Approximate Causal Effect Identification under Weak Confounding

Ziwei Jiang [1] **Lai Wei** [1] **Murat Kocaoglu** [1]

Abstract

Causal effect estimation has been studied by many researchers when only observational data is available. Sound and complete algorithms have been developed for pointwise estimation of identifiable causal queries. For non-identifiable causal queries, researchers developed polynomial programs to estimate tight bounds on causal effect. However, these are computationally difficult to optimize for variables with large support sizes. In this paper, we analyze the effect of "weak confounding" on causal estimands. More specifically, under the assumption that the unobserved confounders that render a query non-identifiable have small entropy, we propose an efficient linear program to derive the upper and lower bounds of the causal effect. We show that our bounds are consistent in the sense that as the entropy of unobserved confounders goes to zero, the gap between the upper and lower bound vanishes. Finally, we conduct synthetic and real data simulations to compare our bounds with the bounds obtained by the existing work that cannot incorporate such entropy constraints and show that our bounds are tighter for the setting with weak confounders.

1. Introduction

Estimating the causal effect has long been a question of great interest in a wide range of fields, such as marketing (Jung et al., 2022), healthcare (Lv et al., 2021; Meilia et al., 2020), social science (Freedman, 2010), and machine learning (Pearl, 2019). The causal relation differs from the statistical association due to the existence of unobserved confounders, variables that affect both the treatment and outcome, which create a spurious association, causing the statistical association to deviate from the true causal effect. An example study of the causal relationship between weekly exercise and cholesterol in various age groups is discussed by Glymour et al.(2016). In this study, the cholesterol level is negatively correlated with the amount of weekly exercise within each age group. But if the age data is not observed, the cholesterol level appears positively correlated with the amount of exercise. This is known as Simpson's paradox, where the confounding variable of age causes the sign reversal. If the variable age is not observed in the data, the true causal effect of exercise on cholesterol is not identifiable. Numerous studies have addressed this problem in different settings (Rosenbaum & Rubin, 1983; Pratt & Schlaifer, 1988; Pearl, 2022).

It is well known that the causal effect can be estimated from observational data if we could control for the confounders, which means the confounders are included in the observational data (Lindley & Novick, 1981; Rubin, 1974; Pearl, 1995). Tian and Pearl (2002) provide conditions for the identifiability of causal queries.

One approach for addressing non-identifiable causal queries involves making additional untestable assumptions either on the variables or on the parametric form of the model. For example, with linear model assumption, instrumental variables can be used to estimate the Average Treatment Effect (ATE) even if unobserved confounders exist (Bowden & Turkington, 1990). However, these approaches are limited to specific settings due to the restriction of variables or the expressive power of the parametric model assumptions. Many recent studies have focused on alleviating this constraint by using machine learning models as the function for the instrumental variable (Singh et al., 2019; Xu et al., 2020).

The other type of approach makes no additional assumptions but attempts to obtain bounds for the causal effect instead of point identification. This is also known as partial identification. With instrumental variables, Kilbertus et al. (2020) studied bounds on ATE without the assumption that unobserved confounders affect variables additively. Padh et al.(2022) extend this idea to continuous treatments. Zhang and Bareinboim (2021) developed a method to bound causal effects on continuous outcomes. Recent works using generative neural networks for partial identification in high dimensional continuous settings (Hu et al., 2021; Balazadeh Meresht et al., 2022).

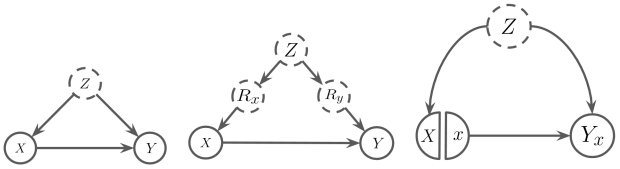
[1]Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. Correspondence to: Ziwei Jiang <jiang622@purdue.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

The simplest non-identifiable setting is when two observed variables, i.e., treatment and the outcome, are confounded by some latent variables. In this paper, we focus on this setting. Tian and Pearl (2000) developed tight bounds in the non-parametric setting using the observational distribution as a constraint. Li and Pearl (2022) derived bounds with nonlinear programming with partially observed confounders, i.e., only the prior distribution of the confounder is known.

A key challenge in causal inference is determining the strength of the confounder, which refers to the degree to which the confounder is associated with the treatment and the outcome. The stronger the association, the more likely it is that the confounder is biasing the estimate of the effect of the exposure on the outcome. Sensitivity analysis is commonly used, especially for parametric models (Cinelli et al., 2019). Many existing studies used information theoretic quantities such as directed information (Etesami & Kiyavash, 2014; Quinn et al., 2015) and relative entropy (Janzing et al., 2013) as measurements of the edge strength. Researchers have used entropy to discover the causal direction in the graphs (Kocaoglu et al., 2017; Compton et al., 2020). Janzing and Schölkopf (2010) developed a theory for causal inference based on the algorithmic independence of the Markov kernels. Vreeken and Budhathoki (2015; 2018) extend this idea by using minimum description length for causal discovery. Another common usage of information theory in causal inference is quantifying the causal influence of variables. Ay and Polani (2008) defined information flow to measure the strength of causal effect based on the causal independence of the variables. Similar to relative entropy or mutual information, the information flow measures the independence between a set of nodes B and A after intervening on another set S. Janzing et al. (2013) studied the causal influence by quantifying the changes in the distribution of a single variable response to the intervention. Geiger et al. (2014) used these measures to formulate the bounds of the confounding variables by showing that the back-door dependence should be greater or equal to the deviation between observed dependence and causal effect with some measure.

We are interested in the problem of estimating causal effect when confounders are "simple," i.e., the entropy of the confounder is small. The information passing through such confounders should not be arbitrarily large, so we should get tighter bounds on the causal effect compared to the methods that cannot utilize this side information. However, it is nontrivial to incorporate low-entropy constraints since entropy is a concave function. Enforcing small entropy as a constraint directly changes the feasible set to a non-convex set. Therefore, the problem cannot be solved directly using the existing formulations. In this paper, we address this problem by quantifying the tradeoff between the strength of the unobserved confounder measured by its entropy and the upper and lower bounds on causal effect.



(a) DAG $\mathcal{G}$ with latent confounder (b) Canonical partition of the DAG $\mathcal{G}$ (c) Single world intervention graph of DAG $\mathcal{G}$

Figure 1. A graph consist of treatment X, outcome Y and an unobserved confounder Z with small entropy.

The main contributions of this paper are as follows:

- We formulate a novel optimization problem to efficiently estimate the bounds of causal effect using counterfactual probabilities and apply the low-entropy confounder constraint using this formulation.
- We examine the conditions on the entropy constraint for the optimization to yield a tighter bound. We analytically show the condition when either or both treatment and outcome are binary variables.
- We conducted experiments using both simulated and real-world data to test our method and demonstrate that our bound is tighter than existing approaches that are unable to incorporate entropy constraints.

2. Background and Notation

Notations. Throughout the paper, we use uppercase letters X, Y, Z to denote the random variables and lowercase letters x_i, y_i, z_i for their states. We use $\{x, x'\}$ to denote the states of binary variables. The Greek letters α, β, θ are used to denote some constant value for the probability mass function or information-theoretic quantities. $|X|$ represents the number of states for a random variable. The uppercase letter with a lowercase letter as the subscript shows an intervened variable, i.e., $P(Y_x = y) := P(y|do(x))$. This notation is also used for counterfactual distributions, e.g., $P(Y_x = y|X = x')$ means the probability of y had we intervened on x given that x' is observed. For a probability mass function $P(Y = y, X = x)$, we write $P(y, x)$ as an abbreviation. For counterfactual distribution $P(Y_x = y, x')$, we keep the notation of a random variable to avoid confusion.

Entropy and mutual information. In this paper, we use the term entropy to refer to the Shannon entropy, which quantifies the average amount of information in the variable. For a discrete random variable X, its Shannon entropy is defined as follows

$$H(X) = \sum_i -P(x_i) \log P(x_i).$$

Mutual information is a concept closely related to entropy. It measures the average amount of information one variable carries about another. For discrete variables, X, Y, the mutual information is defined as

$$I(X;Y) = \sum_i \sum_j P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}.$$

Causal DAG. A directed acyclic graph (DAG) encodes the causal relationship between variables, where the nodes represent variables, and the edges represent their causal relationships. Causal graphs are often used to help identify and understand complex systems. A graphical condition called d-separation can be used to read-off the independences induced by a graph. Pearl (1995) introduced a set of rules called do-calculus for deriving causal queries from observational data.

Structural Causal Model. Pearl (1995) introduced the Structural Causal Model (SCM), a mathematical framework that can be used to study causality and counterfactuals (Pearl, 2009; Zhang et al., 2022). We can describe causal relationships between variables with a set of functions in SCM. More specifically, an SCM is a tuple $\{\mathcal{U}, \mathcal{V}, \mathcal{F}, \mathbb{P}\}$ where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{F}$ is a set of functions with same cardinality of $\mathcal{V}$, and $\mathbb{P}$ is a product probability measure. Each $v \in \mathcal{V}$ is generated by some $f \in \mathcal{F}$ as a function of other variables. The functions in an SCM impose a causal graph.

Canonical Partition. A widely used method for bounding the causal effects is canonical partition. Consider an SCM with the graph in Figure 1a. The binary variables X, Y are generated by the functions $y = f_y(x, u_y)$ and $x = f_x(u_x)$. The latent variables U_x and U_y are not independent since there is latent confounding between X and Y. Balke and Pearl (1997) pointed out that the latent variable U_y can be replaced by a finite-state response variable R_y representing the distinct functions mapping X to Y. All these functions can be represented by a response variable R_y with four states.

Table 1. Response variable R_y

	$X = x_0$	$X = x_1$
$R_y = 0$	$Y = y_0$	$Y = y_0$
$R_y = 1$	$Y = y_0$	$Y = y_1$
$R_y = 2$	$Y = y_1$	$Y = y_0$
$R_y = 3$	$Y = y_1$	$Y = y_1$

Each row corresponds to a function that maps X to Y. For example if $R_y = 1$, we have

$$Y = f_Y(X, R_y = 1) = \begin{cases} y_0 & \text{if } X = x_0 \\ y_1 & \text{if } X = x_1. \end{cases}$$

And for X we have $R_x \in \{1, 2\}$

$$R_x = 0 : X = x_0,$$
$$R_x = 1 : X = x_1.$$

Due to the latent confounder, R_x and R_y are dependent as shown in Figure 1b. The joint distribution $P(R_x, R_y)$ has total of 8 states, denoted by $q_{ij} = P(R_x = i, R_y = j)$, and $p_{ij} = P(x_i, y_j)$. The observational probability can be expressed as $p_{00} = q_{00} + q_{01}$, $p_{01} = q_{02} + q_{03}$, $p_{10} = q_{10} + q_{12}$, and $p_{11} = q_{11} + q_{13}$.

The causal effect $P(y_0|do(x_0))$ is the probability of the function which maps x_0 to y_0. Therefore we have $P(y_0|do(x_0)) = q_{00} + q_{01} + q_{10} + q_{11}$. Combining the above equations, we obtain the bounds of causal effect in a closed-form expression: $p_{00} \leq p(y_0|do(x_0)) \leq 1 - p_{01}$. This method has been used in causal inference problems. Tian and Pearl (2000) apply this to estimate the bounds for the probability of causation given the interventional data. Zhang and Bareinboim (2017) use this method to derive bounds for the multi-arm bandit problem.

This bound holds true for any pair of variables, so it can not incorporate any side information, such as the graph structure or the prior distribution of the unobserved confounder, and might be loose in some cases. For example, consider the distribution $P(X, Y)$ with binary X and Y and a low entropy confounder. Suppose both $P(x, y)$ and $P(x, y')$ are small, and we know that the entropy of the confounder is small, i.e., upper bounded by some small value θ. Intuitively, the causal effect $p(y|do(x))$ should be close to $P(y|x)$. This is validated with experiment in Section 5.1. Without incorporating this information about confounder, the bounds are not very informative since the bounds are close to 0 and 1:

$$0 \approx P(x, y) \leq P(y|do(x)) \leq 1 - P(x, y') \approx 1$$

In Section 3.1, we form an optimization problem that can incorporate such low entropy constraints.

Counterfactual and Single-World Intervention Graph (SWIG). Counterfactual queries are questions of the form *"What would happen if an intervention or action had been taken differently, given what already has happened."* Pearl (2009) introduced counterfactual reasoning with the SCM. A counterfactual query $P(Y_x = y|x')$ reads, "The probability of y had we intervened on x given x' is observed." In general, given an SCM, the counterfactual queries can be estimated with three steps: "abduction," "action," and "prediction."

The first step is to use the observed x' as evidence to update the exogenous variables U. The second step is to apply the intervention by replacing the value in the SCM with x. And lastly, make predictions with the updated SCM.

Richardson and Robins (2013) introduced a graphical representation to link the counterfactual distribution and DAG, called Single World intervention graphs (SWIGs). We can represent the interventional variable Y_x as a node in the DAG and split the treatment variable into nodes X and $X = x$. As shown in Figure 1c, we have Y_x independent from X given Z.

3. Bounding Causal Effect with Entropy Constraint

3.1. Bounds with Canonical Partition

Consider the DAG in Figure 1a, the latent factors can be represented by a joint distribution $P(R_x, R_y)$ with $|R_x| = 2, |R_y| = 4$. The canonical partition representation parameterizes the exogenous variables with R_x, R_y. Therefore the entropy of $H(R_y, R_x)$ does not necessarily reflect the confoundedness of X, Y. For example, we could have a small entropy confounder and Y with large exogenous entropy. In that case, $H(Z) \ll H(R_y)$. Since the unobserved confounders do not appear in the canonical partition representation, applying the entropy constraint for estimating bounds is not straightforward.

To overcome this difficulty, we notice that the causal effect $P(y|do(x))$ is equal to $P(y, x) + \alpha P(y, x') + \beta P(y', x')$ for some unknown α, β. Intuitively, these two parameters can be thought of as the proportion of $p(y, x')$ and $p(y', x)$ that is generated by the function that maps x to y, i.e., from $R_y = \{0, 1\}$. The causal effect attains the Tian-Pearl lower bound if $\alpha = \beta = 0$. In that case, $P(R_x = 1, R_y = 0) = P(R_x = 1, R_y = 1) = 0$. This imposes a constraint on the minimum value of mutual information to attain such distribution. Since the R_x, R_y are conditionally independent of Z, the mutual information $I(R_x; R_y)$ is bounded by the entropy of Z. Under the assumption that the confounder Z is simple, i.e., $H(Z) \leq \theta$. We can apply the entropy constraint to estimate the bounds of the causal effect.

Table 2. Table for the counterfactual distribution

		$P(x)$				
		x_0	...	x_q	...	x_n
$P(Y_{x_q}\|X)$	y_0	b_{00}	...	$P(y_0\|x_q)$	...	b_{0n}
	⋮	⋮	⋱	⋮	⋱	⋮
	y_p	b_{p0}	...	$P(y_p\|x_q)$	...	b_{pn}
	⋮	⋮	⋱	⋮	⋱	⋮
	y_m	b_{m0}	...	$P(y_m\|x_q)$	...	b_{mn}

Theorem 3.1. *Let (X, Y) be the pair of variables in the causal graph in Figure 1a with the joint distribution $P(X, Y)$. Suppose $|X| = n, |Y| = m$. Assuming X and Y are confounded by a set of small entropy unobserved variables Z, i.e., $H(Z) \leq \theta$ for some $\theta \in \mathbb{R}$. The causal effect of x_q on y_p is bounded by $LB \leq P(y_p|do(x_q)) \leq UB$, where*

$$LB/UB = \min / \max \left(\sum_{i=0}^{m^n-1} \sum_{j=0}^{n-1} a_{ij} P(x_j) \right)$$

subject to

$$\sum_{i,j} a_{ij} P(x_j) = 1,$$

$$\sum_{i=0}^{m^{n-1}} a_{iq} P(x_q) = P(y_p, x_q),$$

$$0 \leq a_{ij} \leq 1 \ \forall i, j,$$

$$\sum_{i,j} a_{ij} P(x_j) \log \left(\frac{a_{ij}}{\sum_k a_{ik} P(x_k)} \right) \leq \theta.$$

We formulate the bounds as optimization problems with entropy constraints. a_{ij} is the parameter for the optimization problem. The canonical partition can be naturally generalized to variables in higher dimensions. However, the number of states in the optimization problem quickly becomes intractable with the number of states of the observed variables. For $|X| = n$, $|Y| = m$, the number of possible functions mapping X to Y is m^n, so $|R_y| = m^n$. The total number of parameters is nm^n, which grows exponentially fast as the number of states of X and Y increase. In the next subsection, we present an alternative formulation to estimate bounds.

3.2. Bounds via Counterfactual Probabilities

We propose a new optimization problem using counterfactual probabilities to address the computational challenge in the canonical partition method.

For the causal graph in Figure 1a, the interventional distribution can be represented as $P(Y_x) = P(Y_x, x) + P(Y_x, x')$. By the consistensy property (Robins, 1987), we have $P(Y_x, x) = P(Y, x)$. And by the axiom of probability, $P(y_x, x') \leq P(x')$ for any $y \in Y$.

$$\begin{aligned} P(y, x) &\leq P(Y_x = y) \\ &= P(y, x) + P(y_x, x') \\ &\leq P(y, x) + P(x') \\ &= 1 - P(y', x) \end{aligned} \tag{1}$$

The above derivation shows the bounds from the counterfactual probability are equivalent to Tian-Pearl bounds. Y_x and X are d-separated by the confounder Z, i.e. $Y_x \perp\!\!\!\perp X|Z$.

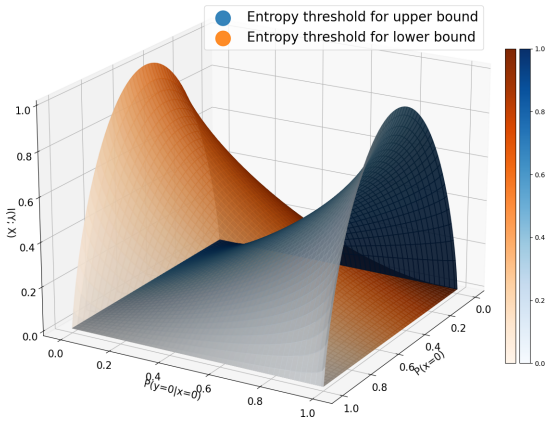


Figure 2. The entropy threshold for obtaining tighter bounds. The thresholds are obtained by sampling $P(x_0)$ and $P(y_0|x_0)$ from 0 to 1 which are the x and y axies in the figure. The orange surface represents the entropy threshold for obtaining a tighter upper bound; the blue surface represents the entropy threshold for obtaining a tighter lower bound. The lightness indicates the gap between the upper and lower bound; the lighter the color, the smaller the gap. Without entropy constraint, the gap depends on the value of $P(x_0)$.

Similar to the argument in Section 3.1, a minimum value of mutual information $I(Y_x; X)$ exists for the causal effect to attain maximum/minimum. By exploiting the d-separation in the SWIG, we can impose the entropy constraint for the optimization problem. We present an optimization problem with entropy constraint based on this method and show that this formulation significantly reduces the number of parameters compared to the canonical partition approach.

Theorem 3.2. *Let (X, Y) be the pair of variables in the causal graph in Figure 1a with the joint distribution $P(X, Y)$. Suppose $|X| = n, |Y| = m$. Assuming X and Y are confounded by a set of small entropy unobserved variables Z, i.e., $H(Z) \leq \theta$ for some $\theta \in \mathbb{R}$. The causal effect of x_q on y_p is bounded by LB $\leq P(y_p|do(x_q)) \leq$ UB, where*

$$LB/UB = \min / \max \left(\sum_j b_{pj} P(x_j) \right)$$

subject to

$$\sum_{i,j} b_{ij} P(x_j) = 1,$$

$$b_{iq} P(x_q) = P(y_i, x_q) \; \forall i,$$

$$0 \leq b_{ij} \leq 1 \; \forall i, j,$$

$$\sum_{i,j} b_{ij} P(x_j) \log \left(\frac{b_{ij}}{\sum_k b_{ik} P(x_k)} \right) \leq \theta.$$

Here b_{ij} are the parameters for the optimization problem. Similar to Section 3.1, we form the causal effect bounds estimation as a maximization and minimization problem. This formulation is more efficient than the canonical partition method in terms of the number of parameters. Consider again the case $|X| = n$, $|Y| = m$; the number of parameters for this optimization problem is nm. This number is significantly smaller than the canonical partition case with nm^n parameters. We will discuss this in more detail in Section 5.4.

4. Condition for Obtaining Tighter Bounds

For Theorem 3.1, the mutual information $I(R_y; X)$ is upper bounded by θ. And for Theorem 3.2, mutual information $I(Y_x; X)$ is upper bounded by θ. In both formulations, the entropy constraint depends on the mutual information between X and another variable. The bounds with entropy constraint will be identical to Tian-Pearl bounds when the upper bound on the confounders entropy is large. We define the greatest value of entropy constraint that yields tighter bounds as the "entropy threshold".

Definition 4.1. Let (X, Y) the pair of variables in the causal graph in Figure 1a. Given an observational distribution $P(X, Y)$ and a causal query $P(y_p|do(x_q))$, the entropy threshold is the greatest entropy constraint such that the bounds obtained from Theorem 3.2 are tighter than the Tian-Pearl bounds.

The entropy threshold depends on the observational distribution $P(X, Y)$. The following lemmas show the entropy threshold when either X and Y are binary variables.

Lemma 4.2. *Let (X, Y) be the pair of binary variables in the causal graph in Figure 1a. Consider $P(Y_x, X)$ for any $x \in X$. Assume, without loss of generality, $P(y|x) \geq P(y'|x)$. Then the following conditions are equivalent:*

1. *$P(Y_x = y)$ attain the Tian-Pearl lower bound,*
2. *$P(Y_x = y')$ attain the Tian-Pearl upper bound,*
3. *$I(Y_x; X)$ is maximized for the given $P(X, Y)$.*

Lemma 4.3. *Let (X, Y) be the pair of variables in the causal graph in Figure 1a, where $|X| = 2$ and $|Y| = m$. The causal effect $P(Y_x = y_p)$ attain the Tian-Pearl upper bound when $P(Y_x = y_p|x') = 1$; attain the Tian-Pearl lower bound with minimum mutual information when $P(Y_x = y_i|x') = \frac{P(Y_x=y_i|X=x)}{\sum_{j \neq p} P(Y=y_j|X=x)}$ for all $i \neq p$.*

Lemma 4.4. *Let (X, Y) be the pair of variables in the causal graph in Figure 1a, where $|Y| = 2$ and $|X| = n$. The causal effect $P(y|do(x_q))$ attain the Tian-Pearl upper bound when $P(Y_{x_q} = y|x_j) = 1, \forall j \neq q$; attain the Tian-Pearl lower bound when $P(Y_{x_q} = y|x_j) = 0, \forall j \neq q$.*

The above lemmas build the link between the bounds of causal effect and the mutual information of counterfactual

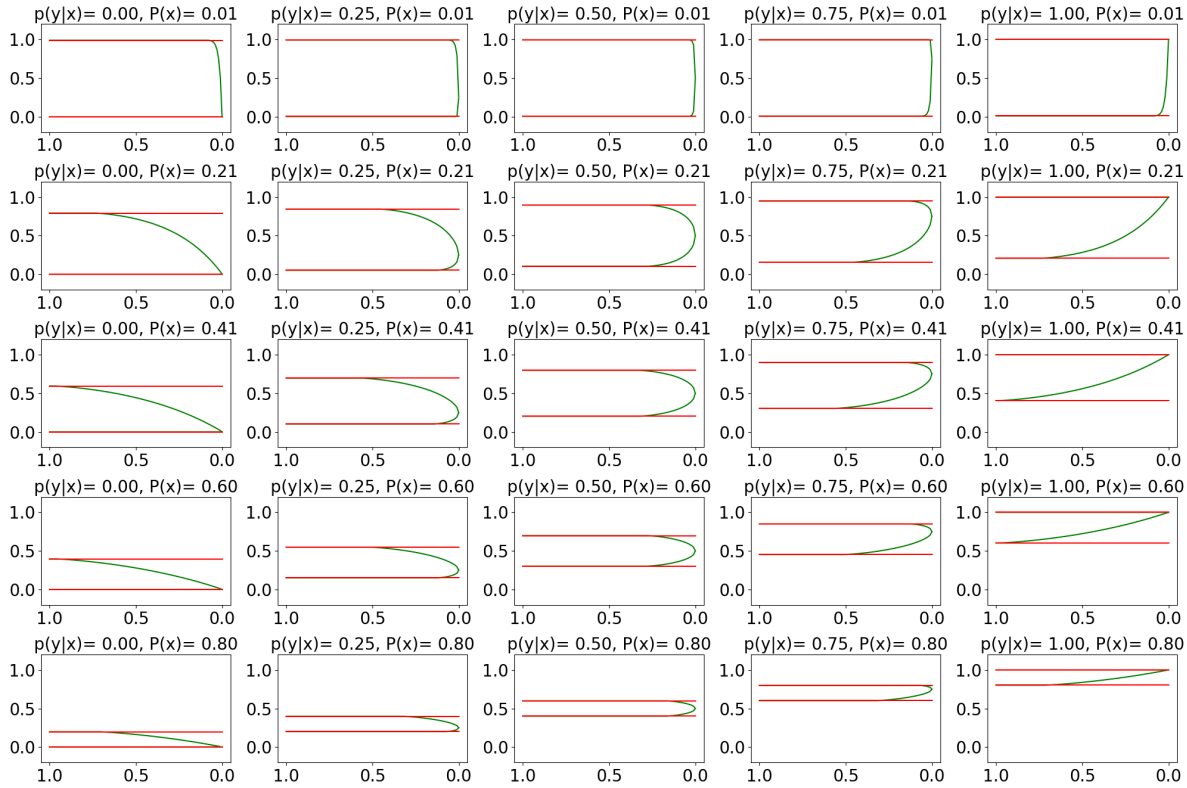


Figure 3. Bounds of the causal effect. The red lines show Tian-Pearl's bounds, and the green lines show our bounds. The x-axis represents the entropy constraint, and the y-axis represents the causal effect $P(y|do(x))$. For each row $P(y|x)$ increases as $P(x)$ is fixed; $P(x)$ increases from top to bottom. The gap between the upper and lower bound decreases monotonically as $P(x)$ increases. The entropy threshold is high when $P(x)$ is close to 0.5 and $P(y|x)$ is close to 1 or 0.

distribution. One can think of b_{ij} as unknown conditional probabilities as shown in Table 2. The q-th column (black) equals the conditional distribution $P(Y|x_q)$, which serves as the constraint from observational distribution. The causal effect is maximized when all entries in p-th row (red) are equal to one; minimized when they are equal to zero.

Next, the following theorem shows the relation between observational distribution $P(X, Y)$ and the entropy threshold.

Theorem 4.5. *Let (X, Y) be a pair of variables in a causal graph G as shown in Figure 1a, where either X or Y is binary. Let (U, V) be two binary variables such that $P(v_0|u_0) = P(y_p|x_q)$, $P(v_1|u_0) = 1 - P(y_p|x_q)$, and $P(u_0) = P(x_q)$. The entropy threshold for the bounds of $P(y_p|do(x_q))$ is equal to $\max(I(U; V))$.*

By Theorem 4.5, we can compute the entropy threshold for a given distribution $P(X, Y)$. Then if we know that the confounder is simple, i.e., with entropy less than the threshold, we can use the entropy constraint to obtain a tighter bound.

Figure 2 shows the entropy threshold for different value of $P(x)$ and $P(y|x)$. The entropy threshold is higher when $P(x)$ is close to 0.5. For fixed $P(x)$, the threshold increases as $P(y|x)$ is close to 0 or 1, which corresponds to the causal effect's lower and upper bound. Without entropy constraint, the gap between bounds is only related to $P(x)$.

Following from Theorem 4.5, the entropy threshold of $P(y_p|do(x_q))$ only depends on the value of $P(x_q)$ and $P(y_p|x_q)$. So we sample $P(x)$ from 0.01 to 0.8 and $P(y|x)$ from 0 to 1. Then let the $p(Y|x')$ be uniform distributions. For each pair of $p(x)$ and $p(y|x)$, we calculate the bounds with entropy constraint for each distribution from 1 to 0. The results in Figure 3 demonstrate the gap between our bounds vanishes as entropy goes to 0. The entropy threshold is small when $P(x)$ is close to 0 or 1 and $P(y|x)$ is close to 0.5. On the other hand, the entropy threshold is high when $P(x)$ is close to 0.5 and $P(y|x)$ is close to 0 or 1. For a fixed conditional probability and entropy constraint, the gap between bounds decreases monotonically with $P(x)$.

5. Experiments

We demonstrated our method with simulated and real-world datasets in this section. First, we show the behavior of the

Table 3. Results of Causal Effect in real-world dataset

DATASET	SUBGROUP	X	Y	H(Z)	OUR BOUNDS	T-P BOUNDS
INSUR	UNDER 5,000 MILES, NORMAL	CAR COST	PROP COST	ACCI		
		100,000	10,000	0.092	[0.000, 0.246]	[0.000, 0.800]
		100,000	100,000	0.092	[0.699, 0.996]	[0.196, 0.996]
		100,000	1,000,000	0.092	[0.004, 0.301]	[0.004, 0.804]
		1,000,000	10,000	0.092	[0.000, 0.044]	[0.000, 0.249]
		1,000,000	100,000	0.092	[0.000, 0.044]	[0.000, 0.249]
		1,000,000	1,000,000	0.092	[0.956, 0.999]	[0.751, 0.999]
ADULT		RELATIONSHIP	INCOME	AGE		
	BELOW HIGH SCHOOL, FULL-TIME	YES	<= 50K	0.21	[0.605, 0.934]	[0.423, 0.934]
	BELOW HIGH SCHOOL, FULL-TIME	NO	<= 50K	0.21	[0.762, 0.985]	[0.496, 0.985]
	BELOW HIGH SCHOOL, FULL-TIME	YES	> 50K	0.21	[0.066, 0.395]	[0.066, 0.577]
	BELOW HIGH SCHOOL, FULL-TIME	NO	> 50K	0.21	[0.015, 0.238]	[0.015, 0.504]
	ABOVE HIGH SCHOOL, PART-TIME	YES	<= 50K	0.41	[0.186, 0.903]	[0.183, 0.903]
	ABOVE HIGH SCHOOL, PART-TIME	NO	<= 50K	0.41	[0.779, 0.982]	[0.703, 0.983]
	ABOVE HIGH SCHOOL, PART-TIME	YES	> 50K	0.41	[0.017, 0.814]	[0.096, 0.817]
	ABOVE HIGH SCHOOL, PART-TIME	NO	> 50K	0.41	[0.017,0.220]	[0.017, 0.297]
	ABOVE HIGH SCHOOL, FULL-TIME	YES	<= 50K	0.12	[0.310, 0.664]	[0.250, 0.734]
	ABOVE HIGH SCHOOL, FULL-TIME	NO	<= 50K	0.12	[0.725, 0.953]	[0.438, 0.953]
	ABOVE HIGH SCHOOL, FULL-TIME	YES	> 50K	0.12	[0.336, 0.690]	[0.266, 0.750]
	ABOVE HIGH SCHOOL, FULL-TIME	NO	> 50K	0.12	[0.046, 0.275]	[0.046, 0.562]

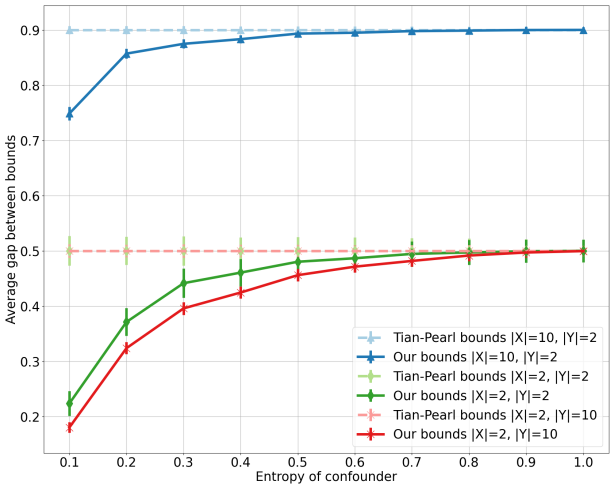
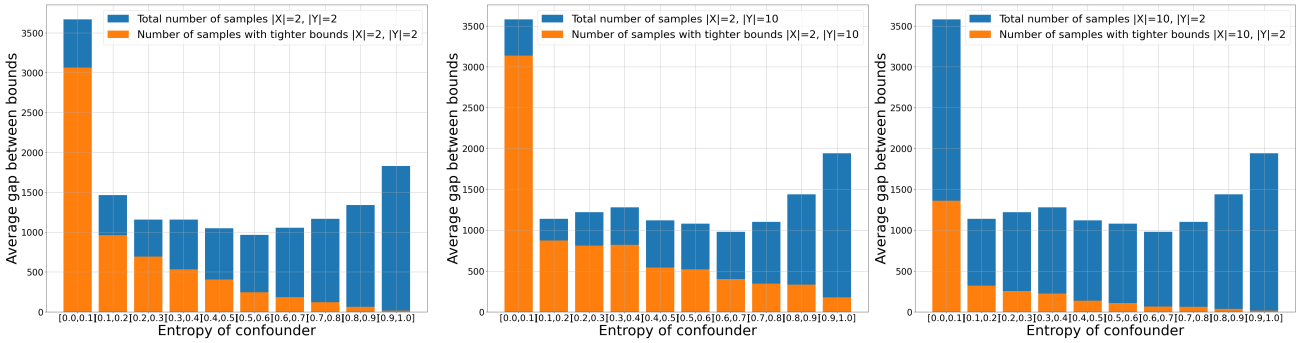


Figure 4. The average gap of our bounds and Tian-Pearl bounds. The x-axis represents the entropy groups and the y-axis represent the average gap in the group. The error bars represent the 95% confidence interval.

bounds with randomly sampled distributions $P(X, Y)$. We change the entropy constraint θ from 1 to 0 for each sampled distribution. We also experiment with the full distribution $P(X, Y, Z)$ where Z is the low entropy confounder and X, Y in high dimensions. We show the experimental results with the real-world dataset Adult (Dua & Graff, 2017). Since our algorithm works for discrete random variables with binary treatment or outcome, we take a subset of features in the graph and modify some features by discretizing continuous variables or combining states with very low probabilities. And finally, we experiment with our method in the finite sample setting and compare two optimization problem formulations.

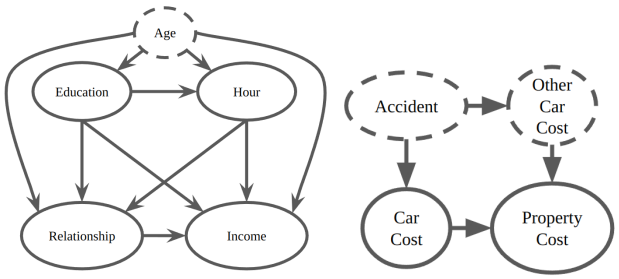
5.1. Randomly Sampled Distributions

First, we want to compare the gaps of our bounds and Tian-Pearl bound. We use randomly sampled data to compare the gaps. We sample the full joint distribution $P(X, Y, Z)$ according to the Figure 1a. Then we treat $P(X, Y)$ as observational data and variable Z as the unobserved confounder and estimate the causal effect using the entropy of Z as θ. The details for sampling the full distribution are in Appendix G. We tested three cases: $(|X| = 2, |Y| = 2), (|X| = 2, |Y| = 10)$ and $(|X| = 10, |Y| = 2)$. For each case, we generate 20000 distributions and compute the bounds $P(y_i|do(x_j))$ for each pair of (i, j). The result is shown in Figure 4. To enhance the interpretability of the results, we group the samples based on the entropy of the confounder and compare the average gap for each entropy group. For example, we consider entropy ranges such as $H(Z) \in [0, 0.1), [0.1, 0.2)$, and so on. Notice that the average gap is smaller when X is binary. This is mainly because a larger portion of samples with $P(x)$ close to 0.5 has a larger entropy threshold. We demonstrate this by plotting the number of samples that yields a tighter bound in Figure 5. When $|X|$ is large, it is less likely to have $P(x)$ close to 0.5, and as the Figure 2 shows, the entropy threshold is low when $P(x)$ is close to 0 or 1, so there is a small number of distributions yields



(a) $|X| = 2, |Y| = 2$ (b) $|X| = 2, |Y| = 10$ (c) $|X| = 10, |Y| = 2$

Figure 5. The number of samples with tighter bounds. The blue bars represent the total number of distributions in each group and the orange bars show the number of distributions with tighter bounds.



(a) ADULT (b) INSURANCE

Figure 6. Causal graphs for the real-world experiments. Only a subset of nodes in the dataset is shown in the figure. Unused variables are omitted.

tighter bounds as shown in Figure 5c. On the other hand, when $|X| = 2$, it is more likely to obtain $P(x)$ that close to 0.5 while $P(y|x)$ is close to the boundary. So the entropy threshold is higher on average, and more distributions with tighter bounds are shown in Figure 5a and Figure 5b.

Next, we will consider experiments in a more realistic setting and see how the entropy constraint could be useful in the real-world problem of causal inference.

5.2. Real-World Dataset Experiment

In this section, we experiment with the INSURANCE dataset (Binder et al., 1997) and the ADULT dataset (Dua & Graff, 2017).

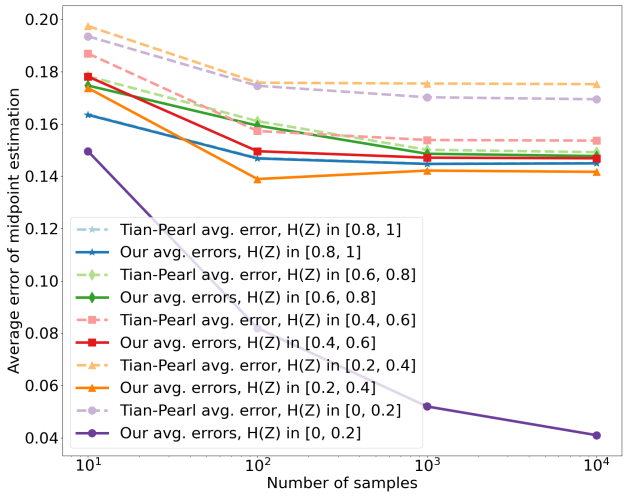
For the INSURANCE dataset, we aim to estimate the causal effect of Car Cost on the expected claim of the Property Cost. We consider the variable Accident as an unobserved variable with known entropy. The Car Cost and Property Cost claim are confounded through the cost of the other car in the accident as shown in Figure 6b. The results in Table 3 indicate narrow bounds on the causal effect when the entropy of the confounder is small. Therefore, we can have confidence in predicting the expected claim based on car cost, even in the presence of the confounding variable.

For the ADULT Dataset (Dua & Graff, 2017), we take a subset of variables from the dataset with the causal graph as shown in Figure 6a. In this experiment, we treat age as a protected feature, which may not be accessible from the dataset, and only the entropy of age is known. If we assume age not having a too complex effect on other variables, i.e., the causal effects of any variable to the income is not much different for groups of people under 65 on average; and similarly for groups of people above 65. The above assumption enables us to discretize the age variable into two categories: "young" and "senior", using a cutting point of 65. Since there are other confounding variables between cause and effect, we take the conditional joint distribution as the subgroup and compute the bounds. Some of the results are summarized in Table 3. One way to interpret the results is to determine whether the causal effect is positive or negative. Our tighter bounds can help in establishing a positive causal effect by comparing the lower bound of $P(Y = 1|do(X = 1))$ with the upper bound of $P(Y = 1|do(X = 0))$. Similarly, for a negative causal effect, we would compare the upper bound of $P(Y = 1|do(X = 1))$ with the lower bound of $P(Y = 1|do(X = 0))$. For instance, in Table 3, for the subgroup of the population with high school or higher education and full-time jobs, the relationship has a positive effect on income. This can be seen by comparing the lower bound of $P(Income > 50K|do(relationship = 1))$ and the upper bound of $P(Income > 50K|do(relationship = 0))$. Our method of bounding causal effect can be used in decision-making processes involving such scenarios.

In the real-world setting, we could use expert knowledge for the complexity of confounders. Even if the confounder has many states, we could still assume the confounder has small entropy if we know many of these states may have a similar effect on the outcome.

5.3. Finite Sample Experiment

In this section, we conducted experiments using our method in the finite data regime, aiming to estimate bounds from finite data samples. We test cases where $(|X| = 2, |Y| = 2)$, $(|X| = 2, |Y| = 10)$, and $(|X| = 10, |Y| = 2)$. Similar to Section 5.1, we generate 1000 distributions for each case. We use $\{10, 10^2, 10^3, 10^4\}$ samples from each distribution and compute the bounds of casual effect $P(y_i|do(x_j))$ for each pair of (i, j) using the empirical distributions. To evaluate the accuracy, we estimate the causal effect with the midpoint of bounds and calculate the average error of each group. The results for binary X, Y are shown in Figure 7, and the rest are shown in the appendix in Figure 8. Our method has a smaller average error than the Tian-Pearl bounds for the cases $H(Z) \leq 0.8$. For $H(Z) \leq 0.2$, the average error drops rapidly as the number of samples increases. This demonstrates that our method improves the causal effect estimation with finite data.



(a) $|X| = 2, |Y| = 2$

Figure 7. The average error of midpoint estimation with finite samples. The dashed lines are average errors with the Tian-Pearl midpoint estimation, and the solid lines are the average error with our midpoint estimation.

5.4. Comparison of Formulations with Canonical Partition and the Counterfactual Distribution

Equation (1) demonstrates the equivalence of two bounds formulations for binary X and Y without entropy constraints. It can be extended to arbitrary discrete variables straightforwardly. For the entropy constraint in Theorem 3.1, we can envision a table similar to Table 2 where p-th column and q-th row are divided into multiple columns and rows. Intuitively, Theorem 3.1 is an over-parameterization version of Theorem 3.2. In terms of partial identification, those two methods are identical. We verify this with an experiment similar to Section 5.1. We apply both methods with $|X|, |Y| = \{2, 4, 8\}$. For each case, we generate 100 distributions and compute the bounds $P(y_i|do(x_j))$ for each pair of (i, j). The experiments indicate that the two approaches have the same optimal values within a precision of three decimal places. The Table 4 provides the number of parameters and average runtime for the two methods. The counterfactual formulation is significantly more efficient when $|X|$ is large.

Table 4. Comparison of methods of Theorem 3.1 and Theorem 3.2

$\|X\|$	$\|Y\|$	Theorem 3.1		Theorem 3.2	
		Num of Param	Ave time	Num of Param	Ave time
2	2	8	0.54s	4	0.19s
2	4	32	3.37s	8	1.03s
2	8	128	18.39s	16	2.94s
4	2	64	13.35s	8	1.62s
8	2	2048	1138.72s	16	5.06s

6. Conclusion

In this paper, we proposed a way to utilize entropy to estimate the bounds of the causal effect. We formulate optimization problems with counterfactual probability, which significantly reduces the number of parameters in the optimization problem. We demonstrate a method to compute the entropy threshold easily so that we can use the entropy threshold as a criterion for applying entropy constraint. For the real-world problem, if we know that two variables are confounded by a confounder with entropy no more than the entropy threshold, we can apply the method and obtain tighter bounds. Another possible scenario where our method can be applied is when the distribution of the confounder is not provided as a joint distribution. Instead, only the marginal distributions $P(Z)$, $P(X, Y)$ are available, as in the example presented by Li and Pearl (2022). In such cases, our method can be utilized to derive tighter bounds.

Our optimization methods work for any discrete X, Y. Only the computation of the entropy threshold requires either X or Y being binary. For future works, it would be worthwhile to explore the tight bounds condition for non-binary X, Y. To obtain entropy thresholds that scale with the number of states of the observed variables, one might need to consider the dependence between different queries $P(y|do(x_j))$ for $j \in |X|$.

Acknowledgements

We thank the reviewers for their valuable feedback. This research has been supported in part by NSF Grant CAREER 2239375.

References

Ay, N. and Polani, D. Information flows in causal networks. *Advances in complex systems*, 11(01):17–41, 2008.

Balazadeh Meresht, V., Syrgkanis, V., and Krishnan, R. G. Partial identification of treatment effects with implicit generative models. *Advances in Neural Information Processing Systems*, 35:22816–22829, 2022.

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

Binder, J., Koller, D., Russell, S., and Kanazawa, K. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.

Bowden, R. J. and Turkington, D. A. *Instrumental variables*. Number 8. Cambridge university press, 1990.

Budhathoki, K. and Vreeken, J. Origo: causal inference by compression. *Knowledge and Information Systems*, 56 (2):285–307, 2018.

Chickering, D. M. and Meek, C. Finding optimal bayesian networks. *arXiv preprint arXiv:1301.0561*, 2012.

Cinelli, C., Kumor, D., Chen, B., Pearl, J., and Bareinboim, E. Sensitivity analysis of linear structural causal models. In *International conference on machine learning*, pp. 1252–1261. PMLR, 2019.

Compton, S., Kocaoglu, M., Greenewald, K., and Katz, D. Entropic causal inference: Identifiability and finite sample results. *Advances in Neural Information Processing Systems*, 33:14772–14782, 2020.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Etesami, J. and Kiyavash, N. Directed information graphs: A generalization of linear dynamical graphs. In *2014 American control conference*, pp. 2563–2568. IEEE, 2014.

Freedman, D. A. *Statistical models and causal inference: a dialogue with the social sciences*. Cambridge University Press, 2010.

Geiger, P., Janzing, D., and Schölkopf, B. Estimating causal effects by bounding confounding. In *UAI*, pp. 240–249, 2014.

Glymour, M., Pearl, J., and Jewell, N. P. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Hu, Y., Wu, Y., Zhang, L., and Wu, X. A generative adversarial framework for bounding confounded causal effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12104–12112, 2021.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Schölkopf, B. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

Jung, Y., Kasiviswanathan, S., Tian, J., Janzing, D., Blöbaum, P., and Bareinboim, E. On measuring causal contributions via do-interventions. In *International Conference on Machine Learning*, pp. 10476–10501. PMLR, 2022.

Kilbertus, N., Kusner, M. J., and Silva, R. A class of algorithms for general instrumental variable models. *Advances in Neural Information Processing Systems*, 33: 20108–20119, 2020.

Kocaoglu, M., Dimakis, A. G., Vishwanath, S., and Hassibi, B. Entropic causal inference. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Lauritzen, S. L. and Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

Li, A. and Pearl, J. Bounds on causal effects and application to high dimensional data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5773–5780, 2022.

Lindley, D. V. and Novick, M. R. The role of exchangeability in inference. *The annals of statistics*, pp. 45–58, 1981.

Lv, B.-M., Quan, Y., and Zhang, H.-Y. Causal inference in microbiome medicine: Principles and applications. *Trends in microbiology*, 29(8):736–746, 2021.

Meilia, P. D. I., Freeman, M. D., Zeegers, M. P., et al. A review of causal inference in forensic medicine. *Forensic Science, Medicine and Pathology*, 16(2):313–320, 2020.

Padh, K., Zeitler, J., Watson, D., Kusner, M., Silva, R., and Kilbertus, N. Stochastic causal programming for bounding treatment effects. *arXiv preprint arXiv:2202.10806*, 2022.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

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

Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62 (3):54–60, 2019.

Pearl, J. Comment: understanding simpson's paradox. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 399–412. 2022.

Pratt, J. W. and Schlaifer, R. On the interpretation and observation of laws. *Journal of Econometrics*, 39(1-2): 23–52, 1988.

Quinn, C. J., Kiyavash, N., and Coleman, T. P. Directed information graphs. *IEEE Transactions on information theory*, 61(12):6887–6909, 2015.

Richardson, T. S. and Robins, J. M. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Robins, J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of chronic diseases*, 40:139S–161S, 1987.

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

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Tian, J. and Pearl, J. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

Tian, J. and Pearl, J. *A general identification condition for causal effects*. eScholarship, University of California, 2002.

Vreeken, J. Causal inference by direction of information. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 909–917. SIAM, 2015.

Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., and Gretton, A. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*, 2020.

Zhang, J. and Bareinboim, E. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 1778–1780, 2017.

Zhang, J. and Bareinboim, E. Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12207–12215, 2021.

Zhang, J., Tian, J., and Bareinboim, E. Partial counterfactual identification from observational and experimental data. In *International Conference on Machine Learning*, pp. 26548–26558. PMLR, 2022.

A. Proof of Theorem 3.1

Recall the Theorem 3.1.

Theorem 3.1. *Let (X, Y) be the pair of variables in the causal graph in Figure 1a with the joint distribution $P(X, Y)$. Suppose $|X| = n, |Y| = m$. Assuming X and Y are confounded by a set of small entropy unobserved variables Z, i.e., $H(Z) \leq \theta$ for some $\theta \in \mathbb{R}$. The causal effect of x_q on y_p is bounded by LB $\leq P(y_p|do(x_q)) \leq$ UB, where*

$$\textit{LB}/\textit{UB} = \min / \max \left(\sum_{i=0}^{m^n-1} \sum_{j=0}^{n-1} a_{ij} P(x_j) \right)$$

subject to

$$\sum_{i,j} a_{ij} P(x_j) = 1,$$

$$\sum_{i=0}^{m^{n-1}} a_{iq} P(x_q) = P(y_p, x_q),$$

$$0 \leq a_{ij} \leq 1 \; \forall i, j,$$

$$\sum_{i,j} a_{ij} P(x_j) \log \left(\frac{a_{ij}}{\sum_k a_{ik} P(x_k)} \right) \leq \theta.$$

Proof. To show the LB and UB bound the causal effect, we first need to show the causal effect lies in the feasible set of the optimization problem.

Let (R_y, R_x) be a canonical partition of (X, Y). Let $a_{ij} = P(R_y = i|R_x = j)$. By the construction of finite response variables, each $P(y_i|x_q)$ equal to the sum of m^{n-1} terms of $P(R_y|R_x)$. Let D_k be a set of indices such that $\sum_{i \in D_k} P(R_y = i|R_x = q) = P(y_i|x_q)$ and $\sum_{i \in D} P(R_y = i|R_x = q) = 1$ where $D = \bigcup_{i \in [n]} D_i$. Then we have the following relations.

$$\sum_{ij} a_{ij} P(x_j) = \sum P(R_y, R_x) = 1$$

$$\sum_{i=0}^{m^{n-1}} a_{iq} P(x_q) = \sum_{i \in D_k} P(R_y = i|R_x = q) P(x_q) = P(y_p, x_q)$$

$$\sum_{i,j} a_{ij} P(x_j) \log \left(\frac{a_{ij}}{\sum_k a_{ik} P(x_k)} \right)$$

$$= I(R_x; R_y) \leq \theta.$$

Since R_x and R_y are d-separated by the confounder, by the data processing inequality, the mutual information between R_x, R_y is less than the entropy of the confounder. So the last inequality holds. Therefore we have $P(y_0|do(x_0))$ is in the feasible set.
Since mutual information is a convex function of the conditional distributions, the set of a_{ij} satisfies $I(R_x; R_y) \leq \theta$ is convex. The objective function and all other constraints are linear functions of a_{ij}, so the optimization problem is convex and obtains global optimal in the feasible set. □

We use the CVXPY package to solve the problem and formulate the constraint according to the Disciplined Convex Programming rules.

B. Proof of Theorem 3.2

Recall the Theorem 3.2.

Theorem 3.2. *Let (X, Y) be the pair of variables in the causal graph in Figure 1a with the joint distribution $P(X, Y)$. Suppose $|X| = n, |Y| = m$. Assuming X and Y are confounded by a set of small entropy unobserved variables Z, i.e.,*

$H(Z) \leq \theta$ for some $\theta \in \mathbb{R}$. The causal effect of x_q on y_p is bounded by LB $\leq P(y_p|do(x_q)) \leq$ UB, where

$$\textit{LB}/\textit{UB} = \min / \max \left(\sum_j b_{pj} P(x_j) \right)$$

subject to

$$\sum_{i,j} b_{ij} P(x_j) = 1,$$
$$b_{iq} P(x_q) = P(y_i, x_q) \; \forall i,$$
$$0 \leq b_{ij} \leq 1 \; \forall i, j,$$
$$\sum_{i,j} b_{ij} P(x_j) \log \left(\frac{b_{ij}}{\sum_k b_{ik} P(x_k)} \right) \leq \theta.$$

Proof. To show the LB and UB bound the causal effect, we first need to show the causal effect lies in the feasible set of the optimization problem.

Let $P(Y_{x_q}, X)$ be the counterfactual distribution for $x_q \in X$. Let $b_{ij} = P(Y_{x_q} = y_i | x_j)$, Then we have the following

$$\sum_{ij} b_{ij} P(x_j) = \sum P(R_y, R_x) = 1$$

$$b_{iq} P(x_q) = P(Y_x = y_i | x_n) P(x_q) = P(y_i, x_q) \; \forall i$$

$$\sum_{i,j} b_{ij} P(x_j) \log \left(\frac{b_{ij}}{\sum_k b_{ik} P(x_k)} \right)$$
$$= I(Y_x; X) \leq \theta.$$

Since Y_x and X are d-separated by the confounder, by the data processing inequality, the mutual information between them is less than the entropy of the confounder. So the last inequality holds. Therefore we have $P(y_0|do(x_0))$ in the feasible set.

Since mutual information is a convex function of the conditional distributions, the set of b_{ij} satisfies $I(Y_x; X) \leq \theta$ is convex. The objective function and all other constraints are linear functions of b_{ij}, so the optimization problem is convex and obtains global optimal in the feasible set.

□

We use the CVXPY package to solve the problem and formulate the constraint according to the Disciplined Convex Programming rules.

C. Proof of Lemma 4.2

Recall the Lemma 4.2

Lemma 4.2. *Let (X, Y) be the pair of binary variables in the causal graph in Figure 1a. Consider $P(Y_x, X)$ for any $x \in X$. Assume, without loss of generality, $P(y|x) \geq P(y'|x)$. Then the following conditions are equivalent:*

1. *$P(Y_x = y)$ attain the Tian-Pearl lower bound,*
2. *$P(Y_x = y')$ attain the Tian-Pearl upper bound,*
3. *$I(Y_x; X)$ is maximized for the given $P(X, Y)$.*

Proof. By the law of total probability, we have that

$$P(Y_x = y) = P(Y_x = y|x)P(x) + P(Y_x = y|x')P(x'),$$

and similarly

$$P(Y_x = y') = P(Y_x = y'|x)P(x) + P(Y_x = y'|x')P(x').$$

From the observational distribution, we have $P(Y_x = y|x) = P(y|x)$, $P(Y_x = y'|x) = P(y'|x)$. Denote $p = P(Y_x = y|x')$, $1 - p = P(Y_x = y'|x')$.

We first show the case $P(y'|x) \leq P(y|x)$.
(1 $\implies$ 2) Assume $P(Y_x = y)$ attain the Tian-Pearl lower bound, i.e. $P(Y_x = y) = P(y, x)$. Since $P(Y_x = y|x) = P(y|x)$, we have $P(Y_x = y|x')P(x') = 0$. Since $P(x') > 0$, $P(Y_x = y|x') = 0$, so $P(Y_x = y'|x') = 1$. Then we have $P(Y_x = y') = P(Y_x = y'|x)P(x) + P(x') = 1 - P(x, y)$ attain the Tian-Pearl upper bound. Thus 1 $\implies$ 2.

(2 $\implies$ 3) Assume $P(Y_x = y')$ attain the Tian-Pearl upper bound, we have $P(Y_x = y'|x') = 1$ and $P(Y_x = y|x') = 0$. We want to show that the mutual information is maximized when $p = 1$. Since $I(Y_x; X)$ is a convex function of $P(Y_x|X)$, it is a convex of p. $I(Y_x; X) = 0$ when $p = P(Y_x = y|x)$, and monotonically increasing for both $p > P(Y_x = y|x)$ and $p < P(Y_x = y|x)$. So $I(Y_x; X)$ obtains the local maximum at two boundaries $p = 0, 1$. To compare those two points, denote $I(Y_x; X)$ as the mutual information if $p = 0$, and I' as the mutual information if $p = 1$. Then we have $I - I' = P(x') \left(\log \frac{P(x')}{1+P(y'|x)} - \log \frac{P(x')}{1+P(y|x)} \right) \leq 0$, since $P(y'|x) \leq P(y|x)$. The global maximum of mutual information is at $p = P(Y_x = y|x') = 1$.

(3 $\implies$ 1) Assumes $I(Y_x; X)$ attain maximum given $P(X, Y)$. The above argument shows that $P(Y_x = y|x') = 1$. So $P(Y_x = y) = P(x) + P(x, y)$ attain the Tian-Pearl upper bound. □

D. Proof of Lemma 4.3

Recall Lemma 4.3

Lemma 4.3. *Let (X, Y) be the pair of variables in the causal graph in Figure 1a, where $|X| = 2$ and $|Y| = m$. The causal effect $P(Y_x = y_p)$ attain the Tian-Pearl upper bound when $P(Y_x = y_p|x') = 1$; attain the Tian-Pearl lower bound with minimum mutual information when $P(Y_x = y_i|x') = \frac{P(Y_x=y_i|X=x)}{\sum_{j\neq p} P(Y=y_j|X=x)}$ for all $i \neq p$.*

Proof. Given $P(Y, X)$, we have $P(Y_x = y_i|x_i) = P(y_i|x)$ for all $i \leq n$. If $P(Y_x = y_p|x') = 1$, then $P(Y_x = y_p)$ attain the Tian-Pearl upper bound:

$$P(Y_x = y_p) = P(Y_x = y_p|x)P(x) + P(Y_x = y_p|x')P(x') = P(y_p, x) + P(x') = 1 - \sum_{i\neq p} P(y_i, x).$$

Next show the minimum mutual information that attain the Tian-Pearl lower bound. $P(Y_x = y_i) = P(Y_x = y_i|x)P(x) + P(Y_x = y_i|x')P(x')$ attain the Tian-Pearl lower bound if $P(Y_x = y_i|x') = 0$ for all $i \neq p$.
Since we fixed $P(Y_x|x) = P(Y|x)$, the domain of the mutual information is to a $(n-1)$simplex Δ^{n-1} of $P(Y_x|x')$. Since $I(Y_x; X)$ is convex with respect to $P(Y_x|X)$, this restricted function is also convex. Clearly, the restricted function obtains minimum when $P(Y_x|x') = P(Y_x|x)$. Since we fixed $P(y_p|x') = 0$, this corresponding to the restricted function on the $(n-2)$-simplex. With a similar argument, this restricted function is also convex. Now we only need to find the local extrema on the $(n-2)$-simplex.

Let $P(Y_x = y_p|x') = 0$, and denote $P(y_i|x) = \alpha_i$ for all $i \leq n$ and $P(Y_x = y_i|x') = \beta_i$ for all $1 \leq i \leq n$. So $P(Y_x) = [\alpha_0 P(x), \alpha_1 P(x) + \beta_1 (P(x'), \ldots, \alpha_n P(x) + \beta_n P(x')]$.
Using the grouping property of entropy, we can write entropy as

$$\begin{aligned} H(Y_x) &= H_b\left(\alpha_0 P(x)\right) + H\left(\frac{\alpha_1 P(x) + \beta_1 P(x')}{1 - \alpha_0 P(x)}, \ldots, \frac{\alpha_n P(x) + \beta_n P(x')}{1 - \alpha_0 P(x)}\right)\left(1 - \alpha_0 P(x)\right) \\ &= H_b\left(\alpha_0 P(x)\right) + H_b\left(\frac{\alpha_1 P(x) + \beta_1 P(x')}{1 - \alpha_0 P(x)}\right)\left(1 - \alpha_0 P(x)\right) \\ &\quad + H\left(\frac{\alpha_2 P(x) + \beta_2 P(x')}{1 - \alpha_0 P(x)}, \ldots, \frac{\alpha_n P(x) + \beta_n P(x')}{1 - \alpha_0 P(x)}\right)\left(\frac{\sum_{i=2}^{n}(\alpha_i P(x) + \beta_i P(x'))}{1 - \alpha_0 P(x)}\right) \end{aligned}$$

Similarly, we can write the conditional entropy as

$$\begin{aligned}
H(Y_x|X) &= P(x)H(Y_x|x) - P(x')H(Y_x|x') \\
&= P(x)H(Y_x|x) - P(x')H(\beta_1, \ldots, \beta_n) \\
&= P(x)H(Y_x|x) - P(x')H_b(\beta_1) - P(x')H\left(\frac{\beta_2}{\sum_{i=2}^n \beta_i}, \ldots, \frac{\beta_n}{\sum_{i=2}^n \beta_i}\right) P\left(\sum_{i=2}^n \beta_i\right)
\end{aligned}$$

the mutual information as

$$\begin{aligned}
I(Y_x; X) &= H(Y_x) - H(Y_x|X) \\
&= H_b\left(\alpha_0 P(x)\right) + H_b\left(\frac{\alpha_1 P(x) + \beta_1 P(x')}{1 - \alpha_0 P(x)}\right)\left(1 - \alpha_0 P(x)\right) \\
&\quad + H\left(\frac{\alpha_2 P(x) + \beta_2 P(x')}{1 - \alpha_0 P(x)}, \ldots, \frac{\alpha_n P(x) + \beta_n P(x')}{1 - \alpha_0 P(x)}\right)\left(\frac{\sum_{i=2}^n (\alpha_i P(x) + \beta_i P(x'))}{1 - \alpha_0 P(x)}\right) \\
&\quad - P(x)H(Y_x|x) - P(x')H_b(\beta_1) - P(x')H\left(\frac{\beta_2}{\sum_{i=2}^n \beta_i}, \ldots, \frac{\beta_n}{\sum_{i=2}^n \beta_i}\right) P\left(\sum_{i=2}^n \beta_i\right)
\end{aligned}$$

Now denote terms that do not involve β_1 as some constant. We can write the mutual information as follows.

$$I(Y_x; X) = C_1 + \left(1 - \alpha_0 P(x)\right) H_b\left(\frac{\alpha_1 P(x) + \beta_1 P(x')}{1 - \alpha_0 P(x)}\right) + C_2 - C_3 - P(x')H_b(\beta_1) - C_4$$

Then take the derivative with respect to β_1 and get

$$\begin{aligned}
\frac{\partial I(Y_x; X)}{\partial \beta_1} &= \left(1 - \alpha_0 P(x)\right)\left(\log \frac{1 - \alpha_0 P(x) - (\alpha_1 P(x) + \beta_1 P(x'))}{\alpha_1 P(x) + \beta_1 P(x')}\right)\frac{P(x')}{1 - \alpha_0 P(x)} - P(x') \log \frac{1 - \beta_1}{\beta_1} \\
&= P(x')\left(\log \frac{1 - (\alpha_0 + \alpha_1)P(x) + \beta_1 P(x')}{\alpha_1 P(x) + \beta_1 P(x')} - \log \frac{1 - \beta_1}{\beta_1}\right)
\end{aligned}$$

Then we can find the local extrema by setting the derivative to zero.

$$\begin{aligned}
\frac{\partial I(Y_x; X)}{\partial \beta_1} &= 0 \\
P(x') \log \frac{1 - (\alpha_0 + \alpha_1)P(x) - \beta_1 P(x')}{\alpha_1 P(x) + \beta_1 P(x')} &= P(x') \log \frac{1 - \beta_1}{\beta_1} \\
\log \frac{1 - (\alpha_0 + \alpha_1)P(x) - \beta_1 P(x')}{\alpha_1 P(x) + \beta_1 P(x')} &= \log \frac{1 - \beta_1}{\beta_1} \\
\frac{1 - (\alpha_0 + \alpha_1)P(x) - \beta_1 P(x')}{\alpha_1 P(x) + \beta_1 P(x')} &= \frac{1 - \beta_1}{\beta_1} \\
(\alpha_1 P(x) + \beta_1 P(x'))(1 - \beta_1) &= (1 - (\alpha_0 + \alpha_1)P(x) - \beta_1 P(x'))\beta_1 \\
\alpha_1 P(x) - \beta_1 \alpha_1 P(x) + \beta_1 P(x') &= (1 - (\alpha_0 + \alpha_1)P(x))\beta_1 \\
(1 - (\alpha_0 + \alpha_1)P(x) + \alpha_1 P(x) - P(x'))\beta_1 &= \alpha_1 P(x) \\
(P(x) - (\alpha_0 + \alpha_1)P(x) + \alpha_1 P(x))\beta_1 &= \alpha_1 P(x) \\
(1 - \alpha_0 - \alpha_1 + \alpha_1)P(x)\beta_1 &= \alpha_1 P(x) \\
\beta_1 &= \frac{\alpha_1}{1 - \alpha_0}
\end{aligned}$$

Repeat the steps for $1 \leq i \leq n$, we can get the local minimum at $\beta_i = \frac{\alpha_i}{1 - \alpha_0}$ for all $1 \leq i \leq n$. Since the mutual information is convex, these points give the global minimum of mutual information. □

E. Proof of Lemma 4.4

Recall the Lemma 4.4

Lemma 4.4. *Let (X, Y) be the pair of variables in the causal graph in Figure 1a, where $|Y| = 2$ and $|X| = n$. The causal effect $P(y|do(x_q))$ attain the Tian-Pearl upper bound when $P(Y_{x_q} = y|x_j) = 1, \forall j \neq q$; attain the Tian-Pearl lower bound when $P(Y_{x_q} = y|x_j) = 0, \forall j \neq q$.*

Proof. Given $P(Y, X)$, we have $P(Y_x = y|x_q) = P(y|x_q)$ for all $y \in Y$. Assumes $P(Y_{x_q} = y)$ attain the Tian-Pearl upper bound, i.e.

$$P(Y_{x_q} = y) = 1 - P(y', x_q) = P(y, x_q) + \sum_{j \neq q}(P(y, x_j) + P(y', x_j)) = P(y_p, x_q) + \sum_{j \neq q} P(x_j).$$

On the other hand, we have

$$P(Y_{x_q} = y_p) = \sum_j P(Y_{x_q} = y_p|x_j)P(x_j).$$

Combines the above two equations, we get $P(Y_{x_q} = y_p|x_j) = 1$ for all $j \neq q$.

For the lower bound, assumes $P(Y_{x_q} = y_p) = P(y_p, x_q)$ by a similar argument as above, we have

$$P(Y_{x_q} = y_p) = \sum_j P(Y_{x_q} = y_p|x_j)P(x_j) = P(y_p, x_q) + \sum_{j \neq q} P(Y_{x_q} = y_p|x_j)P(x_j)$$

So from the above two equations, we get $P(Y_{x_q} = y_p|x_j) = 0$ for all $j \neq q$. □

F. Proof of Theorem 4.5

Recall the Theorem 4.5

Theorem 4.5. *Let (X, Y) be a pair of variables in a causal graph G as shown in Figure 1a, where either X or Y is binary. Let (U, V) be two binary variables such that $P(v_0|u_0) = P(y_p|x_q)$, $P(v_1|u_0) = 1 - P(y_p|x_q)$, and $P(u_0) = P(x_q)$. The entropy threshold for the bounds of $P(y_p|do(x_q))$ is equal to $\max(I(U; V))$.*

Proof. Let $P(U, V)$ be the constructed joint distribution according to the theorem. By Lemma 4.2, assuming $P(y'|x) \leq P(y|x)$, $I(U; V)$ is maximum is equivalent to $P(v_0) = P(v_0|u_0)P(u_0) + P(v_0|u_1)P(u_1)$ attain maximum or minimum. That is when $P(v_0|u_1) = 1$ or $P(v_1|u_1) = 1$

If $P(v_0|u_1) = 1$,

$$I(U; V) = H(V) - H(V|U) = H_b((1 - P(y_p|x_q))P(x_q)) - P(x_q)H_b(P(y_p|x_q)) \tag{2}$$

If $P(v_1|u_1) = 1$,

$$I(U; V) = H(V) - H(V|U) = H_b(P(y_p|x_q)P(x_q)) - P(x_q)H_b(P(y_p|x_q)) \tag{3}$$

First, consider the case where Y is a binary variable and $|X| = n$. By Lemma 4.4, $P(Y_{x_q} = y)$ attain the Tian-Pearl upper bound when $P(Y_{x_q} = y|x_j) = 1$ for all $j \neq q$. So we have

$$Y_x = \begin{cases} y & P(y|x_0)P(x_0) + \sum_{j=1}^{n} P(x_j) \\ y' & P(y'|x_0)P(x_0) \end{cases}.$$

Since $P(Y_{x_q} = y|x_j) = 1$ for all $j \neq q$, $H(Y_{x_q}|x_j) = 0$ for all $j \neq q$. So $H(Y_{x_q}|X) = P(x_q)H_b(P(y|x_q))$. Then we have

$$I(Y_{x_q}; X) = H(Y_{x_q}) - H(Y_{x_q}|X) = H_b(P(y'|x_q)P(x_q)) - P(x_q)H_b(P(y|x_q)).$$

This equals to the Equation (2), so we have $P(Y_{x_q} = y)$ attain the Tian-Pearl upper bound implies $I(U; V)$ obtains maximum.

Again by Lemma 4.4, $P(Y_{x_q} = y)$ attain the Tian-Pearl lower bound when $P(Y_{x_q} = y'|x_j) = 1$ for all $j \neq q$. So we have

$$Y_x = \begin{cases} y & P(y|x_0)P(x_0) \\ y' & P(y'|x_0)P(x_0) + \sum_{j=1}^{n} P(x_j). \end{cases}$$

Since $P(Y_{x_q} = y'|x_j) = 1$ for all $j \neq q$, $H(Y_{x_q}|x_j) = 0$ for all $j \neq q$. So $H(Y_{x_q}|X) = P(x_q)H_b(P(y|x_q))$. Then we have

$$I(Y_{x_q}; X) = H(Y_{x_q}) - H(Y_{x_q}|X) = H_b(P(y|x_q)P(x_q)) - P(x_q)H_b(P(y|x_q)).$$

This equals to the Equation (3), so we have $P(Y_{x_q} = y)$ attains the Tian-Pearl lower bound implies $I(U; V)$ obtains maximum.

We have shown for the binary Y,the causal effect $P(Y_x)$ attains Tian-Pearl bounds implies $I(Y_x; X) = \max(I(U; V))$. Suppose we have $I(Y_x; X) \leq H(Z) < \max(I(U; V))$, by the contraposition, $P(Y_x)$ cannot attains Tian-Pearl bounds.

Now consider the case where X is a binary variable and $|Y| = m$. By Lemma 4.3, the causal effect $P(Y_x = y_p)$ attains Tian-Pearl upper bound when $P(Y_x = y_p|x') = 1$; attains lower bound with minimum mutual information when $P(Y_x = y_i|x') = \frac{P(Y_x=y_i|X=x)}{\sum_{j\neq p} P(Y=y_j|X=x)}$ for all $i \neq p$.

For the upper bound case, assuming $P(Y_x = y_p|x') = 1$, we have $P(Y_x = y_i|x') = 0$ and $H(X|y_i) = 0$ for all $i \neq p$. $H(X|Y) = P(y_p)H(X|y_p)$.

The mutual information is

$$I(Y_x; X) = H_b(x) - P(y_p)H(X|y_p).$$

On the other hand, we can write Equation (2) as

$$I(U; V) = H(U) - H(U|V) = H_b(x) - P(y_p)H(X|y_p) = I(Y_x; X).$$

So we have $P(Y_x)$ attains the Tian-Pearl lower bound implies $I(Y_x; X) = \max(I(U; V))$

Next assuming $P(Y_x = y_i|x') = \frac{P(Y_x=y_i|X=x)}{\sum_{j\neq p} P(Y=y_j|X=x)}$ for all $i \neq p$. We have $P(Y_x = y_p|x) = 0$. Denote $P(Y_x = y_i|x) = \alpha_i$. Using the grouping property of entropy, we could get

$$\begin{aligned}
H(Y_x|X) &= P(x)H(Y_x|x) + P(x')H(Y_x|x') \\
&= P(x)H(\alpha_0, \ldots, \alpha_n) + P(x')H\left(\frac{\alpha_0}{1-\alpha_p}, \ldots, \frac{\alpha_{p-1}}{1-\alpha_p}, \frac{\alpha_{p+1}}{1-\alpha_p}, \ldots, \frac{\alpha_m}{1-\alpha_p}\right) \\
&= P(x)\left[H(\alpha_p) + (1-\alpha_p)H\left(\frac{\alpha_0}{1-\alpha_p}, \ldots, \frac{\alpha_{p-1}}{1-\alpha_p}, \frac{\alpha_{p+1}}{1-\alpha_p}, \ldots, \frac{\alpha_m}{1-\alpha_p}\right)\right] \\
&\quad + P(x')H\left(\frac{\alpha_0}{1-\alpha_p}, \ldots, \frac{\alpha_{p-1}}{1-\alpha_p}, \frac{\alpha_{p+1}}{1-\alpha_p}, \ldots, \frac{\alpha_m}{1-\alpha_p}\right) \\
&= P(x)H(\alpha_p) + (P(x)(1-\alpha_p) + P(x'))\, H\left(\frac{\alpha_0}{1-\alpha_p}, \ldots, \frac{\alpha_{p-1}}{1-\alpha_p}, \frac{\alpha_{p+1}}{1-\alpha_p}, \ldots, \frac{\alpha_m}{1-\alpha_p}\right) \\
&= P(x)H(\alpha_p) + (1-\alpha_p P(x))\, H\left(\frac{\alpha_0}{1-\alpha_p}, \ldots, \frac{\alpha_{p-1}}{1-\alpha_p}, \frac{\alpha_{p+1}}{1-\alpha_p}, \ldots, \frac{\alpha_m}{1-\alpha_p}\right).
\end{aligned}$$

Then we have

$$Y_x = \begin{cases} y_0 & \alpha_0 P(x) + \frac{\alpha_0}{1-\alpha_p} P(x') \\ \vdots & \vdots \\ y_p & \alpha_p P(x) \\ \vdots & \vdots \\ y_m & \alpha_m P(x) + \frac{\alpha_m}{1-\alpha_p} P(x') \end{cases}$$

Again by the grouping property, we have

$$
\begin{aligned}
H(Y_x) &= H_b(\alpha_p P(x)) + (1-\alpha_p P(x)) H\left(\frac{\alpha_0 P(x) + \frac{\alpha_0}{1-\alpha_p} P(x')}{1-\alpha_p P(x)}, \dots\right) \\
&= H_b(\alpha_p P(x)) + (1-\alpha_p P(x)) H\left(\frac{\frac{\alpha_0 P(x)(1-\alpha_p)+\alpha_0 P(x')}{1-\alpha_p}}{1-\alpha_p P(x)}, \dots\right) \\
&= H_b(\alpha_p P(x)) + (1-\alpha_p P(x)) H\left(\frac{\frac{\alpha_0 P(x)(1-\alpha_p)+\alpha_0 (1-P(x))}{1-\alpha_p}}{1-\alpha_p P(x)}, \dots\right) \\
&= H_b(\alpha_p P(x)) + (1-\alpha_p P(x)) H\left(\frac{\frac{\alpha_0 P(x)-\alpha_0\alpha_p P(x)+\alpha_0-\alpha_0 P(x)}{1-\alpha_p}}{1-\alpha_p P(x)}, \dots\right) \\
&= H_b(\alpha_p P(x)) + (1-\alpha_p P(x)) H\left(\frac{\alpha_0-\alpha_0\alpha_p P(x)}{(1-\alpha_p)(1-\alpha_p P(x))}, \dots\right) \\
&= H_b(\alpha_p P(x)) + (1-\alpha_p P(x)) H\left(\frac{\alpha_0(1-\alpha_p P(x))}{(1-\alpha_p)(1-\alpha_p P(x))}, \dots\right) \\
&= H_b(\alpha_p P(x)) + (1-\alpha_p P(x))\, H\left(\frac{\alpha_0}{1-\alpha_p}, \dots, \frac{\alpha_{p-1}}{1-\alpha_p}, \frac{\alpha_{p+1}}{1-\alpha_p}, \dots, \frac{\alpha_m}{1-\alpha_p}\right)
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
I(Y_x; X) &= H(Y_x) - H(Y_x|X) \\
&= H_b(\alpha_p P(x)) + (1-\alpha_p P(x))\, H\left(\frac{\alpha_0}{1-\alpha_p}, \dots, \frac{\alpha_{p-1}}{1-\alpha_p}, \frac{\alpha_{p+1}}{1-\alpha_p}, \dots, \frac{\alpha_m}{1-\alpha_p}\right) \\
&\quad - P(x) H_b(\alpha_p) + (1-\alpha_p P(x))\, H\left(\frac{\alpha_0}{1-\alpha_p}, \dots, \frac{\alpha_{p-1}}{1-\alpha_p}, \frac{\alpha_{p+1}}{1-\alpha_p}, \dots, \frac{\alpha_m}{1-\alpha_p}\right) \\
&= H_b(\alpha_p P(x)) - P(x) H_b(\alpha_p) \\
&= H_b(P(y_p|x_q)P(x_q)) - P(x_q) H_b(P(y_p|x_q))
\end{aligned}
$$

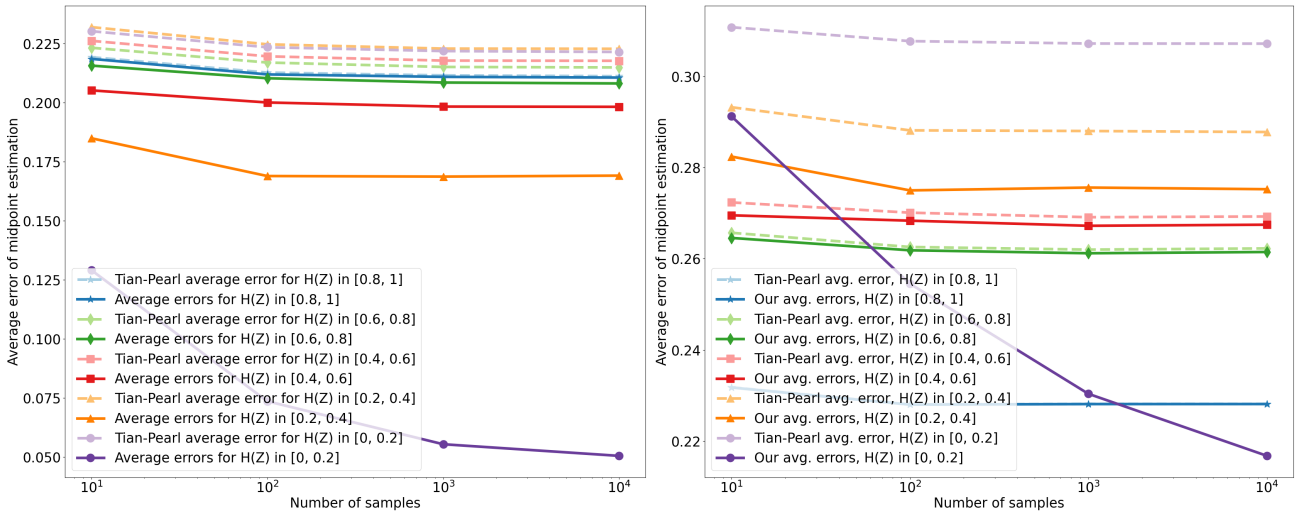
This equals to Equation (3). So the minimum $I(Y_x; X)$ for $P(Y_x = y_p)$ attains Tian-Pearl lower bound is equal to the maximum of $I(U; V)$. For any other distribution where $P(Y_x)$ attains Tian-Pearl lower bound has mutual information greater than $\max(I(U; V))$. Hence $P(Y_x)$ attains Tian-Pearl lower bound implies the $I(Y_x; X) \geq \max(I(U; V))$.

We have shown that for the binary X, the causal effect $P(Y_x)$ attains Tian-Pearl bounds implies $I(Y_x; X) \geq \max(I(U; V))$. Suppose we have $I(Y_x; X) \leq H(Z) < \max(I(U; V))$, by the contraposition, $P(Y_x)$ cannot attains Tian-Pearl bounds. □

G. Sampling the Joint Distribution

Given a DAG as shown in Figure 1a, we first generate $P(Z) \sim Dir(\alpha)$ for some small α value. In this experiment, we use $\alpha = 0.1$. For X with n states, we first construct a vector $\mathbf{v} = \frac{1}{\mathbf{T}}[\mathbf{1}, \frac{1}{\mathbf{2}}, \dots, \frac{1}{\mathbf{n}}]$, where T is normalizing factor such that $\sum \mathbf{v} = \mathbf{1}$. Then for each state of Z, we create a shifted $\mathbf{v_k}$ by rolling the values of $\mathbf{v}$. Then we sample $P(X|z_k) \sim Dir(\mathbf{v_k})$. Similarly, for Y with m states, we construct a vector $\mathbf{u} = \frac{1}{\mathbf{T}}[\mathbf{1}, \frac{1}{\mathbf{2}}, \dots, \frac{1}{\mathbf{m}}]$ and for each x_j, z_k, we sample $P(Y|x_j, z_k) \sim Dir(\mathbf{u_i})$. This procedure was described by Chickering and Meek (2012). They use this method to prevent parent-child relationships between nodes from being uniform for a given DAG.

H. More Finite-sample Experiments



(a) $|X| = 2, |Y| = 10$ (b) $|X| = 10, |Y| = 2$

Figure 8. The average error of midpoint estimation with finite samples

I. Additional Experiment on ASIAN dataset

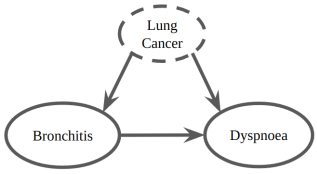


Figure 9. Causal graph for a subset of features from the ASIAN dataset. Unused variables are omitted.

We experiment with the ASIA dataset (Lauritzen & Spiegelhalter, 1988). The causal graph is shown in Figure 9. We compute the bounds of the causal effect of Bronchitis on Dyspnoea with lung cancer as a confounder. The results are summarized in Table 5. In this example, Bronchitis and Dyspnoea are connected through a backdoor path consisting of "smoke" and "lung cancer". In such a case, we can use the confounder with small entropy as a constraint to get tighter bounds, even if other variables in the backdoor path are more complex. The improvement of bounds shows the causal relationship between variables. The lower bound of $Dyspnoea|do(Bronchitis)$ increases from 0.364 to 0.461, and the upper bound of $Dyspnoea|do(NotBronchitis)$ drops from 0.522 to 0.412. Since the bounds lie below 0.5, one can be more certain that Bronchitis has some effect on Dyspnoea.

Table 5. Results of Causal Effect in ASIAN dataset

DATASET	X	Y	H(Z)	OUR BOUNDS	T-P BOUNDS
ASIA	BRONC	DYSP	CANCER		
	YES	YES	0.31	[0.461, 0.914]	[0.364, 0.914]
	YES	NO	0.31	[0.072, 0.412]	[0.072, 0.522]
	NO	YES	0.31	[0.086, 0.539]	[0.086, 0.636]
	NO	NO	0.31	[0.588, 0.928]	[0.478, 0.928]