A Generative Adversarial Framework for Bounding Confounded Causal Effects

Yaowei Hu,[1] **Yongkai Wu,** [2] **Lu Zhang,** [1] **Xintao Wu** [1]

[1] University of Arkansas
[2] Clemson University
yaoweihu@uark.edu, yongkaw@clemson.edu, lz006@uark.edu, xintaowu@uark.edu

Abstract

Causal inference from observational data is receiving wide applications in many fields. However, unidentifiable situations, where causal effects cannot be uniquely computed from observational data, pose critical barriers to applying causal inference to complicated real applications. In this paper, we develop a bounding method for estimating the average causal effect (ACE) under unidentifiable situations due to hidden confounding based on Pearl's structural causal model. We propose to parameterize the unknown exogenous random variables and structural equations of a causal model using neural networks and implicit generative models. Then, using an adversarial learning framework, we search the parameter space to explicitly traverse causal models that agree with the given observational distribution, and find those that minimize or maximize the ACE to obtain its lower and upper bounds. The proposed method does not make assumption about the type of structural equations and variables. Experiments using both synthetic and real-world datasets are conducted.

Introduction

Inferring causal effects from observational data is an important task in many fields, including economics (Zhao, Runfola, and Kemper 2017), healthcare (Hu and Kerschberg 2018; Wang and Wu 2019), fair machine learning (Zhang and Bareinboim 2018; Zhang, Wu, and Wu 2018; Kusner et al. 2017), etc. Pearl's structural causal model (Pearl 2009) is a decent and widely adopted mathematical framework for conducting causal inference. In the structural causal model, intervention is a technique that fixes some variable to certain constant without changing other parts of the model. Facilitated by intervention, the average causal effect (ACE) of one variable A on another variable B can be formulated as the average change in B due to interventions on A. The calculation of the ACE from the observational data depends on a causal graph that specifies the parental relationship among variables. Under the assumption of no hidden confounding (i.e., no unobserved common causes of A and B), the ACE can be calculated using the well-known truncated factorization formula (Pearl 2009). However, the no-hidden confounding assumption is usually over-simplified and needs to be relaxed in practice. When hidden confounders exist, the ACE may not be uniquely calculated from the observational data without further assumptions, known as the unidentifiable situation (Shpitser and Pearl 2008). In an unidentifiable situation, any estimation of ACEs only based on the observational distribution is not guaranteed to be correct, since there can be other underlying causal models that also agree with the observational distribution but result in different ACEs (Avin, Shpitser, and Pearl 2005).

In order to estimate the ACE in unidentifiable situations, researchers seek for bounding approaches. In (Balke and Pearl 1997), the authors develop a constrained optimization problem for discovering bounds from the observational data which are guaranteed to be tightest. The general idea is to shift the randomness of the causal model from the distributions of U to the distributions of mappings so that it covers all possible domains of U, either categorical, continuous, or mixed. Then, the optimization is to search mapping distributions with the maximum/minimum ACEs subject to the constraint that the joint distribution must be consistent with the data. In a recent work (Wu et al. 2019), the authors extend this idea to bound unidentifiable path-specific effects and counterfactual effects in addition to the ACE. However, these methods are limited to categorical observed variables since the number of distinct functional mappings between continuous variables are infinite, leading to an infinite number of variables in the optimization problem.

In this paper, we extend the method in (Balke and Pearl 1997) for bounding ACEs to continuous and possibly high-dimensional variables, thanks to the significant progress of generative models made by the machine learning community in past years. Following the spirit of (Balke and Pearl 1997), we also shift the randomness of the causal model from the distribution of U to the distribution of all possible mappings from A to B. We define a distribution over the space of all mappings from A to B, which summarizes all possible equations from (U, A) to B for all possible U. Since the mapping from A to B can be in arbitrary forms, we use the neural network as a universal estimator of mappings so that the space of all possible mappings is estimated by the parameter space Θ of the neural network. Then, given any distribution over Θ, i.e., $P(\theta)$, both the joint distribution and the ACE can be expressed as functions of $P(\theta)$. In order for the induced joint distribution to fit the observational data, we adopt the generative adversarial learning framework, where a generator is designed to produce $P(\theta)$, and a discriminator is designed to distinguish the generated distribution from the real distribution. Finally, we combine task of finding the highest and lowest values of the ACE with the adversarial learning. The training procedure of the generator is to find a $P(\theta)$ that maximizes/minimizes the ACE, under the condition that the discriminator is unable to identify the generated distribution. In the experiments we demonstrate that, our method can provide more accurate estimations to the ACE than several widely used causal inference methods including instrumental variable estimation (Bowden and Turkington 1984) and propensity score adjusted regression (Abdia et al. 2017).

Handling hidden confounders is one of the most important aspects of inferring ACEs from observational data. Some recent works try to estimate the representation of hidden confounders using latent-variable models (e.g., (Louizos et al. 2017; Chiappa 2019; Madras et al. 2019)). However, these works don't consider the identifiability issue. Grounded on the method in (Balke and Pearl 1997), our work provides a much more reliable estimation of the ACE than previous works. Since we don't make any assumption about the type of structural equations and variables, the proposed method can serve as a general basis of applying the structural causal model to a wider range of applications. Last but not least, in some practical situations it may be reasonable to assume a certain type of equations. Our method can be simply modified to encode such assumption. We include a subsection showing that encoding the linear equation assumption can make the bounds converge to a fixed value.

Related Work

To infer causal effects from observational data, many causal inference methods have been proposed, either based on Pearl's structural causal model framework (Pearl 2009) or Rubin's potential outcome framework (Rubin 2005). Widely adopted classic methods include propensity score based methods (Abdia et al. 2017), causal graph based methods (Pearl 2010), instrumental variable estimation (Bowden and Turkington 1984), etc. In recent years, it has been proposed to use machine learning models to facilitate causal inference. In (Shalit, Johansson, and Sontag 2017), the authors estimated the causal effect under the no-hidden confounding assumption by using a neural network structure that learns a balanced representation for the treated and control distributions. In (Louizos et al. 2017; Chiappa 2019; Madras et al. 2019), a latent representation was learned to summarize the exogenous variable space of the causal model by using deep latent-variable models such as the variational auto-encoding (VAE). Then, causal effects can be estimated based on the latent-variable models via sampling approaches. In (Yoon, Jordon, and Van Der Schaar 2018), the authors used generative adversarial nets (GAN) to generate counterfactual outcomes such that the discriminator cannot distinguish the counterfactual outcomes from the factual outcomes, and then estimate the ACE based on the potential outcome framework. In (Kocaoglu et al. 2018; Xu et al. 2019), the authors also used GAN but propose to arrange the network structure of the generator to preserve the structure of a given causal graph. A recent work (Li et al. 2020) developed a GAN based deconfouding algorithm assuming no hidden confounding.

One critical issue in causal inference is the identifiability, which is not paid enough attention in many of these work. Unidentifiable situations refer to that the unique estimation of certain causal effect on a causal graph is theoretically infeasible (Shpitser and Pearl 2008). In this sense, the ACE in unidentifiable situations cannot be point identified from the data and one can only seek for estimating bounds. The interval estimation of the ACE based on the potential outcome framework has been studied in (Kallus, Mao, and Zhou 2018), where a functional interval estimator is derived from a weighted kernel regression. However, this work only applies to the case where A is a binary treatment. In the structural causal model framework, the reasons of leading to unidentifiable situations have been studied in the causal inference literature, as summarized in (Wu et al. 2019), where hidden confounding is one of the typical reasons. Different bounding techniques have been proposed to address different unidentifiable situations. For example, (Miles et al. 2015) derived bounds for the natural direct effect with a discrete mediator. In (Zhang, Wu, and Wu 2018), the authors proposed a method to bound unidentifiable path-specific effects due to the existence of the "kite graph". In (Wu, Zhang, and Wu 2019), a method for bounding unidentifiable counterfactual effects due to the existence of the "w graph" was developed. However, none of these methods can apply to unidentifiable situations due to hidden confounders. In (Wu et al. 2019), the authors proposed a general framework for bounding all unidentifiable situations, which is extended from (Balke and Pearl 1997) and also inherit its limitation that all endogenous variables must be categorical. Our work addresses the limitation of (Balke and Pearl 1997; Wu et al. 2019) by a novel adaptation of the generative adversarial learning framework. Following the idea of (Wu et al. 2019), our method can also extend to all unidentifiable situations in addition to those caused by hidden confounders. This will be our future work.

Preliminaries

We develop our method based on Pearl's structural causal model framework, which is formally defined as follows.

Definition 1 (Structural Causal Model (Pearl 2009))**.** *A structural causal model $\mathcal{M}$ is represented by a quadriple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ where*

1. $\mathbf{U}$ *is a set of exogenous random variables that are determined by factors outside the model.*
2. $P(\mathbf{U})$ *is a joint probability distribution defined over* $\mathbf{U}$.
3. $\mathbf{V}$ *is a set of endogenous variables that are determined by variables in* $\mathbf{U} \cup \mathbf{V}$.
4. $\mathbf{F}$ *is a set of structural equations from* $\mathbf{U} \cup \mathbf{V}$ *to* $\mathbf{V}$. *Specifically, for each* $V \in \mathbf{V}$, *there is a function* $f_V \in \mathbf{F}$ *mapping from* $\mathbf{U} \cup (\mathbf{V} \backslash V)$ *to* V, *i.e.,* $v = f_V(\mathsf{pa}_V, u_V)$, *where* pa_V *and* u_V *are realization of a set of endogenous variables* $\mathsf{PA}_V \in \mathbf{V} \setminus V$ *and a set of exogenous variables* U_V *respectively.*

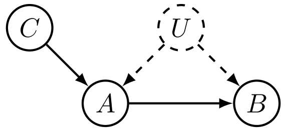

Figure 1: Example: A, B, C are observed variables and U is a hidden variable.

U	C	$A = f_A$
0	0	0
0	1	0
1	0	1
1	1	1

Table 1: Equation $f_A(c, u)$ for determining values of A.

U	A	$B = f_B^1$	$B = f_B^2$
0	0	0	0
0	1	1	0
1	0	0	0
1	1	1	1

Table 2: Equations $f_B^1(a, u)$ and $f_B^2(a, u)$ for determining values of B.

If all exogenous variables in $\mathbf{U}$ are assumed to be mutually independent, then the causal model is called a *Markovian model*; otherwise, it is called a *semi-Markovian model*. For example, a causal model represented by Figure 1 is a semi-Markovian model in which U_A and U_B are completely correlated and denoted by a single exogenous variable U.

In general, $f_V(\cdot)$ can be an equation of any type. In some cases people may assume $f_V(\cdot)$ to be a certain type of equation. For example, if for each node V, $f_V(\cdot)$ is a linear equation, then the causal model is called a *linear causal model*; if $f_V(\cdot)$ is an additive equation, i.e., $v = f_V(\mathsf{pa}_V) + u_V$, then the causal model is called an *additive causal model*.

Each causal model $\mathcal{M}$ is associated with a causal graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of edges. Each node of $\mathcal{V}$ corresponds to a variable of $\mathbf{V}$ in $\mathcal{M}$. Each edge in $\mathcal{E}$, denoted by a directed arrow $\rightarrow$, points from a node $A \in \mathbf{U} \cup \mathbf{V}$ to a different node $B \in \mathbf{V}$ if f_B uses values of A as input. Apparently, two different causal models are associated with the same causal graph if they have the same inputs for all equations.

An intervention on endogenous variable A is defined as the substitution of structural equation $f_A(\mathsf{PA}_A, U_A)$ with a constant a, denoted as $do(A = a)$ or $do(a)$ for short. For another endogenous variable B which is affected by the intervention, its distribution under the intervention, called the post-intervention distribution, is denoted as $P(B|do(a))$. The ACE is defined as the difference of expected values of B under two different interventions (Pearl 2009).

Definition 2. *The average causal effect (ACE) of A on B is given by* $\mathbb{E}[B|do(a_1)] - \mathbb{E}[B|do(a_0)]$.

Given a causal graph, if there exist two or more different causal models associated with the causal graph that agree with an observational distribution but result in different ACE values, then this ACE is said to be *unidentifiable* (Avin, Shpitser, and Pearl 2005). The complete graphical criterion of ACE identifiability has been studied in (Shpitser and Pearl 2008) known as the "hedge criterion".

An unidentifiable example. Figure 1 gives a toy example where the ACE of A and B cannot be uniquely identified due to a hidden confounder U. Consider two causal models with the same causal graph shown in Figure 1. Assume that the two models have the same equations for determining the values of C and A, but differ in the equation for determining the value of B, as shown in Tables 1 and 2. Denoting the probability $P(U = 1)$ as p, we can compute the joint distribution $P(A = a, B = b, C = c)$ represented by each model by summarizing the probabilities of all values of U that lead to $A = a, B = b$ and $C = c$. For example, we can obtain that $P(A = 0, B = 0, C = 0) = (1 - p)P(C = 0)$ for both models since in both models U must be 0 to get $A = 0$ and $B = 0$ when $C = 0$. In fact, one can verify that two models completely agree with the joint distribution $P(A, B, C)$. On the other hand, $P(B = b|do(A = a)) = \sum_{u: f_B(a,u)=b} P(u)$ is computed by summarizing the probabilities of all values of U that lead to $B = b$ when fixing the value of A to a. One can verify that, for the first model (with equation f_B^1), $P(B = 1|do(A = 1)) = 1$ and $P(B = 1|do(A = 0)) = 0$; for the second model (with equation f_B^2), $P(B = 1|do(A = 1)) = p$ and $P(B = 1|do(A = 0)) = 0$. Thus, the ACE is 1 for the first model and p for the second. Assume that the true model is either one of the two. Since they represent the same joint distribution, there is no way to distinguish between them based on the observational data. If we randomly guess the true model from the two, the bias in estimating ACE can vary from 0 to 1 depending on the value of p.

Bounding ACEs via Optimization

According to (Avin, Shpitser, and Pearl 2005), if an ACE is unidentifiable, then there must exist multiple causal models that have the same observational distribution but produce different values of the ACE. Since there is no way to distinguish these causal models using the observational data, we are unable to compute the accurate ACE without further knowledge or assumption about the underlying data generating mechanism. However, from all these possible causal models if we can find the ones that produce the maximal/minimal ACEs, then the maximum/minimum in fact provide the tightest bounds of the unidentifiable ACE. As such, the bounding exercise can be formulated as mathematical optimization problems for maximizing/minimizing the ACE, where we need to examine all possible causal models. The challenge is that, in general we have no knowledge about the equations and exogenous variables of the possible causal models. Thus, for optimization we need to intentionally traverse all possible variable dimensions and domains, all kinds of variable distributions, and all types of equations, which is computationally impossible.

In (Balke and Pearl 1997), the authors proposed to partition the domain of each exogenous variable into a limited number of equivalent classes, each inducing a distinct functional mapping between endogenous variables. These functional mappings are called the *response functions*. Consider an endogenous variable $V \in \mathbf{V}$, whose equation $v = f_V(\mathsf{pa}_V, u_V)$ is a mapping from PA_V, U_V to V. In general, U_V can be with arbitrary domain size, and f_V can be any function. However, for a given value u_V, no matter what dimension and domain U_V has, f_V is a deterministic mapping from endogenous variables PA_V to V, i.e., a response function from PA_V to V. Thus, by denoting response functions as values of a variable R_V, called the *response-function variable*, the distribution of U_V, i.e., $P(u_V)$, can be equivalently partitioned and translated into the distribution $P(r_V)$. Then, as U_V varies along its domain, the only effect it can have on V is to switch the response function among all possible response functions from PA_V to V in the domain of R_V. It is known that we can use $P(\mathbf{u})$ to express the joint distribution $P(\mathbf{v})$ as well as all ACEs (Tian and Pearl 2000). The relationship between U_V and R_V means that we can similarly use $P(\mathbf{r})$ to do the same, where $\mathbf{R}$ denotes the set of all response-function variables. As a result, by expressing $P(\mathbf{v})$ and the ACE using $P(\mathbf{r})$, we can compute the lower or upper bound of the ACE by searching for the $P(\mathbf{r})$ that minimizes or maximizes the ACE, subject to that $P(\mathbf{v})$ agrees with the observational data. This would give us a linear programming problem with $P(\mathbf{r})$ as variables.

As can be seen, if PA_V and/or V are continuous variables, there will be an infinite number of response functions from PA_V to V, resulting a linear programming problem with infinite variables $P(\mathbf{r})$.Thus, the above method is limited to categorical endogenous variables and cannot directly extend to the continuous domain. We tackle this challenge by adopting the generative adversarial learning framework and representing candidate causal models in a parameter space such that it can be searched using state-of-the-art optimization algorithms. Consider to estimate response functions from PA_V to V by neural networks with a certain network structure. This means that we can define the network parameters as the response-function variables for summarizing all possible mappings from PA_V, U_V to V that could be estimated by a fixed-architecture neural network. We then use the implicit generative model to generate the distribution for the response-function variable such that the domain of the distribution is represented as the parameter space of the generative model. We parameterize the causal model by expressing it with response-function variables.Finally, the parameterzied causal model is used to formulate an adversarial learning problem for computing the bounds of the ACE.

In the following, we explain our method in details.

Parameterizing Causal Models

Specifically, for each endogenous variable V, a neural network $v = h_V(\mathsf{pa}_V; \theta_V)$ with input pa_V and parameters $\theta_V \in \Theta_V$ is used as a universal estimator of response functions from PA_V to V, i.e., we treat Θ_V as the response variable. Thus, the domain of U_V is estimated by parameter space Θ_V of the neural network, and distribution $P(u_V)$ is correspondingly represented by the distribution over all parameters values, i.e., $P(\theta_V)$. Then, to traverse distributions over U_V is simulated by traversing distributions over Θ_V. As a special case, if $\mathsf{PA}_V = \emptyset$, then we directly let $v = \theta_V$ to represent a trivial mapping.

To generate different distributions for θ_V, we adopt the implicit generative model (Mohamed and Lakshminarayanan 2016). An implicit generative model defines a stochastic procedure to generate data by transforming some random noise to the data via some deterministic function. In our problem, random noise $\mathbf{z}_V$ is taken as input and transformed into θ_V via a neural network $G_V(\mathbf{z}_V)$. By convention, we refer to this neural network as a generator. Since θ_V represents the response function for computing v, with $G_V(\mathbf{z}_V)$ we in fact define an implicit generative model for generating v from $\mathbf{z}_V$, i.e., $v = h_V(\mathsf{pa}_V; G_V(\mathbf{z}_V))$. As a result, by updating the generator, since parameter space Θ_V approximately represents the domain of U_V, we can explicitly traverse possible distributions over U_V no matter what dimensions and domains it has.

Based on above discussions, to parameterize a causal model, we define an implicit generative model for each $V \in \mathbf{V}$ as follows.

Definition 3. *For a causal model $\forall V \in \mathbf{V}, v = f_V(\mathsf{pa}_V, u_V)$, its parameterized version is given by*

$$\forall V \in \mathbf{V}, v = h_V(\mathsf{pa}_V; G_V(\mathbf{z}_V))$$

where generators $G_V(\mathbf{z}_V)$ contain parameters that are to be learned from data.

For simplicity, we denote the parameterized causal model as $\mathbf{v} = G(\mathbf{z})$ where $\mathbf{z}$ denotes the set of noise terms for all endogenous variables.

Encoding independence assumptions. It is worth noting that, since Θ_V is a representation of U_V, it should inherit the independence relationship between U_V and other exogenous variables. That is to say, Θ_{V_1} and Θ_{V_2} should be (in)dependent if U_{V_1} and U_{V_2} are known to be (in)dependent. We address this issue by using the same random noise for generators G_{V_1} and G_{V_2} if U_{V_1} and U_{V_2} are dependent. More formally, as shown in (Tian and Pearl 2002), any causal graph can be decomposed into a number of disjoint components, called c-components, such that any pair of exogenous variables are correlated if they belong to the same component and independent if they belong to different components. C-component factorization (Tian and Pearl 2002) provides a way to group all correlated variables together. For example, in Figure 1, A, B belong to one c-component and C belongs to another c-component. In our method, we adopt the c-component factorization to decompose the causal graph into c-components. Then, we share the same noise term among generators corresponding to variables within each c-component.

Optimizing ACEs

Our objective is to find causal models that bound the ACE of interest. Given a parameterized causal model defined in Definition 3, computing the ACE from it is straightforward. For any intervention $do(a')$, we directly perform it to modify the parameterized causal model as:

$$a = a';\ \forall V \neq A, v = h_V(\mathsf{pa}_V; G_V(\mathbf{z}_V)).$$

Then, we estimate the value of an ACE of A on B by sampling B from the intervened parameterized causal model. We denote it as $\mathrm{ACE}(G;a_1,a_0)$ since the estimated value depends on all generators. As a result, we can learn the lower bound by minimizing $\mathrm{ACE}(G;a_1,a_0)$, and learn the upper bound by maximizing $\mathrm{ACE}(G;a_1,a_0)$ or minimizing $-\mathrm{ACE}(G;a_1,a_0)$.

Meanwhile, we want the causal models searched in above learning process to be confined to those agree with a given observational distribution $P(\mathbf{v})$. The parameterized causal model allows us to compare the generated distribution with the observational distribution, and use the discrepancy to update generators. Similar to generative adversarial networks (GAN) (Goodfellow et al. 2014), we make use of a discriminator D to distinguish observational data from that generated by the parameterized causal model. We train the discriminator to maximize its probability of assigning correct labels, which is then used as a measure of the discrepancy between the generated distribution and the observational distribution. Symbolically, it is given by $\max_D V(G,D)$ where

$$V(G,D){=}\mathbb{E}_{\mathbf{v}\sim P(\mathbf{v})}[\log D(\mathbf{v})]{+}\mathbb{E}_{\mathbf{z}\sim P(\mathbf{z})}[\log(1{-}D(G(\mathbf{z})))].$$

When the generated distribution is equivalent to the observational distribution, we reach the theoretical minimal value of $\max_D V(G,D)$. We denote this value by m, which varies for different versions of the GAN model (e.g., it is equal to $-\log 4$ in the vanilla GAN (Goodfellow et al. 2014)).

Combining above two partial objectives, to obtain the lower bound (similarly for the upper bound), we would like to learn generators G that minimize $\mathrm{ACE}(G;a_1,a_0)$ subject to that $\max_D V(G,D) \leq m{+}\eta$. Here η is a threshold which specifies how close we want the generated distribution is to the observational distribution. Ideally η should be set to 0 but in practice we can set it to a small value to allow some room for imperfect fitting. By introducing the Lagrange multiplier λ, this constrained optimization problem can be converted to an unconstrained optimization problem, given by

$$\min_G \max_{\lambda\geq 0} \left\{\mathrm{ACE}(G;a_1,a_0) + \lambda\left(\max_D V(G,D) - m - \eta\right)\right\}.$$

Since the first term is irrelevant to D, we can pull $\max_D$ out of the sum. Finally, the optimization problem is defined as:

Problem 1. *Given a causal graph and the data, the lower bound (similarly for the upper bound) of the ACE of A on B is computed by solving the optimization*

$$\min_G \max_{\lambda\geq 0} \max_D \left\{\mathrm{ACE}(G;a_1,a_0) + \lambda\left(V(G,D) - m - \eta\right)\right\}.$$

The pseudocode of above procedure is given in Algorithm 1 in the supplementary file.

The training procedure is as follows. We continually sample mini-batches of noise samples $\mathbf{z}$. For each noise sample, we compute the expressions of B prior and post to the intervention $do(a)$ based on the parameterized causal model. The average of the post-intervention expressions of B is used to compute $\mathrm{ACE}(G;a_1,a_0)$, and the average of the prior-intervention expressions of B is used to compute $V(G,D)$. Then, we recurrently update the discriminator, λ, and generators based on the gradients of the objective function.

Example

We use the toy example shown in Figure 1 to demonstrate how to formulate the optimization problem given a causal graph. In general, the causal model of this example consists of three structural equations: $c = f_C(\mathbf{u}_C)$, $a = f_A(c, \mathbf{u}_A)$, $b = f_B(a, \mathbf{u}_B)$. We cannot observe the hidden confounder U, but assume that we know A and B are confounded. Thus, we define three neural networks (generators): G_C, G_A, and G_B. By conducting the c-composition factorization, we obtain two c-components: $\{C\}$ and $\{A, B\}$. So, generator G_C uses one noise term $\mathbf{z}_1$ and generators G_A and G_B share another noise term $\mathbf{z}_2$. To construct the parameterized causal model, since C has no parent, $G_C(\mathbf{z}_1)$ directly generates c to represent a trivial mapping. On the other hand, a is generated by the response function, a neural network $h_V(c; G_A(\mathbf{z}_2))$ with c as the input and $G_A(\mathbf{z}_2)$ as the parameters. Similarly, b is generated by the neural network $h_V(a; G_B(\mathbf{z}_2))$ with a as the input and $G_B(\mathbf{z}_2)$ as the parameters. As a result, the parameterized causal model is given by

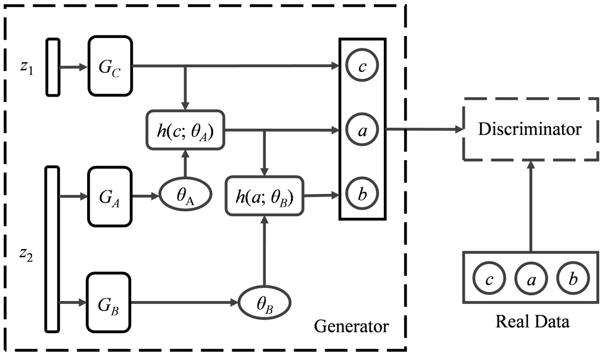

Figure 2: The network architecture with three generators G_A, G_B, G_C for the causal graph shown in Figure 1. A discriminator is used to measure the difference between the generated data and the real data.

$$\begin{array}{l} c = f_C(\mathbf{u}_C) \\ a = f_A(c, \mathbf{u}_A) \\ b = f_B(a, \mathbf{u}_B) \end{array} \xRightarrow{\text{parameterized}} \begin{array}{l} c = G_C(\mathbf{z}_1) \\ a = h_V(c; G_A(\mathbf{z}_2)) \\ b = h_V(a; G_B(\mathbf{z}_2)) \end{array}$$

The model structure is shown in Figure 2, and the ACE of A on B is given by

$$\begin{aligned}\mathrm{ACE}(G;a_1,a_0) = &\mathbb{E}_{\mathbf{z_2}\sim P(\mathbf{z_2})}\left[h_V(a_1; G_B(\mathbf{z_2}))\right] \\ &- \mathbb{E}_{\mathbf{z_2}\sim P(\mathbf{z_2})}\left[h_V(a_0; G_B(\mathbf{z_2}))\right].\end{aligned}$$

Implementation Considerations

In practice, the generative adversarial learning does not guarantee to converge to the global optimum (Goodfellow et al. 2014), which means that we may not be able to find the optimal solution to Problem 1. However, the proposed bounding method is still meaningful even if the optimal solution cannot be obtained. Recall that the fundamental of the bounding method is to search for all the causal models that agree with the observational distribution and find the maximal/minimal ACE among them. Thus, we can treat the optimizing process as a constructive way of estimating the ACE. That is to say, after the discriminator loss $V(G,D)$ becomes stable, each intermediate solution provides a feasible estimation of the ACE, which can be used to construct the bounds. Formally, we examine the intermediate solutions where constraint $\max_D V(G,D) \leq m + \eta$ is satisfied. Then, we take multiple intermediate solutions to compute the mean and variance, and use the one sided confidence interval to estimate the bound. This estimation is meaningful as the causal models that agree with the observational distribution rarely fall outside the interval. Experiments show that this approach can provide good estimations to the ACE.

Our method relies on the input of the causal graph. Many algorithms have been proposed to discovery the causal structure with possible hidden confounding, such as the well-known Fast Causal Inference (FCI) algorithm (Spirtes, Meek, and Richardson 1995) and its extension tsFCI (Entner and Hoyer 2010). We can adopt these algorithms for learning the causal graph from data in practice.

Linear Causal Models: A Special Case

In some situations we may assume that the mapping from PA_V to V for each variable V is a certain type of equation. For example, linear causal models assume that all structural equations in the model are linear. In this case, we can adopt a simplified generative model to encode this assumption. Specifically, for each variable V, we define the response function as the inner product between a parameter vector and the input, i.e., $v = G_V(\mathbf{z}_V) \cdot [\mathsf{pa}_V, 1]^T$. The parameter vector $G_V(\mathbf{z}_V) = [\theta_V^{(1)}, \theta_V^{(2)}, \cdots, \theta_V^{(|\mathsf{pa}_V|)}, g_V(\mathbf{z}_V)]$ consists of two part: one is a set of parameters $\theta_V^{(1)}, \theta_V^{(2)}, \cdots$ that are irrelevant to the noise, and the other is a neural network $g_V(\mathbf{z}_V)$ that maps the noise to a single variable. Upon the definition of the response function, the joint distribution and the ACE could be expressed, and the optimization problem could be formulated and solved similarly to Problem 1.

Encoding the equation assumption can reduce the search space for finding causal models and shrink the bounding range. The following proposition shows that, for linear causal models, if there exists an instrumental variable for ACE of A on B, i.e., a variable that affects A and affects B only by influencing A, then the ACE estimated from Problem 1 would converge to a fixed value as the generated distribution converges to the observational distribution. This result is consistent to the well-known instrumental variable formula (Bowden and Turkington 1984). Please refer to the supplementary file for the proof.

Proposition 1. *Let C be an instrumental variable for ACE of A on B, then both bounds computed from Problem 1 will converge to $\frac{\mathrm{cov}(B,C)}{\mathrm{cov}(A,C)}(a_1 - a_0)$ if the generated distribution converges to the observational distribution.*

Experiments

We implement and evaluate the proposed bounding method. We use both synthetic data and a real-world dataset, Adult (Dheeru and Karra Taniskidou 2017). In addition to our method, we also evaluate following baseline methods for inferring the ACE. Previous bounding and estimation methods (Wu et al. 2019; Louizos et al. 2017) are not included as they cannot handle continuous or high dimension treatment attributes.

- **Linear/logistic regression**: We build a linear/logistic regression on the outcome using all observed variables, and then compute the ACE based on the coefficient of the treatment variable.
- **Instrumental variable** estimation: We implement this method following the classic instrumental variable formula (Bowden and Turkington 1984).
- **Propensity score** adjusted regression: We adopt the propensity score adjusted regression explained in (Abdia et al. 2017) and follow the method in (Hirano and Imbens 2004) to handle continuous variables.

Experimental Settings In the implementation of our method, we use one hidden layer with 16 nodes for all generators $G_V(\cdot)$ and neural networks $h_V(\cdot;\cdot)$, and use ReLU as the activation function. For the discriminator, we adopt the framework of Wasserstein GAN (Arjovsky, Chintala, and Bottou 2017), which leverages the Wasserstein distance between observational and generated distributions. The benefits of using the Wasserstein GAN are to stabilize the training and provide a meaningful loss metric to indicate properties of convergence. In addition, the adaptive gradient clipping (Belghazi et al. 2018) is also used to stabilize the training. The threshold η in the constraint is set to 0.001, and we take 50 solutions satisfying the constraint to compute the mean and variance. The **upper bound** is computed as mean $+$ std, and the **lower bound** is computed as mean $-$ std.

Synthetic Data

We manually build a causal graph (shown in Figure 3) with 5 continuous endogenous variables, Z, X, W, V, Y. We assume that the exogenous variables associated with X and V are confounded by a hidden variable U_1; the exogenous variables associated W and V are confounded by another hidden variable U_2; and other pairs of exogenous variables are independent. To completely specify the causal model, all exogenous variables are Gaussian or mixed-Gaussian noises with zero mean and variance$= 0.1$.

To illustrate the capability of coping with nonlinear causal model, we design the causal model as follows.

$$u_1 = \epsilon_1, \quad u_2 = \epsilon_2, \quad z = \text{Uniform}(\theta_1^z, \theta_2^z) + \epsilon_z$$
$$x = \theta_0^x + \theta_1^x z + \theta_2^x u_1 + \epsilon_x,\ w = \theta_0^w + \theta_1^w x^2 + \theta_2^w u_2 + \epsilon_w$$
$$v = \theta_0^v + \theta_1^v u_1 + \theta_2^v u_2 + \epsilon_v,\ y = \theta_0^y + \theta_1^y w + \theta_2^y v + \epsilon_y$$

where $\epsilon_1, \epsilon_2, \epsilon_z, \epsilon_x, \epsilon_w, \epsilon_v, \epsilon_y$ are independent Gaussian noises. After specifying the causal model, we generate 10,000 examples *without* hidden variables U_1, U_2 which are then used by all methods for estimating the ACE. The ground truth of the ACE is directly obtained by performing the intervention on the causal model. For instrumental variable estimation, Z is treated as the instrumental variable.

Results The experimental results are shown in Figure 4. ACEs are obtained by performing different interventions, i.e., $\mathbb{E}[Y|do(x_1)] - \mathbb{E}[Y|do(x_0)]$ with different x_1 and x_0. For demonstration, we select five interventions and report the corresponding ACEs. We see that our upper bound and lower bound cover the ground truth in all interventions. However, other baseline methods cannot produce accurate estimations and fall outside the bounds in most cases. This

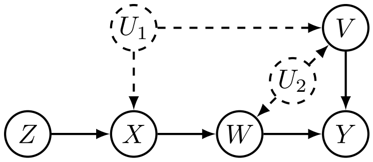

Figure 3: The causal graph of synthetic data: Z, X, W, V, Y are observed variables and U_1, U_2 are hidden variables.

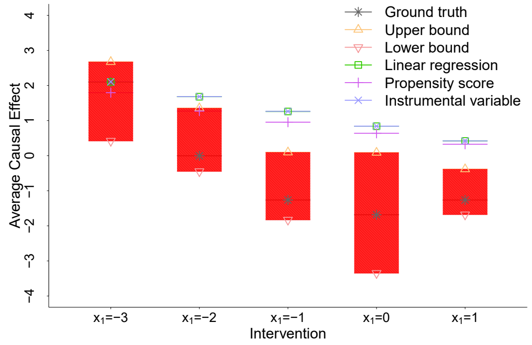

Figure 4: Average causal effects with different interventions ($x_0 = 2$) on the nonlinear synthetic dataset.

may be due to the fact that their assumptions (e.g., strong ignorability for propensity score, and linear assumption for linear regression and instrumental variable formula) do not hold in our general setting. Especially, note that when $x_1 = 1$, our bounds can help estimate the correct sign of the ACE, while other methods produce opposite estimations.

We also evaluate the special linear case as described in the previous section and obtain similar results. Please refer to the supplementary file for details.

Adult Dataset

We further conduct evaluations on the real-world Adult dataset.The Adult dataset consists of 48,842 tuples with 11 attributes. We make use of the causal graph in (Zhang, Wu, and Wu 2018) shown in Figure 5. With all the attributes observed, we evaluate the ACE of edu_level (which ranges from 0 to 16 with 0 the lowest level) on income (1 if $> 50K$ and 0 if $\leq 50K$) as the ground truth by using the code from (Xu et al. 2019).

In order to simulate hidden confounders, we deliberately hide 5 attributes (denoted by dotted nodes in Figure 5), and treat the remaining attributes (denoted by solid nodes) as observed to compute the ACE. In this setting, we can easily see that native_country and race are two hidden confounders. Meanwhile, marital_status is also a hidden confounder since it is a confounder between edu_level and occupation while the latter is an intermediate on a causal path from edu_level to income, resulting confounding effects of edu_level on income. As a result, the ACE of edu_level on income is unidentifiable. Our method is implemented using the sub-graph induced by the observed attributes. All baseline methods are computed using observed attributes only. Since there is no instrumental variable for the ACE of edu_level on income, the instrumental variable estimation is not performed.

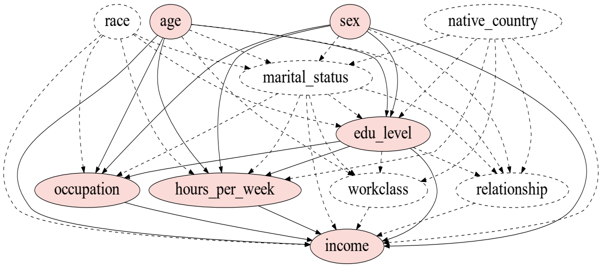

Figure 5: The causal graph of the Adult dataset.

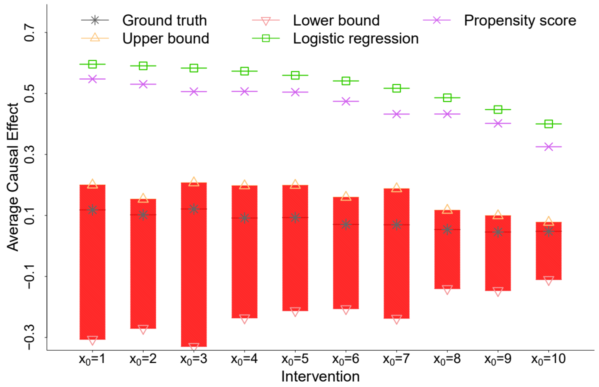

Figure 6: Average causal effects with different interventions ($x_1 = 16$) on the Adult dataset.

Results The experimental results are shown in Figure 6. ACEs are evaluated by performing different interventions $\mathbb{E}[Y|do(X = x_1)] - \mathbb{E}[Y|do(X = x_0)]$, and we report 10 interventions where $x_1 = 16$ and x_0 ranges from 1 to 10. In general, the ground truth decreases from 0.11 to 0.04 as x_0 increases. It falls in the range of the upper bound and lower bound in all interventions, which validate the efficacy of our method. However, other baseline methods including the logistic regression and propensity score cannot produce accurate estimations and fall outside the bounds in all cases.

Conclusions

We proposed a bounding method for estimating average causal effects (ACEs) from observational data with hidden confounding. The method parameterizes the causal model using implicit generative models, and builds an adversarial network to formulate a constrained optimization problem for computing the bounds. We showed that encoding the linear assumption can make the bounds converge to a fixed value. Experiments using both synthetic and real-world datasets showed that our method provides more accurate estimations than several widely used causal inference methods.

Reproducibility. The source code is available at https://github.com/yaoweihu/Bound-Confounded-Causal-Effects.

Acknowledgments

This work was supported in part by NSF 1646654, 1920920, and 1946391.

Ethics Statement

Causal inference is a fundamental technique for inferring causal effects among variables. In past years, it has been widely adopted by computer scientists for addressing ethical issues in artificial intelligence such as fair machine learning. These works usually rely on analyzing causal effects of sensitive attributes such as gender or race on decisions such as hiring or admission. Such analysis can be performed either on the training data so that the data could modified before it is used for training, or on model outputs so that it could be used to regulate the learning procedure. However, when hidden confounders exist, how to estimating causal effects is a big challenge. Failing to correctly estimate causal effects may undermine these causal-based fair machine learning techniques. The contribution of our paper is that it provides a reliable approach for bounding the causal effects under the existence of hidden confounders. It can be potentially used to assess fairness of data or machine learning models and facilitate the research in fair machine learning.

References

Abdia, Y.; Kulasekera, K.; Datta, S.; Boakye, M.; and Kong, M. 2017. Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal* 59(5): 967–985.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* .

Avin, C.; Shpitser, I.; and Pearl, J. 2005. Identifiability of path-specific effects. In *Proceedings of IJCAI'05*, 357–363.

Balke, A.; and Pearl, J. 1997. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92(439): 1171–1176.

Belghazi, M. I.; Baratin, A.; Rajeswar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, R. D. 2018. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062* .

Bowden, R.; and Turkington, D. 1984. *Instrumental Variables*. Cambridge, UK: Cambridge University Press.

Chiappa, S. 2019. Path-specific counterfactual fairness. In *Proceedings of AAAI'19*, volume 33, 7801–7808.

Dheeru, D.; and Karra Taniskidou, E. 2017. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences.

Entner, D.; and Hoyer, P. O. 2010. On causal discovery from time series data using FCI. *Probabilistic graphical models* 121–128.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in NeurIPS'14*, 2672–2680.

Hirano, K.; and Imbens, G. W. 2004. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164: 73–84.

Hu, H.; and Kerschberg, L. 2018. Evolving Medical Ontologies Based on Causal Inference. In *2018 IEEE/ACM ASONAM'18*, 954–957. IEEE.

Kallus, N.; Mao, X.; and Zhou, A. 2018. Interval estimation of individual-level causal effects under unobserved confounding. *arXiv preprint arXiv:1810.02894* .

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2018. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *International Conference on Learning Representations*.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in NeurIPS'17*, 4066–4076.

Li, Y.; Kuang, K.; Li, B.; Cui, P.; Tao, J.; Yang, H.; and Wu, F. 2020. Continuous Treatment Effect Estimation via Generative Adversarial De-confounding. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, 4–22. PMLR.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. In *Advances in NeurIPS'17*, 6446–6456.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2019. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of FAT'19*, 349–358. ACM.

Miles, C. H.; Kanki, P.; Meloni, S.; and Tchetgen, E. J. T. 2015. On partial identification of the pure direct effect. *arXiv preprint arXiv:1509.01652* .

Mohamed, S.; and Lakshminarayanan, B. 2016. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483* .

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

Pearl, J. 2010. An introduction to causal inference. *The international journal of biostatistics* 6(2).

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

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3076–3085. JMLR. org.

Shpitser, I.; and Pearl, J. 2008. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research* 9(Sep): 1941–1979.

Spirtes, P.; Meek, C.; and Richardson, T. 1995. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, 499–506.

Tian, J.; and Pearl, J. 2000. Probabilities of causation: Bounds and identification. In *Proceedings of UAI'00*, 589–598. Morgan Kaufmann Publishers Inc.

Tian, J.; and Pearl, J. 2002. A general identification condition for causal effects. In *AAAI/IAAI*, 567–573.

Wang, M. D.; and Wu, H. 2019. Tutorial: Causal Inference in Biomedical Data Analytics–Basics and Recent Advances. In *Proceedings of ACM-BCB'19*, 558–558.

Wu, Y.; Zhang, L.; and Wu, X. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of IJCAI'19*, 10–16.

Wu, Y.; Zhang, L.; Wu, X.; and Tong, H. 2019. PC-Fairness: A Unified Framework for Measuring Causality-based Fairness. In *Advances in NeurIPS'19*, 3399–3409.

Xu, D.; Wu, Y.; Yuan, S.; Zhang, L.; and Wu, X. 2019. Achieving causal fairness through generative adversarial networks. In *Proceedings of IJCAI'19*.

Yoon, J.; Jordon, J.; and Van Der Schaar, M. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Zhang, J.; and Bareinboim, E. 2018. Fairness in decision-making–the causal explanation formula. In *32nd AAAI Conference on Artificial Intelligence*.

Zhang, L.; Wu, Y.; and Wu, X. 2018. Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. *IEEE Transactions on Knowledge and Data Engineering* 31(11): 2035–2050.

Zhao, J.; Runfola, D. M.; and Kemper, P. 2017. Simulation study in quantifying heterogeneous causal effects. In *Winter Simulation Conference*, 2650–2661.