TREATMENT EFFECT ESTIMATION WITH CONFOUNDER BALANCED INSTRUMENTAL VARIABLE REGRESSION

Anonymous authors
Paper under double-blind review

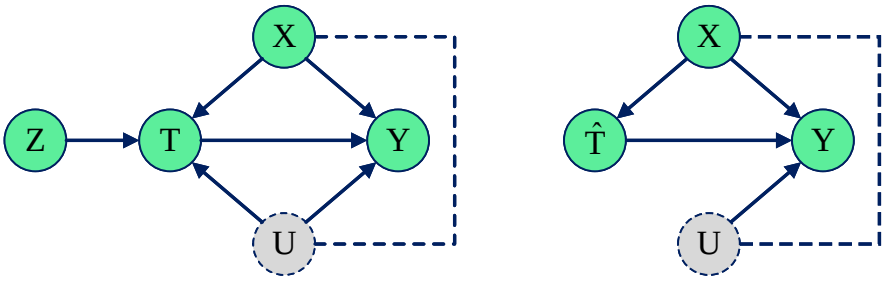
ABSTRACT

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 they need to assume the additive separability of noise and are limited to the linear setting. Recently, many nonlinear IV regression variants were proposed by regressing the treatment with IVs and confounders in the first stage, leading to confounding bias between the predicted treatment and outcome in the second stage. In this paper, we propose a Confounder Balanced IV Regression (CB-IV) algorithm to jointly remove the bias from the unmeasured confounders with IV regression and achieve better bias-variance trade-off in imbalanced treatment distributions due to the observed confounders by balancing for treatment effect estimation. Specifically, CB-IV algorithm consists of three main modules: (1) treatment regression: regressing the treatment with IVs and confounders like previous nonlinear IV methods for removing the confounding from unmeasured confounders; (2) confounder balancing: learning a balanced representation of confounders to eliminate the bias induced by the observed confounders (3) outcome regression: regressing the outcome with the predicted treatment and the balanced confounders representation for treatment effect estimation. To the best of our knowledge, this is the first work to combine confounder balancing in IV regression for treatment effect estimation. Moreover, we theoretically prove that CB-IV algorithm is also effective under the multiplicative assumption rather than the additive separability assumption. Extensive experiments demonstrate that CB-IV algorithm outperforms the state-of-the-art methods, including IV regression and confounder balancing methods, for treatment effect estimation.

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 unconfounderness 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 unconfounderness 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.

A classical method to address the bias induced by unmeasured confounder is IV regression methods (Pearl et al., 2000; Wright, 1928a; Heckman, 2008; Stock & Trebbi, 2003). As shown in Figure 1(a), let T denotes the treatment, Y refers 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 influence Y via T. In IV regression, two-stage least squares (2SLS) regression (Pearl et al., 2000; Angrist & Imbens, 1995; Angrist & Krueger, 2001) is a classical statistical method with the following two stages: In stage 1, 2SLS performs linear regression from the instruments Z to the treatments T; then in stage 2, it performs linear regression from the conditional expectation of the treatments $\mathbb{E}[T \mid Z]$ (obtained from the stage 1) to the outcomes Y. However, 2SLS and other variants of IV regression methods (Stock et al., 2002; Baum et al., 2003; Carrasco et al., 2007; Buhlmann et al., 2014), require strong assumptions, either linearity or



(a) Causal Structural with Unmeasured Confounders (b) Stage 2 in the nonlinear IV methods

Figure 1: (a) Causal structural with unmeasured confounders. (b) Causal structure of stage 2 regression in the nonlinear IV methods. The observed confounders would be affected both the predicted treatment $\hat{T}$ and outcome Y, leading to confounding bias in stage 2 regression. In these figures, green nodes denote observable variables, and gray nodes mean unmeasured variables. The arrows with solid line point from the cause variables to the effect variables; The dashed lines without arrow mean that the causal direction between the two variables is unknown.

additive separability of instruments Z, confounders X and noise (i.e., unmeasured confounders) U. Moreover, in nonlinear scenarios, these methods cannot effectively extract instruments information in the first stage, and the conditional expectation $\mathbb{E}[T|Z] = \mathbb{E}[f(Z,X)|Z]$ may be a constant 0 or weak association with Y (see Theory in Section A and Experiment in Section E in Appendix).
To address the above problems, recent nonlinear IV regression variants (Hartford et al., 2017; Xu et al., 2020; Singh et al., 2019; Muandet et al., 2019) learn a joint mapping from the instruments Z and observed confounders X to the conditional distribution of the treatment T in stage 1, i.e., $P(T|Z,X) = f(Z,X) + \mathbb{E}[U|X]$. Then, these methods resample the predicted treatment $\hat{T}$ from the conditional distribution $P(T|Z,X)$ obtained in stage 1 and perform nonlinear regression from the resampled treatment $\hat{T}$ and confounders X to the outcomes Y in stage 2, i.e., $\mathbb{E}[Y|Z,X] = \mathbb{E}[h(T,X)|Z,X] = \mathbb{E}[h(\hat{T},X)] = \mathbb{E}[g(\hat{T},X)] + \mathbb{E}[U|X]$, which only holds when the noise U is additive (Bareinboim & Pearl, 2012). From the processes of these methods, we know that the observed variables X would affect the predicted treatment $\hat{T}$ in stage 1, and also influences the outcome Y, therefore, X would bring confounding bias between the predicted treatment $\hat{T}$ and the outcome Y for the regression in stage 2 as shown in figure 1(b), leading to poor performance of these methods. Fortunately, the unobserved confounders U will no longer confound the causal relationship between $\hat{T}$ and Y in stage 2 (see figure 1(b)), and we only need to analyze and adjust the observed confounders X.

In this paper, we propose a Confounder Balanced IV Regression (CB-IV) algorithm [1] to further remove the confounding bias from the observed confounders by balancing in IV regression for treatment effect estimation. Specifically, CB-IV algorithm contains the following three main components: (1) treatment regression: given Z and X, identify conditional probability distribution of the treatment variable T (i.e., $\hat{T} \sim P(T|Z,X) = f_1(Z,X) + \mathbb{E}[f_2(X,U)|X]$) for removing the confounding from unmeasured confounders, where we relax the assumption of additive on noise U; (2) confounder balancing: learn a balanced representation of observed confounders $C = f_\theta(X)$, which is independent with the predicted treatment $\hat{T} \sim P(T|Z,X)$ to reduce the confounding from the observed variables as shown in figure 1(b); and (3) outcome regression: regressing the outcome Y on the predicted treatment $\hat{T}$ and representation of confounders C (i.e., $\mathbb{E}(Y|Z,X) = \mathbb{E}[h(\hat{T},X)] = \mathbb{E}[g_1(\hat{T},X)] + \mathbb{E}[g_2(\hat{T})g_3(U)|C] = \mathbb{E}[g_1(\hat{T},X)] + \mathbb{E}[g_2(\hat{T})]\mathbb{E}[g_3(U)|C]$, which only holds when $C \perp g_2(\hat{T})$) for counterfactual inference and treatment effect estimation. Based on this, we relax the additive noise assumption. The main contributions in this paper are as follows:

- We study the problem of treatment effect estimation from observational data in the presence of unmeasured confounders, and we find that previous IV-based methods are either limited to the linear setting or would suffer from the bias from the observed confounders.

[1] Code: https://www.dropbox.com/sh/zwhp4bogdlhuqtj/AADgFcCLi-FfzRo7DQVTVFV1a?dl=0

- We propose Confounder Balanced IV regression (CB-IV) method to jointly remove the bias from unmeasured confounders by IV regression and observed confounders by balancing. Moreover, with confounder balancing in IV regression, we can relax the additive separability assumption in IV-based methods.
- Extensive experiments on both synthetic and real-world datasets demonstrate the effectiveness of the proposed algorithm. Under the multiplicative assumption (defined in Eq. (4)), CB-IV algorithm works well without additive separability assumption hold.

2 Related Works

2.1 Causal Representation Learning for confounder balance

Inspired by traditional confounder balance works, such as propensity score methods(Rosenbaum & Rubin, 1983; Rosenbaum, 1987; Li et al., 2016; 2020), re-weighting methods(Zubizarreta, 2015; Athey et al., 2018; He & Garcia, 2009), Doubly Robust (Funk et al., 2011) and backdoor criterion (Pearl, 2009), 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 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, this is a trade-off, i.e., $\mathbb{E}[Y|X,T] = g(C,T), C \perp T, C = f_\theta(X)$. CFR-ISW (Hassanpour & Greiner, 2019a) learns the representation C with a context-aware importance sampling weight. SITE (Yao et al., 2018) preserves local similarity and balances the distributions of the representation C simultaneously. DR-CFR (Hassanpour & Greiner, 2019b) and DeR-CFR (Wu et al., 2020) propose a disentanglement framework to identify the representation of confounders from all observed variables. CEVAE (Louizos et al., 2017) and GANITE (Yoon et al., 2018) use deep generative models to estimate the joint distribution for causal inference. More discussion on confounder balance is given in Section G in Appendix.

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 unmeasured confounding bias.

2.2 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; Angrist & Imbens, 1995; Angrist & Krueger, 2001) 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., 2019) implement 2-stage regression via mapping X to a reproducing kernel Hilbert space (RKHS) and performing kernel ridge regression. DFIV (Xu et al., 2020) adopts deep neural nets to replace the kernel counterparts. Based on the optimally weighted Generalized Method of Moments (GMM), AGMM (Lewis & Syrgkanis, 2018) and DeepGMM(Bennett et al., 2019) construct a structural function via minimizing the loss of the sample averages of the moment conditions. Given Z and X, 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), 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 (See Proof 1(b) for details). Based on the two-stage regression of IV methods, we propose to use the above confounder balance techniques to adjust the observed confounder and reduce the variance

in stage 2. This is the first provably efficient algorithm that combines the IV method with the confounder balance technique using deep representation learning to the best of our knowledge.

3 METHODOLOGY

3.1 PROBLEM SETTING AND PRELIMINARIES

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 $t_i \in \mathcal{T}$ where $\mathcal{T} \subset \mathbb{R}$, a outcome variable $y_i \in \mathcal{Y}$ where $\mathcal{Y} \subset \mathbb{R}$, instrumental variables $z_i \in \mathcal{Z}$ where $\mathcal{Z} \subset \mathbb{R}^{m_Z}$, and confounders $x_i \in \mathcal{X}$ where $\mathcal{X} \subset \mathbb{R}^{m_X}$. Besides, there are some confounders $u_i \in \mathcal{U}, \mathcal{U} \subset \mathbb{R}^{m_U}$, that simultaneously affect both t_i and y_i, 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 $\mathcal{X}$, instrumental variables $\mathcal{Z}$ and unobserved confounders $\mathcal{U}$. The causal relationship can be represented with the following model (Figure 1(a)):

$$\{Z, X, U\} \to T; \{T, X, U\} \to Y; Z \perp U, X; X \not\perp U \tag{1}$$

Definition 1 ***The average treatment effect*** *ATE is defined as:*

$$ATE = \mathbb{E}[Y \mid do(T = 1), X] - \mathbb{E}[Y \mid do(T = 0), X] \tag{2}$$

where the $do(\cdot)$ operator indicates that we have intervened to data.

Definition 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 X / U, i.e., $Z \perp X, U$*

Besides, homogeneity and monotonicity assumptions (the structural equation model) in causal effect are often used in the analysis of instrumental variables (Hernán & Robins, 2010; Wright, 1928b; Goldberger, 1972; Wooldridge, 2010).

To precisely estimate the treatment effect/causal relationship, most of the previous IV methods (Hartford et al., 2017; Xu et al., 2020; Singh et al., 2019; Muandet et al., 2019) require the additive noise assumption (i.e., unmeasured noise gets added to the intended results $\{T, Y\}$) and model the causal relationship as follows:

$$T = f(Z, X) + U, Y = g(T, X) + U, Z \perp U, X, \mathbb{E}[U] = 0 \tag{3}$$

where $f(\cdot)$ and $g(\cdot)$ are continuous structure functions, and U is an additive noise term.

In this paper, we model the causal relationship more general and relax the additive separability assumption to the multiplicative assumption, as follows:

$$T = f_1(Z, X) + f_2(Z) f_3(X, U), Y = g_1(T, X) + g_2(T) g_3(U) + g_4(X, U), Z \perp U, X \tag{4}$$

where $f_{1\cdots3}(\cdot)$ and $g_{1\cdots4}(\cdot)$ are continuous functions. In the structural function of Y, $g_2(T)g_3(U)$ denotes the multiplicative terms of U with T (e.g., $U^2T - UT + U$), and we define it as **the multiplicative assumption**. The same principle can be applied to the structural function of T. 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 case can be modeled similarly (Section C).

Definition 3 ***The Latent Outcome Function*** *$h(T, X)$ can be defined under the multiplicative assumption (4), as follows:*

$$\mathbb{E}[Y \mid do(T), X] = \mathbb{E}[h(T, X)] = \mathbb{E}\left[g_1(T, X) + g_2(T)\mathbb{E}[g_3(U) \mid X] + \mathbb{E}[g_4(X, U) \mid X]\right] \tag{5}$$

3.2 THEORETICAL ANALYSIS AND DISCUSSION

Theorem 1 *(Identification of treatment effects). If the learned representation of observed confounders $C = f_\theta(X)$ is independent with the predicted treatment $\hat{T}$, then the latent outcome function $h(T, X)$ can be identified with instrumental variables Z and representation C:*

$$h(T, X) = g_1(T, X) + g_2(T)\mathbb{E}[g_3(U)|C] + \mathbb{E}[g_4(X, U)|C], C = f_\theta(X) \tag{6}$$

where, $\mathbb{E}[g_3(U)|C]$ and $\mathbb{E}[g_4(X,U)|C]$ are constant for the specified X. The proof is given in Section B in Appendix.

Then, the corresponding Average Treatment Effect (**ATE**) estimation can be written as:

$$\begin{aligned} ATE &= \mathbb{E}[h(T=1,X) - h(T=0,X)] && (7) \\ &= \mathbb{E}\left[g_1(1,X) - g_1(0,X)\right] + \mathbb{E}\left[g_2(1) - g_2(0)\right]\mathbb{E}[g_3(U)|C] && (8) \end{aligned}$$

Recent IV methods (Hartford et al., 2017; Newey & Powell, 2003) regress a conditional treatment distribution $\hat{P}(T \mid Z, X)$ using $\{Z, X\}$ in the treatment regression stage, then learn the latent outcome function $h_\xi(T, X)$ from $\{T, X\}$ to Y directly:

$$\mathbb{E}[Y \mid Z, X] = \int h_\xi(T, X) d\hat{P}(T \mid Z, X) \tag{9}$$

Obviously, in the complicated setting (Eq. (4)), these methods do not meet the identification conditions of Theorem 1 and would be fooled by confounders X. Because X cause $\hat{P}(T \mid Z, X)$ in the treatment regression stage, X would be related to $\hat{T} \sim \hat{P}(T \mid Z, X)$ and Y.

Inspired by confounder balance works (Section 2.1), our algorithm (CB-IV) learn a balanced representation $C = f_\theta(X)$ independent of the predicted treatment $\hat{T} \sim \hat{P}(T \mid Z, X)$ and estimate treatment effects simultaneously in the outcome regression stage. Without loss of generality, we take the binary treatment case as an example to detail our algorithm.

$$\mathbb{E}[Y \mid Z, X] = \sum_{t \in \{0,1\}} h_\xi(T = t, C)\hat{P}(T = t \mid Z, X), C \perp T \mid \hat{P}(T \mid Z, X), C = f_\theta(X) \tag{10}$$

Then, we transform the problem into a optimization problem to minimize $MSE(Y - \mathbb{E}[Y \mid Z, X])$, which can be estimated by the train data $\mathbb{D} = \{z_i, x_i, t_i, y_i\}_{i=1}^n$:

$$\min_{h_\xi \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \left(y_i - \sum_{t \in \{0,1\}} h_\xi(t, f_\theta(x_i))\hat{P}(t \mid z_i, x_i) \right)^2 \tag{11}$$

where $\mathcal{H}$ is a function space of h_ξ, and $f_\theta(X)$ is the learned representation of confounders (i.e., C). Thus, the ATE can be estimated by $\hat{ATE} = \mathbb{E}[h_\xi(T = 1, f_\theta(X)) - h_\xi(T = 0, f_\theta(X))]$.

3.3 Algorithm and Optimization

IV regression is the classical method for addressing the unmeasured confounders, but recent nonlinear IV-based methods suffer the bias from the observed confounders as shown in figure 1(b), leading to poor performance 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. Without loss of generality, we take the binary treatment case as an example to detail three main components in the proposed CB-IV algorithm:

(1) Treatment Regression. In this part, we propose to regress treatment T with IVs Z and observed confounders X directly because Z and X are independent. Specifically, we estimate the conditional probability distribution of the treatments $\hat{P}(T|Z, X)$ with a logistic regression network $\pi_\mu(z_i, x_i)$ for each unit i:

$$\min_\mu \mathcal{L}_1 = -\frac{1}{n} \sum_{i=1}^n \left(t_i \log\left(\pi_\mu(z_i, x_i)\right) + (1 - t_i)\left(1 - \log\left(\pi_\mu(z_i, x_i)\right)\right)\right) \tag{12}$$

where $\pi_\mu(z_i, x_i) = \hat{P}(t = 1|z_i, x_i)$, μ is the learnable parameter of π.

(2) Confounder Balancing: In this component, we aim to remove the confounding bias induced by X as shown in figure 1(b). For binary treatment, Sriperumbudur et al. (2009); Johansson et al. (2016); Shalit et al. (2017) proposed integral probability metrics (IPMs) to minimize the discrepancy of distributions from different treatment arms. As for continuous treatment, Yuan et al. (2021); Cheng et al. (2020) adopt mutual information to control the representation learning. In this paper,

we propose to learn a representation of confounders (i.e., $C = f_\theta(X)$), and adopt the Wasserstein distance (Cuturi & Doucet, 2014) to measure the discrepancy of distributions to achieve $C \perp \hat{T}$:

$$\min_{\theta} \text{disc}(\hat{t}, f_\theta(x_i)) = \text{Wass}(\{f_\theta(x_i)\hat{P}(t_i \mid z_i, x_i)\}_{i:t_i=0}, \{f_\theta(x_i)\hat{P}(t_i \mid z_i, x_i)\}_{i:t_i=1}) \tag{13}$$

where $\{f_\theta(x_i)\hat{P}(t_i \mid z_i, x_i)\}_{i:t_i=k}, k \in \{0, 1\}$ 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 stronger version that would force $f_\theta(x_i)$ and original T to be independent directly:

$$\min_{\theta} \text{disc}(\hat{t}, f_\theta(x_i)) = \text{Wass}(\{f_\theta(x_i)\}_{i:t_i=0}, \{f_\theta(x_i)\}_{i:t_i=1}) \tag{14}$$

More discussion on confounder balance and Wass distance is given in Section G in Appendix.

(3) Outcome Regression. Finally, we propose to regress the outcome with the predicted 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). This phenomenon also exists in the IV based methods. We propose to learn $h_{\xi^0}(f_\theta(X))$ and $h_{\xi^1}(f_\theta(X))$ as two different head to estimate the treated outcomes $Y(T = 1, X)$ and control outcomes $Y(T = 0, X)$, which will also help to learn independent representation $C = f_\theta(X)$ and reduce the confounding bias:

$$\min_{\xi^0, \xi^1} \quad \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \sum_{\hat{t}\in\{0,1\}} h_{\xi^{\hat{t}}}(f_\theta(x_i))\hat{P}(\hat{t} \mid z_i, x_i)\right)^2 \tag{15}$$

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

Optimization: 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:

$$\min_{\mu} \mathcal{L}_1 = -\frac{1}{n}\sum_{i=1}^{n} \left(t_i \log\left(\pi_\mu(z_i, x_i)\right) + (1 - t_i)\left(1 - \log\left(\pi_\mu(z_i, x_i)\right)\right)\right) \tag{16}$$

$$\min_{\theta, \xi^0, \xi^1} \mathcal{L}_2 = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \sum_{\hat{t}\in\{0,1\}} h_{\xi^{\hat{t}}}(f_\theta(x_i))\hat{P}(\hat{t} \mid z_i, x_i)\right)^2 + \alpha\, \text{disc}(\hat{t}, f_\theta(x_i)) \tag{17}$$

where α is a trade-off hyper-parameter. For the Treatment Regression, we use stochastic gradient descent (SGD, (Duchi et al., 2011)) to train the logistic regression network π_μ with loss $\mathcal{L}_1$. 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 $\mathcal{L}_2$ jointly. To prevent overfitting, we add a regularization term to regularize the prediction functions h_{ξ^0}, h_{ξ^1} with a small l_2 weight decay. Then, the average treatment effect can be estimated by $\hat{ATE} = \mathbb{E}[h_{\xi^1}(f_\theta(X)) - h_{\xi^0}(f_\theta(X))]$.
The details of pseudo-code (see Algorithm 1) and the network structures (see Table 3) of our algorithm are provided in Section D.1 in Appendix. Besides, the discussion of hyper-parameters α (see Figure 3) is detailed in Section D.2 in Appendix.

4 Experiments

4.1 Baselines

We compare the proposed algorithm (**CB-IV**) with two group 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**[2](Muandet et al., 2019): KernelIV and DualIV implement 2-stage regression with different dictionaries of basis functions from reproducing kernel Hibert spaces (RKHS); (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., 2020): 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 with representation*: (1) **DFL** (Xu et al., 2020): DFL, an ablation experiment of DFIV, performs the nonlinear outcome regression directly without using instrumental variables; (2) **DirectRep** and **CFR** (Johansson et al., 2016;

[2]The codes of KernelIV and DualIV are available at https://github.com/krikamol/DualIV-NeurIPS2020.

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 versions of **CB-IV** without confounder banalcing, and **DirectRep** and **CFR** are the ablation versions of **CB-IV** without IV regression. 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. The continuous treatment experiments are given in Section F in Appendix.

4.2 Experiments on Synthetic Datasets

4.2.1 Dataset.

Similar to (Hassanpour & Greiner, 2019b), we generate the synthetic datasets as follows:

- The latent variables $\{Z, X, U\}$:

$$Z_1, \cdots Z_{m_Z} \sim \mathcal{N}(0, \boldsymbol{I}_{m_Z}), X_1, \cdots X_{m_X}, U_1, \cdots U_{m_U} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{m_X+m_U}) \tag{18}$$

  where m_Z, m_X and m_U are the dimensions of instruments, observed confounders and unobserved confounders respectively. $\boldsymbol{I}_{m_Z}$ denotes m_Z degree identity matrix, $\boldsymbol{\Sigma}_{m_X+m_U} = \boldsymbol{I}_{m_X+m_U} * 0.95 + \mathbb{1}_{m_X+m_U} * 0.05$ 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 variables T:

$$P(T \mid Z, X) = \frac{1}{1+\exp\left(-(\sum_{i=1}^{m_Z} Z_i X_i + \sum_{i=1}^{m_X} X_i) + \sum_{i=1}^{m_U} U_i)\right)}, T \sim Bernoulli(P(T \mid Z, X)), m_X > m_Z \tag{19}$$

  where $Bernoulli(P(T \mid Z, X))$ is the true logging policy of the treatments T.

- The outcome variabls Y:

$$Y(T, X, U) = \frac{T}{m_X+m_U}(\sum_{i=1}^{m_X} X_i^2 + \sum_{i=1}^{m_U} U_i^2) + \frac{1-T}{m_X+m_U}(\sum_{i=1}^{m_X} X_i + \sum_{i=1}^{m_U} U_i) \tag{20}$$

$$= \frac{1}{m_X+m_U}(\sum_{i=1}^{m_X}((X_i^2 - X_i)T + X_i) + \sum_{i=1}^{m_U}((U_i^2 - U_i)T + U_i)) \tag{21}$$

  where $T \in \{0, 1\}$ in the binary treatment settings.

Next, we will verify the effectiveness of our model in different data dimensions.

4.2.2 Results.

In this paper, 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 confounders. And we sample 10000 units from **Syn-1-4-4**,**Syn-2-4-4**,**Syn-2-10-4**,**Syn-2-4-10** and perform 10 replications to report the mean and the standard deviation (std) of the bias of the average treatment effect (ATE) estimation in Table 1, where within-sample error is computed over the training sets and out-of-sample error over the test set. From the results, we have following observations: (1) More valid IVs would bring more accuracy on treatment effect estimation by comparing with the results of **Syn-1-4-4** and **Syn-2-4-4**. (2) High dimension of unmeasured confounder would lead to poor performance of confounder balancing based methods by comparing with the results of **Syn-2-4-4** and **Syn-2-4-10**. (3) The existence of observed confounders would make the IV based methods make huge error on treatment effect estimation, even worse than the confounder balancing based methods. This is because current IV based methods ignored the bias of observed confounders in their second stage regression. (4) Considering confounder balancing in IV regression, our CB-IV improved considerably over the traditional IV-based methods and achieved 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 size (500, 1000, 5000, 10000) on **Syn-2-4-4** to study its impact on model performance. Figure 4.2.2 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 (<3000). As the number of data increases, the variance of CB-IV will decrease linearly. When the amount of data exceeds 3000, 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 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.

Table 1: The bias (mean $\pm$ std) of ATE estimation on Synthetic data (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	1.0551 $\pm$ 0.0105	1.0571 $\pm$ 0.0080	1.0920 $\pm$ 0.0091	1.0196 $\pm$ 0.0076
DeepIV-GMM	0.9336 $\pm$ 0.0107	0.8744 $\pm$ 0.0192	0.7684 $\pm$ 0.0232	0.9253 $\pm$ 0.0172
KernelIV	0.4954 $\pm$ 0.0557	0.4573 $\pm$ 0.0541	0.7649 $\pm$ 0.0283	0.6239 $\pm$ 0.0625
DualIV	1.4689 $\pm$ 0.0721	1.4233 $\pm$ 0.0764	1.7189 $\pm$ 0.0756	1.5344 $\pm$ 0.0727
OneSIV	0.8228 $\pm$ 0.0752	0.6613 $\pm$ 0.0955	0.6886 $\pm$ 0.0540	0.8504 $\pm$ 0.0727
DFIV	0.8515 $\pm$ 0.0097	0.8602 $\pm$ 0.0071	0.8506 $\pm$ 0.0072	0.8858 $\pm$ 0.0090
DFL	0.8401 $\pm$ 0.0020	0.8507 $\pm$ 0.0021	0.8380 $\pm$ 0.0015	0.8308 $\pm$ 0.0045
DirectRep	0.1720 $\pm$ 0.0173	0.1630 $\pm$ 0.0084	0.1181 $\pm$ 0.0173	0.1994 $\pm$ 0.0160
CFR	0.1717 $\pm$ 0.0160	0.1582 $\pm$ 0.0151	0.1050 $\pm$ 0.0196	0.1980 $\pm$ 0.0182
DRCFR	0.1514 $\pm$ 0.0557	0.1359 $\pm$ 0.0337	**0.0630 $\pm$ 0.0439**	0.1542 $\pm$ 0.0317
CB-IV	**0.0381 $\pm$ 0.0712**	**0.0160 $\pm$ 0.0470**	0.0774 $\pm$ 0.0413	**0.0092 $\pm$ 0.0646**
Out-of-Sample				
Method	**Syn-1-4-4**	**Syn-2-4-4**	**Syn-2-10-4**	**Syn-2-4-10**
DeepIV-LOG	1.0549 $\pm$ 0.0101	1.0572 $\pm$ 0.0081	1.0931 $\pm$ 0.0091	1.0197 $\pm$ 0.0076
DeepIV-GMM	0.9334 $\pm$ 0.0106	0.8744 $\pm$ 0.0194	0.7682 $\pm$ 0.0229	0.9252 $\pm$ 0.0173
KernelIV	0.4953 $\pm$ 0.0552	0.4581 $\pm$ 0.0525	0.7652 $\pm$ 0.0278	0.6245 $\pm$ 0.0627
DualIV	1.4722 $\pm$ 0.0791	1.4671 $\pm$ 0.0764	1.7321 $\pm$ 0.0722	1.5131 $\pm$ 0.0664
OneSIV	0.8224 $\pm$ 0.0759	0.6612 $\pm$ 0.0950	0.6904 $\pm$ 0.0527	0.8512 $\pm$ 0.0735
DFIV	0.8514 $\pm$ 0.0091	0.8602 $\pm$ 0.0070	0.8507 $\pm$ 0.0071	0.8857 $\pm$ 0.0091
DFL	0.8401 $\pm$ 0.0020	0.8506 $\pm$ 0.0019	0.8383 $\pm$ 0.0016	0.8308 $\pm$ 0.0043
DirectRep	0.1721 $\pm$ 0.0160	0.1635 $\pm$ 0.0090	0.1160 $\pm$ 0.0154	0.1991 $\pm$ 0.0143
CFR	0.1717 $\pm$ 0.0146	0.1586 $\pm$ 0.0185	0.1029 $\pm$ 0.0187	0.1977 $\pm$ 0.0160
DRCFR	0.1511 $\pm$ 0.0548	0.1365 $\pm$ 0.0348	**0.0617 $\pm$ 0.0450**	0.1538 $\pm$ 0.0321
CB-IV	**0.0374 $\pm$ 0.0750**	**0.0165 $\pm$ 0.0456**	0.0748 $\pm$ 0.0401	**0.0096 $\pm$ 0.0640**

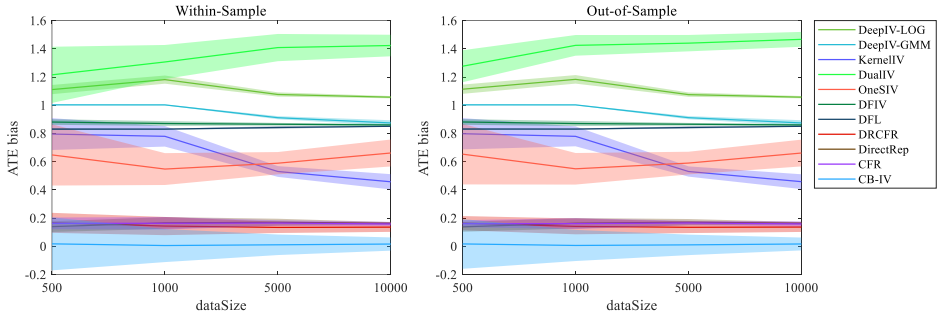


Figure 2: Performance of CB-IV on Syn-2-4-4 by varying data size

4.3 Experiments on Real-World Datasets

4.3.1 Dataset.

We also check the performance of CB-IV methods with experiments on two real-world datasets, which are adopted in Yao et al. (2018); Wu et al. (2020): 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 $P(T \mid Z, X) = \frac{1}{1+\exp\left(-(\sum_{i=1}^{m_Z} Z_i X_i + \sum_{i=1}^{m_X} X_i) + \sum_{i=1}^{m_U} U_i)\right)}, T \sim Bernoulli(P(T \mid Z, X))$.

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

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

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

Table 2: The bias (mean $\pm$ std) of ATE estimation on real-world data (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	2.8736 ± 0.0577	2.6227 ± 0.0651	0.0135 ± 0.0215	0.0237 ± 0.0111
DeepIV-GMM	3.7760 ± 0.0316	3.7396 ± 0.0402	0.0194 ± 0.0047	0.0221 ± 0.0041
KernelIV	3.0605 ± 0.3054	2.9941 ± 0.4634	-	-
DualIV	0.5925 ± 0.2212	0.6581 ± 0.2427	-	-
OneSIV	1.7249 ± 0.3752	1.7411 ± 0.3422	0.0083 ± 0.0191	0.0080 ± 0.0167
DFIV	3.5543 ± 0.0891	3.6218 ± 0.1038	0.0268 ± 0.0005	0.0265 ± 0.0003
DFL	3.2018 ± 0.0496	3.1991 ± 0.0374	0.0624 ± 0.0586	0.0847 ± 0.0049
DirectRep	0.0675 ± 0.0562	0.4600 ± 0.0711	0.0167 ± 0.0171	0.0193 ± 0.0251
CFR	0.0854 ± 0.0579	0.4826 ± 0.0642	0.0115 ± 0.0167	0.0223 ± 0.0176
DRCFR	0.0553 ± 0.0644	0.4336 ± 0.0692	0.0114 ± 0.0221	0.0118 ± 0.0174
CB-IV	$\mathbf{0.0117 \pm 0.3882}$	$\mathbf{0.1601 \pm 0.2499}$	$\mathbf{0.0067 \pm 0.0271}$	$\mathbf{0.0014 \pm 0.0249}$
Out-of-Sample				
Method	**IHDP-2-6-0**	**IHDP-2-4-2**	**Twins-5-8-0**	**Twins-5-5-3**
DeepIV-LOG	2.8760 ± 0.0553	2.6226 ± 0.0692	0.0140 ± 0.0208	0.0238 ± 0.0111
DeepIV-GMM	3.7768 ± 0.0350	3.7388 ± 0.0416	0.0193 ± 0.0047	0.0221 ± 0.0040
KernelIV	3.0703 ± 0.3063	3.0232 ± 0.4401	-	-
DualIV	0.5642 ± 0.2663	0.7147 ± 0.3547	-	-
OneSIV	1.7287 ± 0.3725	1.7351 ± 0.3430	0.0082 ± 0.0191	0.0081 ± 0.0168
DFIV	3.5538 ± 0.0904	3.6225 ± 0.1061	0.0268 ± 0.0005	0.0265 ± 0.0003
DFL	3.2038 ± 0.0496	3.1994 ± 0.0376	0.0624 ± 0.0584	0.0846 ± 0.0046
DirectRep	0.0608 ± 0.0817	0.4571 ± 0.0759	0.0162 ± 0.0175	0.0194 ± 0.0253
CFR	0.0785 ± 0.0810	0.4804 ± 0.0687	0.0110 ± 0.0163	0.0225 ± 0.0180
DRCFR	0.0450 ± 0.0953	0.4321 ± 0.0673	0.0113 ± 0.0219	0.0118 ± 0.0174
CB-IV	$\mathbf{0.0150 \pm 0.3927}$	$\mathbf{0.1578 \pm 0.2540}$	$\mathbf{0.0065 \pm 0.0270}$	$\mathbf{0.0015 \pm 0.0247}$

* 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.

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. (20).

4.3.2 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 2, 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.

5 CONCLUSION

The majority of instrumental variable methods ignore the confounding bias in the second stage in nonlinear scenarios. A promising direction is to implement confounder balance. We confirm this and extend the instrumental variable methods from the additive separability assumptions to a more general scenario with multiplicative assumption through our theoretical and experimental analysis. This leads us to a Confounder Balanced IV Regression (CB-IV) algorithm for causal effect estimation with unobserved confounders. Extensive experiments show that the proposed method achieves state-of-the-art performance in the treatment effect estimation.

In this paper, we mainly focus on treatment effect estimation and have not examined statistical inference yet. Inference after deep neural network training is generally very challenging (Farrell et al., 2021). We leave this to future exploration.

REFERENCES

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

Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects, 1995.

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

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. *arXiv preprint arXiv:1210.4842*, 2012.

Christopher F Baum, Mark E Schaffer, and Steven Stillman. Instrumental variables and gmm: Estimation and testing. *The Stata Journal*, 3(1):1–31, 2003.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *arXiv preprint arXiv:1905.12495*, 2019.

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

Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751, 2007.

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

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, pp. 1779–1788. PMLR, 2020.

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

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

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

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7): 761–767, 2011.

Arthur S Goldberger. Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pp. 979–1001, 1972.

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

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

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

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

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

Miguel A Hernán and James M Robins. Causal inference, 2010.

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

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

Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 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.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. In *IJCAI*, pp. 3768–3774, 2016.

Xiao-Hui Li, Caleb Chen Cao, Yuhan Shi, Wei Bai, Han Gao, Luyu Qiu, Cong Wang, Yuanyuan Gao, Shenjia Zhang, Xun Xue, et al. A survey of data-driven and knowledge-aware explainable ai. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

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

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*, 2017.

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. *arXiv preprint arXiv:1910.12358*, 2019.

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

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

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

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

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

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

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

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

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *arXiv preprint arXiv:1906.00232*, 2019.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

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

James H Stock, Jonathan H Wright, and Motohiro Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20 (4):518–529, 2002.

Alex Strehl, John Langford, Sham Kakade, and Lihong Li. Learning from logged implicit exploration data. *arXiv preprint arXiv:1003.0120*, 2010.

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

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

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

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

Anpeng Wu, Kun Kuang, Junkun Yuan, Bo Li, Pan Zhou, Jianrong Tao, Qiang Zhu, Yueting Zhuang, and Fei Wu. Learning decomposed representation for counterfactual inference. *arXiv preprint arXiv:2006.07040*, 2020.

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

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Junkun Yuan, Anpeng Wu, Kun Kuang, Bo Li, Runze Wu, Fei Wu, and Lanfen Lin. Auto iv: Counterfactual prediction via automatic instrumental variable decomposition. *arXiv preprint arXiv:2107.05884*, 2021.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

A NONLINEAR CASE

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

Proof 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.

Proof 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:

$$\begin{aligned}
ATE_Z &= \mathbb{E}[Y|Z_1, X] - \mathbb{E}[Y|Z_0, X] \\
&= [\mathbb{E}[g(T_1', X)] + \mathbb{E}[U|X]] - [\mathbb{E}[g(T_0', X)] + \mathbb{E}[U|X]] \\
&= \mathbb{E}[g(T_1', X)] - \mathbb{E}[g(T_0', X)]
\end{aligned}$$

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., $\mathbb{E}[\hat{T}X^2 + X + U|\hat{T}, X], \hat{T} = \mathbb{E}[T|Z,X]$. 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. In other words, T is related to X, so there may be multiple different solutions $\hat{g}$ of $\arg\min_{g'}\{\mathbb{E}[\hat{T}X^2 + X + U|\hat{T}, X] - g'(T,X)\}$ and $\hat{g}$ may be different from true structural function g.

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

B THEOREMS

Theorem *(Identification of treatment effects). If the learned representation of observed confounders* $C = f_\theta(X)$ *is independent with the predicted treatment* $\hat{T} \sim P(T \mid Z, X)$*, then the latent outcome function* $h(T,X)$ *can be identified with instrumental variables* Z *and representation* C*:*

$$h(T,X) = g_1(T,X) + g_2(T)\mathbb{E}[g_3(U)|C] + \mathbb{E}[g_4(X,U)|C], C = f_\theta(X) \tag{22}$$

Proof *In this paper, we model the causal relationship more general and relax the additive separability assumption to the multiplicative assumption (Eq. (4)):*

$$T = f_1(Z, X) + f_2(Z)f_3(X, U), Y = g_1(T, X) + g_2(T)g_3(U) + g_4(X, U), Z \perp U, X$$

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

$$\begin{aligned}
\mathbb{E}[T|Z, X] &= \mathbb{E}[f_1(Z, X) + f_2(Z)f_3(X, U)|Z, X] \\
&= \mathbb{E}[f_1(Z, X)|Z, X] + \mathbb{E}[f_2(Z)f_3(X, U)|Z, X] \\
&= f_1(Z, X) + \mathbb{E}[f_2(Z)|Z, X]\mathbb{E}[f_3(X, U)|X] \\
&= f_1(Z, X) + f_2(Z)\mathbb{E}[f_3(X, U)|X]
\end{aligned}$$

where $\mathbb{E}[f_1(Z, X)|Z, X] = f_1(Z, X)$ and $\mathbb{E}[f_2(Z)|Z, X] = f_2(Z)$, because Z and X are independent. We define $\hat{T} = \mathbb{E}[T|Z, X]$.

Outcome Regression Stage, we perform linear/nonlinear regression from $\{Z, C\}$ to Y with $C \perp g_2(T)$ using deep neural networks:

$$\begin{aligned}
\mathbb{E}[Y|Z, C] &= \mathbb{E}[g_1(T, X) + g_2(T)g_3(U) + g_4(X, U)|Z, C] \\
&= \mathbb{E}[g_1(f_1(Z, X) + f_2(Z)f_3(X, U), X) + g_2(T)g_3(U)|Z, C] + \mathbb{E}[g_4(X, U)|C] \\
&= \mathbb{E}[g_1(f_1(Z, X) + f_2(Z)f_3(X, U), X)|Z, C] + \mathbb{E}[g_2(T)g_3(U)|Z, C] + \mathbb{E}[g_4(X, U)|C] \\
&= \mathbb{E}[g_1(\mathbb{E}[T|Z, X], X)|Z, C] + \mathbb{E}[g_2(\mathbb{E}[T|Z, X])g_3(U)|Z, C] + \mathbb{E}[g_4(X, U)|C] \\
&= \mathbb{E}[g_1(\hat{T}, X)|Z, C] + \mathbb{E}[g_2(\hat{T})g_3(U)|Z, C] + \mathbb{E}[g_4(X, U)|C] \\
&= g_1(\hat{T}, X) + g_2(\hat{T})\mathbb{E}[g_3(U)|C] + \mathbb{E}[g_4(X, U)|C]
\end{aligned}$$

As for step 3 to step 4:

$$\begin{aligned}
\mathbb{E}[g_1(\mathbb{E}[T|Z, X], X)|Z, C] &= \mathbb{E}[g_1(f_1(Z, X) + f_2(Z)\mathbb{E}[f_3(X, U)|X], X)|Z, C] \\
&= \mathbb{E}[g_1(f_1(Z, X) + f_2(Z)f_3(X, U), X)|Z, C] \\
&= \mathbb{E}[g_1(T, X)|Z, C]
\end{aligned}$$

where $\mathbb{E}[f_3(X, U)|X]$, only related to X, is a constant for the specified X/C. 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). An example of unique solution can be found in Proof 1 (b) in Section 1.

As for step 5 to step 6:

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

where $g_2(\mathbb{E}[T|Z, X])$ only related to Z and X, and is independent of $g_3(U)$ conditional on Z and C. Note that $g_2(\mathbb{E}[T|Z, X])$ and $g_3(U)$ are conditionally related conditional on X, but conditionally independent given C.

Summarily, the latent outcome function is $h(T, X) = g_1(T, X) + g_2(T)\mathbb{E}[g_3(U)|C]$, and can be identified by IVs.

C Binary Treatment and Binary Outcome Case

In this paper, we model the causal relationship more general and relax the additive separability assumption to the multiplicative assumption, as follows:

$$T = f_1(Z, X) + f_2(Z)f_3(X, U), Y = g_1(T, X) + g_2(T)g_3(U) + g_4(X, U), Z \perp U, X \quad (23)$$

Table 3: 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
	α	0.01/0.001	0.1	0.001/0.0001

where $f_{1\cdots3}(\cdot)$ and $g_{1\cdots4}(\cdot)$ are continuous functions. In the structural function of Y, $g_2(T)g_3(U)$ denotes the multiplicative terms of U with T (e.g., $U^2T - UT + U$). The same principle can be applied to the structural function of T. 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:

$$T \sim Bernoulli(P(T)), \text{where } P(T) = \frac{1}{1+\exp^{-(f_1(Z,X)+f_2(Z)f_3(X,U))}},$$
$$Y \sim Bernoulli(P(Y)), \text{where } P(Y) = \frac{1}{1+\exp^{-(g_1(T,X)+g_2(T)g_3(U)+g_4(X,U))}},$$
$$\log\frac{P(T)}{1-P(T)} = f_1(Z,X) + f_2(Z)f_3(X,U), \log\frac{P(Y)}{1-P(Y)} = g_1(T,X) + g_2(T)g_3(U) + g_4(X,U), Z \perp U, X \quad (24)$$

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 Pseudo-Code and Hyper-parameters

D.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:

$$\min_{\mu} \mathcal{L}_1 = -\frac{1}{n}\sum_{i=1}^{n}\left(t_i \log\left(\pi_\mu(z_i, x_i)\right) + \left(1 - t_i\right)\left(1 - \log\left(\pi_\mu(z_i, x_i)\right)\right)\right) \quad (25)$$

$$\min_{\theta,\xi^0,\xi^1} \mathcal{L}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{\hat{t}\in\{0,1\}} h_{\xi^{\hat{t}}}(f_\theta(x_i))\hat{P}(\hat{t} \mid z_i, x_i)\right)^2 + \alpha \,\text{disc}(\hat{t}, f_\theta(x_i)) \quad (26)$$

where α 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 π_μ 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 π_μ with a loss $\mathcal{L}_1$ 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 $\mathcal{L}_2$ jointly. To prevent overfitting, we add a regularization term to regularize the prediction functions h_{ξ^0}, h_{ξ^1} with a small l_2 weight decay.

Algorithm 1 Two(2)-Stage Representation learning with Instrumental Variables

Input: Observational data $\mathbb{D} = \{z_i, x_i, t_i, y_i\}_{i=1}^n$, The 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: $\mathcal{L}_1$ and $\mathcal{L}_2$
Components: A logistic regression network $\pi_\mu(\cdot$; a representation learning network $f_\theta(\cdot)$; two-head outcome regression networks $h_{\xi^0}(\cdot)$ and $h_{\xi^1}(\cdot)$.
Treatment Regression Stage:
for $i = 1, 2, 3, ...$ **do**
    $\{z_i, x_i\}_{i=1}^n \to \pi_\mu(z_i, x_i) \to \hat{P}(t = 1 \mid z_i, x_i)$
    $\mathcal{L}_1 = -\frac{1}{n}\sum_{i=1}^n \left(t_i \log\left(\pi_\mu(z_i, x_i)\right) + (1 - t_i)\left(1 - \log\left(\pi_\mu(z_i, x_i)\right)\right)\right)$
    update $\mu \leftarrow \mathrm{SGD}\{\mathcal{L}_1\}$
end for
Outcome Regression Stage:
for $i = 1, 2, 3, ..., \mathcal{I}$ **do**
    $\{x_i\}_{i=1}^n \to C_i = f_\theta(x_i)$
    $\{z_i, x_i\}_{i=1}^n \to \pi_\mu(z_i, x_i) \to \hat{P}(t = 1 \mid z_i, x_i)$
    $\{f_\theta(x_i), t_i\}_{i=1}^n \to \mathrm{disc}(\hat{t}, f_\theta(x_i))$
    $\mathcal{L}_2 = \frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{\hat{t}\in\{0,1\}} h_{\xi^{\hat{t}}}(f_\theta(x_i))\hat{P}(\hat{t} \mid z_i, x_i)\right)^2 + \mathrm{disc}(\hat{t}, f_\theta(x_i))$
    update $\theta, \xi^0, \xi^1 \leftarrow \mathrm{Adam}\{\mathcal{L}_2\}$
end for

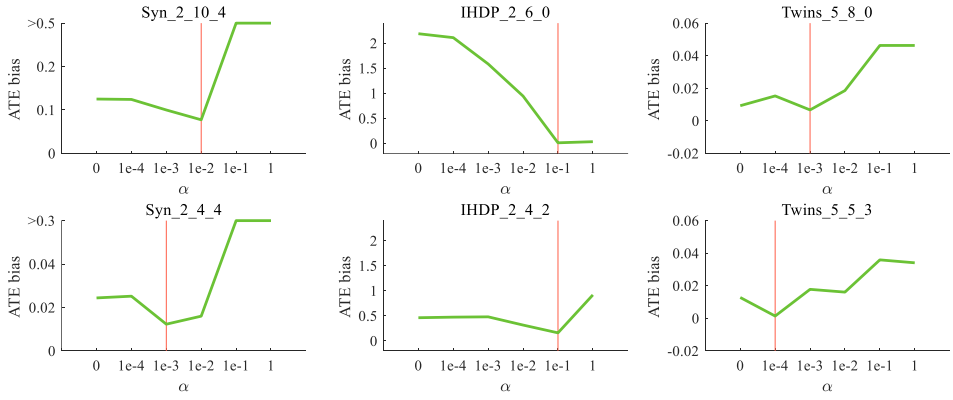


Figure 3: Hyper-parameter sensitivity analysis on Data-m_Z-m_X-m_U. The green lines show the ATE bias of the hyper-parameter α within the specified range $\{0, 0.0001, 0.001, 0.01, 0.1, 1\}$. The red line indicates the parameters chosen by CB-IV.

Table 3 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.

D.2 HYPER-PARAMETERS ANALYSIS ON DATA-m_Z-m_X-m_U

Given the multi-term objective function (Eq. (17)) in CB-IV, we study the confounder balance item (Eq. (13)/(14)) on the average treatment effect estimation of different datasets (Data-m_Z-m_X-m_U) by changing hyper-parameter α 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.

E 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 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 4. 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.

F THE CONTINUOUS TREATMENT EXPERIMENTS

F.1 DEMAND DATASETS WITH DIFFERENT SAMPLE SIZE

In demand Datasets (that applied in DeepIV (Hartford et al., 2017), KernelIV (Singh et al., 2019), DualIV (Muandet et al., 2019) and DFIV (Xu et al., 2020)), we report mean squared error (MSE) and its standard deviations over 10 trials with different data sizes (500, 1000, 5000, 10000): 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 $X_1 \in \{1, \ldots, 7\}$, $X_2 \sim \text{unif}(0, 10)$, $Z, U \sim \text{N}(0, 1)$ and $E \sim \text{N}(0.5U, 0.75)$. 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 $\{U, E\}$.

Like the binary treatment studies in this paper, on this classical simulation data Demand (Table F.2), the balanced representation methods without using IV still perform much better than the pure IV-based methods. Considering confounder balancing in IV regression, our method CB-IV improved considerably over the traditional IV-based methods and achieved better performance than confounder balancing methods in most settings. Nevertheless, our method still relies on large sam-

Table 4: The bias (mean ± std) of average treatment effect estimation on Synthetic data (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	1.0551 ± 0.0057	1.0545 ± 0.0072	1.0588 ± 0.0093	1.0571 ± 0.0080
DeepIV-GMM	0.8617 ± 0.0164	0.9915 ± 0.0066	0.9607 ± 0.0059	0.8744 ± 0.0192
KernelIV	0.9639 ± 0.0698	0.8654 ± 0.1742	0.8897 ± 0.1573	0.4573 ± 0.0541
DualIV	0.6582 ± 0.5607	1.6109 ± 0.4953	1.7628 ± 0.0423	1.4233 ± 0.0764
OneSIV	1.0477 ± 0.0304	1.1760 ± 0.0457	1.0529 ± 0.0448	0.6613 ± 0.0955
DFIV	1.0028 ± 0.0096	0.8945 ± 0.0037	0.8377 ± 0.0066	0.8602 ± 0.0071
DFL	0.8422 ± 0.0016	0.8428 ± 0.0019	0.8423 ± 0.0017	0.8507 ± 0.0021
DirectRep	0.1630 ± 0.0084	0.1630 ± 0.0084	0.1783 ± 0.0224	0.1630 ± 0.0084
CFR	0.1582 ± 0.0151	0.1582 ± 0.0151	0.1775 ± 0.0234	0.1582 ± 0.0151
DRCFR	**0.1359 ± 0.0337**	**0.1359 ± 0.0337**	0.1414 ± 0.0536	0.1359 ± 0.0337
CB-IV	0.4953 ± 0.2631	**0.5294 ± 0.0996**	**0.1145 ± 0.0717**	**0.0160 ± 0.0470**
Out-of-Sample				
Method	**Syn(Z)(X)**	**Syn(X)(X)**	**Syn(Z,X)(Z,X)**	**Syn(Z,X)(X)**
DeepIV-LOG	1.0552 ± 0.0054	1.0546 ± 0.0075	1.0591 ± 0.0097	1.0572 ± 0.0081
DeepIV-GMM	0.8618 ± 0.0164	0.9915 ± 0.0066	0.9606 ± 0.0059	0.8744 ± 0.0194
KernelIV	0.9634 ± 0.0699	0.8651 ± 0.1767	0.9164 ± 0.1573	0.4581 ± 0.0525
DualIV	0.8002 ± 0.3073	1.6063 ± 0.5008	1.7601 ± 0.0371	1.4671 ± 0.0527
OneSIV	1.0478 ± 0.0302	1.1763 ± 0.0453	1.0526 ± 0.0448	0.6612 ± 0.0950
DFIV	1.0027 ± 0.0095	0.8944 ± 0.0037	0.8375 ± 0.0065	0.8602 ± 0.0070
DFL	0.8421 ± 0.0016	0.8427 ± 0.0017	0.8421 ± 0.0015	0.8506 ± 0.0019
DirectRep	0.1635 ± 0.0090	0.1635 ± 0.0090	0.1787 ± 0.0192	0.1635 ± 0.0090
CFR	0.1586 ± 0.0185	0.1586 ± 0.0185	0.1777 ± 0.0233	0.1586 ± 0.0185
DRCFR	**0.1365 ± 0.0348**	**0.1365 ± 0.0348**	0.1416 ± 0.0517	0.1365 ± 0.0348
CB-IV	0.4929 ± 0.2614	**0.5285 ± 0.0994**	**0.1144 ± 0.0714**	**0.0165 ± 0.0456**

ples. The contribution of this paper is to find this phenomenon and give a practical solution, and we relax the additive assumption.

F.2 Demand datasets with different structual functions of T

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 γ and λ 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 F.2) shows that if the information of instrumental variables and confounders increases, all methods will become worse, but the balanced representation methods without using IV still perform much better than the pure IV based methods. If we only increase the information of the instrumental variable, the results of the pure IV based methods and 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 regression them. Anyway, the confounding bias comes from the treatment regression stage is a very important problem in IV based methods.

G Discussion on confounder balancing

G.1 Confounder Balancing

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. This change could result in confounding bias: $\mathbb{P}(T|X) \neq \mathbb{P}(T)$. If we directly regress $\mathbb{E}[Y|T, X] = h_\xi(T, X)$, in binary treatment case , such as the hospital scenario, most patients (have an injection) in the treated group have severe comorbidity, i.e., $\mathbb{P}(T = \text{injection}|X =$

Table 5: The MSE (mean ± std) of latent outcome estimation on Demand data

Within-Sample				
Method	**500**	**1000**	**5000**	**10000**
DeepIV-LOG	-	-	-	-
DeepIV-GMM	7197.0858 ± 591.5079	11199.8894 ± 6482.5072	3163.3388 ± 266.4328	1356.3735 ± 343.5231
KernelIV	3078.2122 ± 647.2202	2363.3228 ± 270.7994	1692.1801 ± 72.6865	1526.4373 ± 141.7145
DualIV	13462.7471 ± 4882.1326	12839.8616 ± 5159.1546	28532.6462 ± 15774.3332	>30000
OneSIV	6196.9547 ± 1931.3269	6879.3032 ± 1940.6865	8784.7186 ± 1200.5437	8203.8744 ± 1120.1937
DFIV	240.0821 ± 381.7838	152.4014 ± 52.8385	198.9294 ± 30.6243	195.2834 ± 9.3424
DFL	141.4824 ± 26.4270	173.2734 ± 29.9088	196.8437 ± 17.8268	195.9884 ± 11.1385
DirectRep	138.7284 ± 24.0162	153.4422 ± 16.6723	193.0451 ± 12.8752	191.2359 ± 5.5144
CFR	126.9027 ± 20.9857	161.7175 ± 20.9926	191.6562 ± 10.2437	193.3015 ± 5.5614
DRCFR	705.7547 ± 462.9351	503.0686 ± 240.5934	419.0754 ± 126.1294	427.2194 ± 162.0811
CB-IV	**117.6441 ± 23.2538**	**142.0652 ± 16.1174**	**164.6670 ± 7.4433**	**165.0155 ± 5.9588**
Out-of-Sample				
Method	**500**	**1000**	**5000**	**10000**
DeepIV-LOG	-	-	-	-
DeepIV-GMM	7249.5025 ± 465.7548	11470.8863 ± 6643.9238	3360.7893 ± 483.8971	1006.6206 ± 313.7140
KernelIV	2859.5013 ± 660.9105	2280.9724 ± 547.9235	1142.8346 ± 170.3749	994.9508 ± 146.2092
DualIV	12101.9675 ± 3948.9629	12455.8961 ± 2916.7411	27940.9859 ± 14022.5994	>30000
OneSIV	6539.1699 ± 1788.4552	7088.5011 ± 1846.4845	8883.0686 ± 988.7041	8330.8031 ± 1026.3255
DFIV	764.4473 ± 415.1062	404.9916 ± 133.0858	214.4949 ± 30.6644	190.5521 ± 8.9768
DFL	358.7445 ± 47.3268	261.3345 ± 35.6856	192.7698 ± 14.4607	182.9253 ± 11.5256
DirectRep	271.8334 ± 25.7621	**222.3286 ± 9.5751**	199.8683 ± 5.4527	193.9514 ± 7.3804
CFR	**266.2449 ± 28.4544**	225.9142 ± 11.7594	195.8037 ± 11.3383	192.0922 ± 8.9325
DRCFR	799.8020 ± 467.5651	621.7465 ± 275.9710	511.0712 ± 155.0455	532.4370 ± 199.5613
CB-IV	291.4765 ± 39.3366	229.1330 ± 42.2210	**179.4130 ± 4.2211**	**172.9054 ± 5.3395**

* 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., 2019) and DFIV (Xu et al., 2020). The difference is that they scale the results by log10, but we don't.

Table 6: The MSE (mean ± std) of latent outcome estimation on different Demand datasets (Demand-γ-λ)

Within-Sample			
Method	**Demand-0-1**	**Demand-0-5**	**Demand-5-1**
DeepIV-LOG	-	-	-
DeepIV-GMM	1356.3735 ± 343.5231	3102.8901 ± 744.4496	1465.5604 ± 253.3932
KernelIV	1526.4373 ± 141.7145	5772.8724 ± 413.1272	1428.5166 ± 227.3451
DualIV	>30000	>30000	>30000
OneSIV	8203.8744 ± 1120.1937	30854.0811 ± 3647.2961	7892.6823 ± 2009.5216
DFIV	195.2834 ± 9.3424	1205.0481 ± 1740.5136	197.2478 ± 16.8028
DFL	195.9884 ± 11.1385	1159.9531 ± 1902.0860	200.3554 ± 8.9157
DirectRep	191.2359 ± 5.5144	888.6762 ± 1077.6299	440.0853 ± 117.3984
CFR	193.3015 ± 5.5614	465.3831 ± 181.4856	449.6735 ± 161.0288
DRCFR	427.2194 ± 162.0811	391.6148 ± 28.2101	405.8180 ± 105.9513
CB-IV	**165.0155 ± 5.9588**	**234.1836 ± 30.0674**	**167.7809 ± 6.7831**
Out-of-Sample			
Method	**Demand-0-1**	**Demand-0-5**	**Demand-5-1**
DeepIV-LOG	-	-	-
DeepIV-GMM	1006.6206 ± 313.7140	2829.4425 ± 724.6786	1151.6218 ± 284.1778
KernelIV	994.9508 ± 146.2092	5435.9011 ± 435.2851	1004.7321 ± 216.7744
DualIV	>30000	>30000	>30000
OneSIV	8330.8031 ± 1026.3255	18508.7687 ± 2341.9042	7856.9271 ± 1977.9162
DFIV	190.5521 ± 8.9768	668.3026 ± 566.7304	196.2839 ± 16.6671
DFL	182.9253 ± 11.5256	597.6806 ± 622.1575	189.7124 ± 7.4217
DirectRep	193.9514 ± 7.3804	689.6526 ± 692.1083	489.9140 ± 121.1920
CFR	192.0922 ± 8.9325	417.2996 ± 123.5452	469.7471 ± 140.7833
DRCFR	532.4380 ± 199.5613	497.3451 ± 26.3724	470.5751 ± 143.4208
CB-IV	**172.9054 ± 5.3395**	**224.3519 ± 18.0629**	**165.8571 ± 7.1423**

severe comorbidity) $>$ $\mathbb{P}(T = \text{injection}|X = \text{mild comorbidity})$. 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 $\mathbb{P}(X|T=1)=\mathbb{P}(X|T=0)$.

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(Zubizarreta, 2015; Athey et al., 2018; He & Garcia, 2009), Doubly Robust (Funk et al., 2011) or backdoor criterion (Pearl, 2009) 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., 2021). 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.

If we directly regress $\mathbb{E}[Y|T,X]=h_\xi(T,X)$, 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**: Confounders are the cause of the treatment variable, the information of treatment variables may be replaced by confounders in outcome regression, resulting in the consistency of the predicted potential outcomes from different treatments for the specified X, i.e., $h_\xi(0,X_{T=t})=h_\xi(1,X_{T=t})=h_\xi(X_{T=t})$, $X_{T=t}$ 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 learning balanced 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 affect the treatments and does not affect the outcomes directly, can be used to eliminate the unmeasured confounding bias (Pearl et al., 2000; Wright, 1928a; 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:

$$\epsilon(h,\theta) \leq 2\left(\epsilon_F^{t=0}(h,\theta) + \epsilon_F^{t=1}(h,\theta) + B_\theta IPM_{\mathrm{G}}\left(p_\theta^{t=1}, p_\theta^{t=0}\right) - 2\sigma_Y^2\right) \tag{27}$$

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_θ is a constant; σ_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 (See Proof 1(b) for details). 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. 17) as the loss function in the outcome regression stage:

$$\min_{\theta,\xi^0,\xi^1} \mathcal{L}_2 = \frac{1}{n}\sum_{i=1}^n \left(y_i - \sum_{\hat{t}\in\{0,1\}} h_{\xi^{\hat{t}}}(f_\theta(x_i))\hat{P}(\hat{t} \mid z_i, x_i)\right)^2 + \alpha \,\mathrm{disc}(\hat{t}, f_\theta(x_i))$$

In mathematical, the optimization goal $\mathcal{L}_2$ is 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 $\mathbb{E}[Y|T,X] = h_\xi(T,X)$, nonparametric 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.