where, $C = f_\theta(X)$. We adopt an alternating training strategy to iteratively optimize $Q_\psi(\hat{T} \mid C)$ and the network $C = f_{\theta(X)}$ to implement balanced representation in the Confounder Balancing.

**Outcome Regression:** Finally, we propose to regress the outcome with the estimated treatment $\hat{T} \sim P(T|Z, X)$ obtained in treatment regression module and the representation of confounders $C = f_\theta(X)$ obtained in confounder balancing module:

$$\mathcal{L}_Y = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \left(y_i - h_\xi(\hat{t}_i, f_\theta(x_i))\right)^2 \tag{29}$$

where $\hat{t}_i \sim \hat{P}(T|Z, X)$ and $f_\theta(x_i)$ are derived from treatment regression module and confounder balancing module, respectively.

**Optimization:** Like the optimization of the previous IV regression method, we also set two-stage optimization for our algorithm. In the first stage, we optimize the treatment regression networks $\{\mu_\psi, \sigma_\psi\}$ by minimizing the loss $\mathcal{L}_T$ ( Eq. (26) or Eq. (27) ). In the second stage, we adopt an alternating training strategy to iteratively optimize the variational distribution $Q_\psi(\hat{T} \mid C)$ and the outcome regression modules with representation $C = f_{\theta(X)}$, with the following loss:

$$\min_\psi \mathcal{L}_\psi \quad = \quad \operatorname{disc}(\hat{T}, f_\theta(X)), \tag{30}$$

$$\min_{\theta,\xi} \mathcal{L}_{\theta,\xi} \quad = \quad \mathcal{L}_Y + \alpha \operatorname{disc}(\hat{T}, f_\theta(X)), \tag{31}$$

where $\alpha$ is a trade-off hyper-parameter. We minimize $\mathcal{L}_\psi$ by using stochastic gradient descent to update the parameters of the variational distribution $Q_\psi(\hat{T} \mid C)$, and then, minimize $\mathcal{L}_{\theta,\xi}$ to implement confounder balance and estimate counterfactual outcome.

# E. Pseudo-Code and Hyper-parameters

## E.1. Pseudo-Code and Network Structures

We formulate the regression problems into optimization problems, and optimize them sequentially (Alternating training strategy is also an option). The optimization loss functions of the two regression networks are:

$$\min_\mu \mathcal{L}_T = -\frac{1}{n} \sum_{i=1}^{n} \left(t_i \log\left(\pi_\mu(z_i, x_i)\right) + (1 - t_i)\left(1 - \log\left(\pi_\mu(z_i, x_i)\right)\right)\right) \tag{32}$$

$$\min_{\theta,\xi^0,\xi^1} \mathcal{L}_Y + \alpha \mathcal{L}_C = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{\hat{t} \in \{0,1\}} h_{\xi^{\hat{t}}}(f_\theta(x_i)) \hat{P}(\hat{t} \mid z_i, x_i)\right)^2 + \alpha \operatorname{disc}(\hat{T}, f_\theta(X)) \tag{33}$$

*Table 5.* Network structures of CB-IV on Data-$m_Z$-$m_X$-$m_U$.

| Stage | Setting | Syn | IHDP | Twins |
|---|---|---|---|---|
| **Treatment Regression** | Loss | log | log | log |
| | Epoch | 3 | 3 | 3 |
| | Batchsize | 500 | 500 | 500 |
| | MLPLayers | [128,64] | [128,64] | [128,64] |
| | Activation | ReLU | ReLU | ReLU |
| | BatchNorm | True | True | True |
| | Learning_Rate | 0.05 | 0.05 | 0.05 |
| | Optimizer | SGD | SGD | SGD |
| **Outcome Regression** | Loss | MSE | MSE | log |
| | Epoch | 3000 | 100 | 200 |
| | Batchsize | 256 | 100 | 100 |
| | MLPLayers_R | [256]$*$3 | [200]$*$3 | [256]$*$3 |
| | MLPLayers_Y | [256]$*$5 | [100]$*$3 | [128]$*$5 |
| | Activation | ELU | ELU | ELU |
| | BatchNorm | False | False | False |
| | Learning_Rate | 0.0005 | 0.0005 | 0.0005 |
| | Optimizer | Adam | Adam | Adam |
| | $\alpha$ | 0.01/0.001 | 0.1 | 0.001/0.0001 |

where $\alpha$ is a trade-off hyper-parameter.

For the Treatment Regression, we use multi-layer perceptrons with ReLU activation function and BatchNorm as our logistic regression network $\pi_\mu$ and the network has two hidden layers with 128, 64 units, respectively. Then, We use stochastic gradient descent (SGD, (Duchi et al., 2011)) to train the network $\pi_\mu$ with a loss $\mathcal{L}_T$ for three epochs with a batch size of 500.

For the Outcome Regression and Confounder Balancing, we use Adam (Kingma & Ba, 2014) to train the three networks $f_\theta, h_{\xi^0}, h_{\xi^1}$ with loss $\mathcal{L}_Y + \alpha\mathcal{L}_C$ jointly. To prevent overfitting, we add a regularization term to regularize the prediction functions $h_{\xi^0}, h_{\xi^1}$ with a small $l_2$ weight decay.

Table 5 shows the details of the structure networks of CB-IV in different datasets. In the Treatment Regression Stage, the Loss would be an MSE-loss for continuous treatments and a log-loss for binary treatments, and the treatment network has multiple hidden layers with [MLPLayers] units. In the Treatment Regression Stage, the Loss would be an MSE-loss for continuous outcomes and a log-loss for binary outcomes. The representation network has multiple hidden layers with [MLPLayers_R] units, and the outcome network has multiple hidden layers with [MLPLayers_Y] units. Algorithm 1 shows the pseudo-code of our methods (CB-IV).

Hardware used: Ubuntu 16.04.5 LTS operating system with 2 * Intel Xeon E5-2678 v3 CPU, 384GB of RAM, and 4 * GeForce GTX 1080Ti GPU with 44GB of VRAM.

Software used: Python with TensorFlow 1.15.0, NumPy 1.17.4, and MatplotLib 3.1.1.

### E.2. Hyper-parameters Analysis on Data-$m_Z$-$m_X$-$m_U$

Given the multi-term objective function (Eq. (7)) in CB-IV, we study the confounder balance item (Eq. (5)) on the average treatment effect estimation of different datasets (Data-$m_Z$-$m_X$-$m_U$) by changing hyper-parameter $\alpha$ in the scope $\{0, 0.0001, 0.001, 0.01, 0.1, 1\}$. The result in Figure 3 demonstrates the confounder balance item is necessary for CB-IV. Combined with the two-head outcome functions, CB-IV indeed learn an effective independent representation and accurately estimate the average treatment effect.

## F. The Experiments about Different Variables Used in Different Stage

According to the preliminaries, we confirm that it is not sufficient to use instruments only in the first stage of the IV methods. In this section, we use **Syn(vars used in stage 1)(vars used in stage 2)** to represent that the regression variables we would

**Algorithm 1** CB-IV: Instrumental Variable Regression with Confounder Balancing

```
Input: Observational data $\mathbb{D} = \{z_i, x_i, t_i, y_i\}_{i=1}^{n}$; Maximum number of iterations $\mathcal{I}$
Output: $\hat{Y}_0 = h_{\xi^0}(f_\theta(X)), \hat{Y}_1 = h_{\xi^1}(f_\theta(X))$
Loss function: $\mathcal{L}_T$ and $\mathcal{L}_Y + \alpha\mathcal{L}_C$
Components: Logistic regression network $\pi_\mu(\cdot)$; Representation learning network $f_\theta(\cdot)$; Two-head outcome regression networks $h_{\xi^0}(\cdot)$ and $h_{\xi^1}(\cdot)$.
Treatment Regression Stage:
for itr $= 1$ to $\mathcal{I}$ do
    $\{z_i, x_i\}_{i=1}^{n} \to \pi_\mu(z_i, x_i) \to \hat{P}(t = 1 \mid z_i, x_i)$
    $\mathcal{L}_T = -\frac{1}{n}\sum_{i=1}^{n}\left(t_i \log\left(\pi_\mu(z_i, x_i)\right) + (1 - t_i)\left(1 - \log\left(\pi_\mu(z_i, x_i)\right)\right)\right)$
    update $\mu \leftarrow \mathrm{SGD}\{\mathcal{L}_T\}$
end for
Outcome Regression Stage:
for itr $= 1$ to $\mathcal{I}$ do
    $\{x_i\}_{i=1}^{n} \to C_i = f_\theta(x_i)$
    $\{z_i, x_i\}_{i=1}^{n} \to \pi_\mu(z_i, x_i) \to \hat{P}(t = 1 \mid z_i, x_i)$
    $\{f_\theta(x_i), t_i\}_{i=1}^{n} \to \mathrm{disc}(\hat{T}, f_\theta(X))$
    $\mathcal{L}_Y + \alpha\mathcal{L}_C = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{\hat{t}\in\{0,1\}} h_{\xi^{\hat{t}}}(f_\theta(x_i))\hat{P}(\hat{t} \mid z_i, x_i)\right)^2 + \alpha\,\mathrm{disc}(\hat{T}, f_\theta(X))$
    update $\theta, \xi^0, \xi^1 \leftarrow \mathrm{Adam}\{\mathcal{L}_Y + \alpha\mathcal{L}_C\}$
end for
```

use in the two stages of the instrumental variable method, respectively. Then we sample 10000 units from *Syn-2-4-4* to construct the datasets Syn(vars used in stage 1)(vars used in stage 2) perform 10 replications. For example, *Syn(Z)(X)* means that we perform logistic regression from the instruments $Z$ to the treatments $T$ in the first stage for all IV methods. We estimate the causal effect of the treatments $T$ on outcomes $Y$ using observed confounders $X$ in the second stage for all IV methods or in the outcome regression stage of representation methods.

We report the mean and the standard deviation on the bias of average treatment effect (ATE) estimation on different data settings in the Table 6. We find that almost all methods achieve the best results on *Syn(Z,X)(X)*, compared with *Syn(Z)(X)*,*Syn(X)(X)* and *Syn(Z,X)(Z,X)*, which is in line with our expectations. Comparing the results of *Syn(Z)(X)* and *Syn(Z,X)(X)*, all IV methods, including CB-IV, are no longer effective in the setting *Syn(Z)(X)*, DRCFR will achieve the best average treatment effect estimation. In addition, the results of DeepIV and DFIV methods are poor and almost unchanged on all data. The result confirms that these IV methods would be no longer effective, using only instrumental variables $Z$ or only observed confounding variables $X$ in the first stage.

In reality, we may not identify which variables we observed are instrumental variables $Z$ and which are confounders $X$. Fortunately, our proposed model is still valid in this case. The result of setting *Syn(Z,X)(Z,X)* shows CB-IV, using all observed variables $\{Z, X\}$ in stage 1 and learning a balanced representation of all observed variables $\{Z, X\}$ to implement causal effect estimation in stage 2, can still obtain a SOTA results. Moreover, the confounder balance methods (DirectRep,CFR and DRCFR) transiently balances the representation of instrumental variables $Z$, the performance will degrade. The traditional instrumental variable methods (DeepIV,OneSIV and DFIV) cannot identify causal effects in this scenario.

## G. Discussion on Confounder Balancing

### G.1. Confounder Bias and Confounder Balancing

Estimating the causal effect is crucial for explanatory analysis and decision-making across many domains (Li et al., 2022; Zhang et al., 2020; 2021). The gold standard approach for treatment effect estimation is to perform Randomized Controlled Trials (RCTs), where different treatments are randomly assigned to units. Unlike RCTs, the treatment $T$ in the observational studies is not randomly assigned; instead depends on confounders $X$. As introduced in Chapter 3.3 in Causality (Pearl, 2009b), this change could result in confounding bias: $\mathbb{P}(T|X) \neq \mathbb{P}(T)$. Specifically, in binary treatment cases, confounder $(X)$ may become imbalanced between different treatment arms $\hat{T}$, e.g., $P(X \mid \hat{T} = 1) \neq P(X \mid \hat{T} = 0)$, leading to confounding bias $\epsilon = \mathbb{E}[Y(\hat{T} = 1) - Y(\hat{T} = 0) \mid X] - \Big(\mathbb{E}[Y(\hat{T} = 1)) \mid \hat{T} = 1, X] - \mathbb{E}[Y(\hat{T} = 0) \mid \hat{T} = 0, X]\Big)$, where

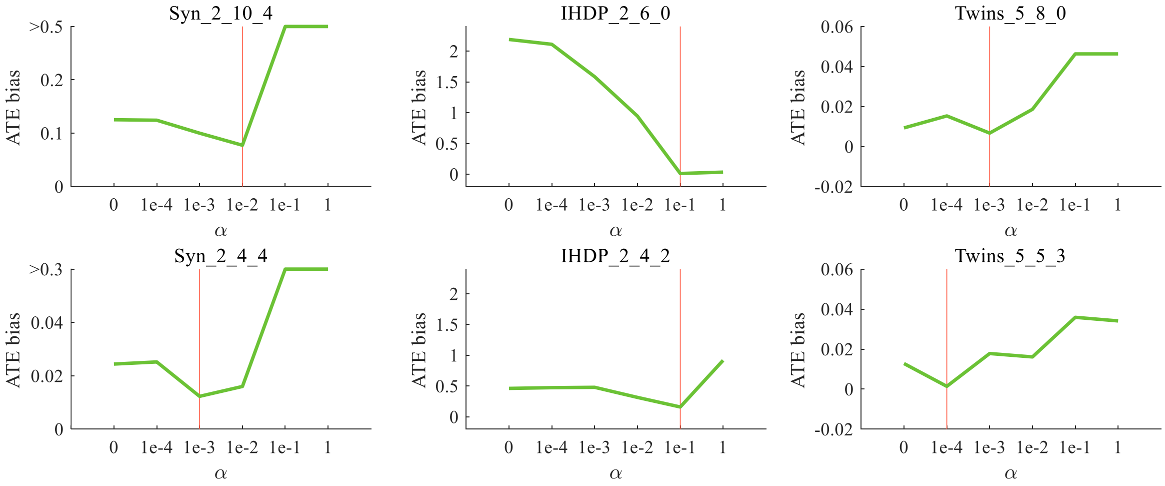

*Figure 3.* Hyper-parameter sensitivity analysis on Data-$m_Z$-$m_X$-$m_U$. The green lines show the ATE bias of the hyper-parameter $\alpha$ within the specified range $\{0, 0.0001, 0.001, 0.01, 0.1, 1\}$. The red line indicates the parameters chosen by CB-IV.

$Y(t)$ denotes the potential outcome.

**Definition G.1.** **Confounding bias** between the treatment and outcome can be defined as the bias of treatment effect estimation when the confounders exist (Pearl, 2009b).

For example, in the hospital scenario, most patients (have an injection) in the treated group have severe comorbidity, i.e., $\mathbb{P}(T = \text{injection}|X = \text{severe comorbidity}) > \mathbb{P}(T = \text{injection}|X = \text{mild comorbidity})$. If we directly regress $\mathbb{E}[Y|T, X] = h_\xi(T, X)$, then, the potential injection output estimation for patients with mild comorbidity will be biased towards the actual results of patients with severe comorbidity due to the confounding bias. Thus, **confounder balancing** means that we try to balance the distributions of confounders $X$ between different treatment arms $T$ to simulate the results of Randomized Controlled Trials (RCTs), i.e., $\mathbb{P}(T = 1|X) = \mathbb{P}(T = 0|X)$, equivalent to $\mathbb{P}(X|T = 1) = \mathbb{P}(X|T = 0)$.

In the current non-linear IV regression models, the observed confounders $X$ would affect both the estimated treatment $\hat{T}$ and the outcome $Y$ in the outcome regression in stage 2 as shown in Figure 1. Then, it would bring confounding bias between $\hat{T}$ and $Y$ if the outcome regression model is misspecified. To address the confounding bias from observable confounders, traditional confounder balance works, such as propensity score methods(Rosenbaum & Rubin, 1983; Rosenbaum, 1987; Li et al., 2016; 2020), re-weighting methods(Athey et al., 2018; He & Garcia, 2009), Doubly Robust (Funk et al., 2011) or backdoor criterion (Pearl, 2009a) to control the confounders' distributions. CFR (Johansson et al., 2016; Shalit et al., 2017) formulates the problem of confounder balance as a covariate shift problem and regards the treated group as the source domain and the control group as the target domain for domain adaptive balance in observational data. In this paper, we use ”balanced” representation learning to tackle the problem.

**Discussion on direct regression**: In the randomized controlled trial setting, two distributions of confounders in treated and control group are same, i.e., $\mathbb{P}(X|T = 0) = \mathbb{P}(X|T = 1) = \mathbb{P}(X)$. We can estimate the potential control and treated outcome well enough by directly implementing neural network regression from the treatments and confounders to the outcomes, i.e., $\mathbb{E}[Y|T, X] = h_\xi(T, X)$. However, in the observational study, estimating causal effects from observational data is different from supervised learning (Yuan et al., 2022). This is close to “learning from logged bandit feedback” (Strehl et al., 2010), with the distinction that we do not have access to the action generator model.

When we directly regress $\mathbb{E}[Y|T, X] = h_\xi(T, X)$, If if the regression model is misspecified, there will be two vital problems: (1) **Finite Samples**:The neural network, without any regularization, may be overfitted on the limited training data. In binary treatment case , such as the hospital scenario, most patients (have an injection) in the treated group have severe comorbidity, i.e., $\mathbb{P}(X = \text{severe comorbidity}|T = \text{injection}) > \mathbb{P}(X = \text{mild comorbidity}|T = \text{injection})$ . Then, the potential injection output estimation for patients with mild comorbidity will be biased towards the actual results of same patients with severe comorbidity due to the confounding bias. (2) **Treatment Indicator might get lost**: $\hat{T}$ is a mediator in chains $X \to \hat{T} \to Y$, and **the information of treatment might got lost** in high-dimension confounders (Johansson et al., 2016; Shalit et al., 2017), resulting in the consistency of the predicted potential outcomes from different treatments for the specified $X$, i.e.,

*Table 6.* The results of ATE estimation, including bias (mean(std)), in binary treatment cases on Synthetic data with different settings (Syn(vars used in stage 1)(vars used in stage 2)).

| | | **Within-Sample** | | |
|---|---|---|---|---|
| **Method** | **Syn(Z)(X)** | **Syn(X)(X)** | **Syn(Z,X)(Z,X)** | **Syn(Z,X)(X)** |
| **DeepIV-LOG** | 1.055(0.006) | 1.054(0.007) | 1.059(0.009) | 1.057(0.008) |
| **DeepIV-GMM** | 0.862(0.016) | 0.992(0.007) | 0.961(0.006) | 0.874(0.019) |
| **KernelIV** | 0.964(0.070) | 0.865(0.174) | 0.890(0.157) | 0.457(0.054) |
| **DualIV** | 0.658(0.561) | 1.611(0.495) | 1.763(0.042) | 1.423(0.076) |
| **OneSIV** | 1.048(0.030) | 1.176(0.046) | 1.053(0.045) | 0.661(0.096) |
| **DFIV** | 1.003(0.010) | 0.894(0.004) | 0.838(0.007) | 0.860(0.007) |
| **DFL** | 0.842(0.002) | 0.843(0.002) | 0.842(0.002) | 0.851(0.002) |
| **DirectRep** | 0.163(0.008) | 0.163(0.008) | 0.178(0.022) | 0.163(0.008) |
| **CFR** | 0.158(0.015) | 0.158(0.015) | 0.177(0.023) | 0.158(0.015) |
| **DRCFR** | **0.136(0.034)** | **0.136(0.034)** | 0.141(0.054) | 0.136(0.034) |
| **CB-IV** | 0.495(0.263) | 0.529(0.100) | **0.115(0.072)** | **0.016(0.047)** |

| | | **Out-of-Sample** | | |
|---|---|---|---|---|
| **Method** | **Syn(Z)(X)** | **Syn(X)(X)** | **Syn(Z,X)(Z,X)** | **Syn(Z,X)(X)** |
| **DeepIV-LOG** | 1.055(0.005) | 1.055(0.007) | 1.059(0.010) | 1.057(0.008) |
| **DeepIV-GMM** | 0.862(0.016) | 0.992(0.007) | 0.961(0.006) | 0.874(0.019) |
| **KernelIV** | 0.963(0.070) | 0.865(0.177) | 0.916(0.157) | 0.458(0.052) |
| **DualIV** | 0.800(0.307) | 1.606(0.501) | 1.760(0.037) | 1.467(0.053) |
| **OneSIV** | 1.048(0.030) | 1.176(0.045) | 1.053(0.045) | 0.661(0.095) |
| **DFIV** | 1.003(0.009) | 0.894(0.004) | 0.838(0.006) | 0.860(0.007) |
| **DFL** | 0.842(0.002) | 0.843(0.002) | 0.842(0.002) | 0.851(0.002) |
| **DirectRep** | 0.164(0.009) | 0.164(0.009) | 0.179(0.019) | 0.164(0.009) |
| **CFR** | 0.159(0.018) | 0.159(0.018) | 0.178(0.023) | 0.159(0.018) |
| **DRCFR** | **0.137(0.035)** | **0.137(0.035)** | 0.142(0.052) | 0.137(0.035) |
| **CB-IV** | 0.493(0.261) | 0.528(0.099) | **0.114(0.071)** | **0.017(0.046)** |

$h_\xi(0, X_{T=t}) = h_\xi(1, X_{T=t}) = h_\xi(X_{T=t})$, $X_{T=t}$ denotes variables from the group $T = t$.

In finite samples, confounder balance is a important regularization on the outcome regression model. Converting $\mathbb{P}(X|T = 1) > \mathbb{P}(X|T = 0)$ to $\mathbb{P}(f_\theta(X)|T = 1) = \mathbb{P}(f_\theta(X)|T = 0) = \mathbb{P}(f_\theta(X))$ via balancing the distributions of confounders $X$ between different treatment arms $T$, we can enforce the representation distribution of training samples to approximate that of the population and keep $T$ not replaced by $X$ in the outcome regression stage. When we balance the representations, although the representations $C = f_\theta(X)$will lose information predictive of $\hat{T}$, we will emphasize the information of $\hat{T}$. Even under the ideal condition, we expect that the discarded information in $X$ can be can reconstructed by representation $C$ and $T$, it's a trade-off in balanced learning representations. Besides, we use the "balanced" representation to bound the expected treatment effect estimation error (Shalit et al., 2017): $\epsilon(h, \Phi) \leq 2\left(\epsilon_F^{t=0}(h, \Phi) + \epsilon_F^{t=1}(h, \Phi) + B_\Phi IPM_{\mathrm{G}}\left(p_\Phi^{t=1}, p_\Phi^{t=0}\right) - 2\sigma_Y^2\right)$. "Balanced" representation means that the gain is from decreasing the bias of the population, including the bias of counterfactual estimation, at the price of a small increase in the estimation bias of common samples in data.

"Balanced" representation (Johansson et al., 2016; Shalit et al., 2017) has good performance and can capture complex relationships among treatments, observed confounders, and outcomes, but it requires the unconfoundedness assumption. For example, physical fitness (i.e., unobserved confounders $U$) may not be recorded in the historical data. The causal effects of the treatments on outcomes are not identifiable from data with unmeasured confounders. To address this challenge, the patients' income, an instrumental variable (IV) $Z$ that only affects the treatments and does not affect the outcomes directly, can be used to eliminate the unmeasured confounding bias (Pearl et al., 2000; Wright, 1928; Heckman, 2008; Stock & Trebbi, 2003).

### G.2. About the Wasserstein Distance

For representation balancing, CFR (Johansson et al., 2016; Shalit et al., 2017) and DR-CFR (Hassanpour & Greiner, 2019b) adopt Maximum Mean Discrepancy (MMD) and Wasserstein distance (Wass) to calculate the dissimilarity of distributions from different treatment arms and fit a balanced representation by minimizing the discrepancy. For the sake of fairness, we uniformly use Wass distance as the discrepancy metrics for CFR, DR-CFR, and CB-IV in the experimental comparison. Wass distances ($W_p(\mu, \nu) \overset{\text{def}}{=} \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega^2} D(x,y)^p d\pi(x,y)\right)^{1/p}, p \in [1, \infty)$ and probability measures $\mu, \nu \in$ Borel probability measures$P(\Omega)$) have many favorable properties, documented both in theory (Villani, 2009; Cuturi & Doucet, 2014) and practice (Pele & Werman, 2009). Besides, Wass distance have consistent estimators which can be efficiently computed in the finite sample case (Shalit et al., 2017; Sriperumbudur et al., 2012) and Wass distance is a common measure in deep learning: many algorithm breakthroughs (Arjovsky et al., 2017; Cuturi & Doucet, 2014) benefit from it. However, there is no known way or a simple method for some function families to compute the integral probability metric or its gradients efficiently. Therefore, this paper adopts the Wass distance in binary treatment cases for fairness and expects better performance. As for continuous treatment cases, we learn a ”balanced” representation via mutual information minimization constraints CLUB (Cheng et al., 2020). The experiments and the theory (Shalit et al., 2017) both prove that a ”balanced” representation facilitates tighter expected error bounds in the enormous sample size.

In binary treatment cases, $\mathbb{P}(C|T=0) = \mathbb{P}(C|T=1)$ if and only if $IPM = Wass(C_{T=0}, C_{T=1}) = 0$. Obviously, in binary case, $IPM = 0$ means that the distributions of representation $C$ are the same in the treated group and the control group, i.e., $\mathbb{P}(C|T=0) = \mathbb{P}(C|T=1) = \mathbb{P}(C)$. The learned representation $C$ is independent of $T$. In continuous treatment cases, we can regard the minimization of mutual information between representation $C$ and treatment $T$ as $C \perp T$.

### G.3. Error Bounds with Representation Balancing

Shalit et al. (2017) gives a novel, and intuitive generalization-error bound showing that the expected treatment effect estimation error is bounded by the standard generalization-error and the distance between the treated and control distributions induced by the representation:

$$\epsilon(h,\theta) \leq 2\left(\epsilon_F^{t=0}(h,\theta) + \epsilon_F^{t=1}(h,\theta) + B_\theta IPM_{\mathrm{G}}\left(p_\theta^{t=1}, p_\theta^{t=0}\right) - 2\sigma_Y^2\right) \tag{34}$$

where $\epsilon_F^{T=t}(h,\theta) = \int_{\mathcal{X}} \ell_2(y, h(T=t, f_\theta(x)))p^{T=t}(x)dx$ for $t \in \{0,1\}$; $p^{T=t}(x)$ denotes the PDF of $x$ given $T=t$; $p_\theta^{T=t} = \{f_\theta(x_i)\}_{i:t_i=t}$; $B_\theta$ is a constant; $\sigma_Y^2$ is the expected variance of $Y$.

The instrumental variable deals with unobserved confounders, as shown in Figure 1(b), variables $X$, common causes of the conditional treatments $\hat{T}$ and outcomes $Y$, are confounders and not deconfounded in stage 2 of these nonlinear IV regression methods (Example A.1 in Section A and corresponding Experiments in Section F in Appendix). Based on the two-stage regression of IV methods, we propose to use confounder balance techniques to reduce the error in the outcome regression stage. Consequently, we use $\mathcal{L}_2$ (Eq. 7) as the loss function in the outcome regression stage:

$$\min_{\theta,\xi^0,\xi^1} \mathcal{L}_Y + \alpha\mathcal{L}_C = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{\hat{t}\in\{0,1\}} h_{\xi^{\hat{t}}}(f_\theta(x_i))\hat{P}(\hat{t} \mid z_i, x_i)\right)^2 + \alpha \operatorname{disc}(\hat{T}, f_\theta(X)) \tag{35}$$

In mathematical, the optimization goal $\mathcal{L}_Y$ and $\mathcal{L}_C$ are consistent with error bound $2\left(\epsilon_F^{t=0}(h,\theta) + \epsilon_F^{t=1}(h,\theta) + B_\theta IPM_{\mathrm{G}}\left(p_\theta^{t=1}, p_\theta^{t=0}\right) - 2\sigma_Y^2\right)$. If we directly regress $\mathbb{E}[Y|T,X] = h_\xi(T,X)$, non-parametric models without prior knowledge may have poor prediction performance for samples that rarely appear in the data (overfiting). Thus, confounder balance is a great regularization on the outcome regression model. We bound the error $\epsilon(h,\theta)$ by minimizing $\epsilon_F^{t=0}(h,\theta) + \epsilon_F^{t=1}(h,\theta)$ and $IPM_{\mathrm{G}}\left(p_\theta^{t=1}, p_\theta^{t=0}\right)$ simultaneously. Combining with IV methods and confound balance methods, we eliminate the confounding bias from observed confounders and unmeasured confounders.