Approximate Causal Effect Identification under Weak Confounding

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

Abstract

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

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. 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.

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.

The main contributions of this paper are as follows:

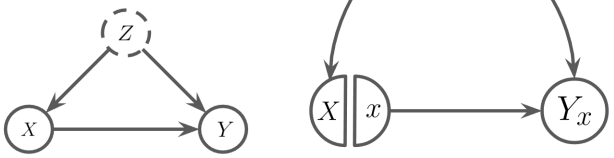
- We formulate a novel optimization problem to efficiently estimate the bounds of causal effect using counterfactual probabilities, and apply the low-entropy confounder constraint using this formulation.
- We examine the conditions on the entropy constraint for the optimization to yield a tighter bound. We analytically show the condition when either or both of X, Y (Figure 1) are binary variables.
- We experiment with our method on both simulated and real-world data, and show that our bound is much tighter than the existing which cannot incorporate such entropy constraints.

2. Background and Notations

Notations. This paper uses uppercase letters X, Y, Z to denote the random variables and lowercase letters x_i, y_i, z_i for their states. We use $\{x, x'\}$, $\{y, y'\}$ 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

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

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2023 by the author(s).



(a) DAG $\mathcal{G}$ with latent confounder

(b) 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.

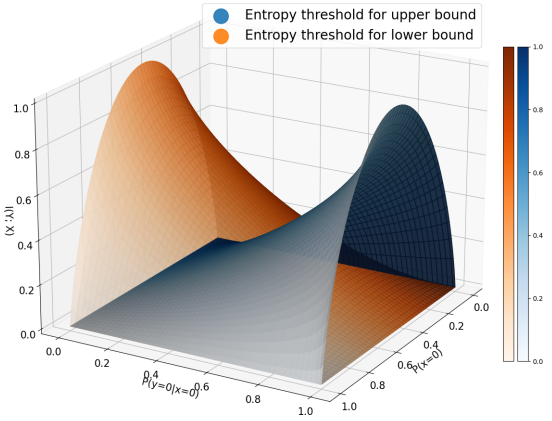


Figure 2. The entropy threshold to obtain tighter bounds. The thresholds are obtained by sampling $P(x_0)$ and $P(y_0|x_0)$ from 0 to 1 which are the x and y axies in the figure. The orange surface represents the entropy threshold for obtaining a tighter upper bound; the blue surface represents the entropy threshold for obtaining a tighter lower bound.

an intervened variable, i.e., $P(Y_x = y) := P(y|do(x))$. This notation is 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.

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 1b, we have Y_x independent from X given Z.

3. Bounding Causal Effect with Entropy Constraint

3.1. Bounds via Counterfactual Probabilities

We propose an optimization problem using counterfactual probabilities to utilize the entropy of the unobserved confounder.

For the causal graph in Figure 1, 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 $y \in Y$.

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

If there were no confounder between X and Y, the interventional distribution would be identical to the conditional distribution $P(Y_x) = P(Y|x)$ for all $x \in X$. Therefore $P(Y_x = y, x') = P(Y_x = y)P(x') = P(y|x')P(x')$. Y_x and X are d-separated by the confounder Z, i.e. $Y_x \perp\!\!\!\perp X|Z$.

By the data processing inequality, the mutual information of Y_x and X is bounded by the entropy of confounder Z. Under the assumption that the confounder Z is weak, i.e., $H(Z) \leq \theta$. 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.

Theorem 3.1. *Let (X, Y) be the pair of variables in the causal graph in Figure 1 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

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

Here b_{ij} are the parameters for the optimization problem. We form the causal effect bounds estimation as a maximization and minimization problem.

4. Condition for Obtaining Tighter Bounds

For Theorem 3.1, the optimization problem has a constraint on $I(Y_x; X)$. 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 (Tian & Pearl, 2000) 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 1. 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 3.1 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 1. 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 1, 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 1, 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 $P(Y_{x_q} = y|x_j) = 0, \forall j \neq q$.*

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

Theorem 4.5. *Let (X, Y) be a pair of variables in a causal graph G as shown in Figure 1, 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 $P(x)$.

5. Experiments

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

5.1. Randomly Sampled Distributions

To compare the bounds with the actual causal effect, we sample the full joint distribution $P(X, Y, Z)$ according to the Figure 1. We treat Z as an unobserved variable and use $P(X, Y)$ as observational data and the entropy of Z as the constraint. The details for sampling the full distribution are in Appendix G. We tested three cases: $(|X| = 2, |Y| = 2)$, $(|X| = 2, |Y| = 10)$ and

Table 1. Results of Causal Effect in ADULT dataset

DATASET	SUBGROUP	X	Y	H(Z)	E-C BOUNDS	T-P BOUNDS
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]

$(|X| = 10, |Y| = 2)$. We generate $20K$ samples for each case and compute the entropy constraint bounds. The result is shown in Figure 6. The samples are grouped according to the entropy of the confounder. We compare the average gap for each group. The error bars represent the 95% confidence interval. The number of samples in each group is shown in Figure 7. Note the asymmetric behavior of $|X|$ and $|Y|$. 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 are small number of distributions yields tighter bounds as shown in Figure 7(c). On the other hand, when $|Y| = 10, |X| = 2$, 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 there are more distributions with tighter bounds as shown in Figure 7(b).

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.

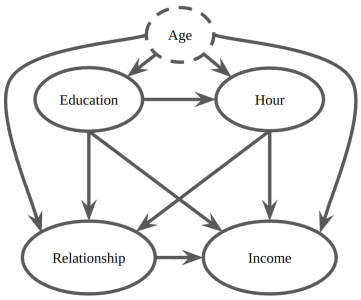


Figure 3. Causal graph for a subset of features from the ADULT dataset.

5.2. Real-World Dataset Experiment

In this section, we experiment with the ADULT Dataset (Dua & Graff, 2017). We take a subset of features from the dataset with the causal graph as shown in Figure 3. 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. Under such an assumption, we convert the age variable to a binary variable as "young" or "senior" people by using 65 as a cutting point. Since other confounding variables exist 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 1.

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

6. Conclusion

In this paper, we proposed a way with counterfactual probability to utilize entropy as a constraint to estimate the bounds of the causal effect. 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. We show the relationship of the entropy threshold with the observed distribution experimentally. We experiment with our method with simulated and real-world data.

References

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

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.

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.

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.

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

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

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.

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

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

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 1 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*

$$
\begin{aligned}
&\textit{LB}/\textit{UB} = \min / \max \left(\sum_j b_{pj} P(x_j) \right) \\
&\textit{subject to} \\
&\sum_{i,j} b_{ij} P(x_j) = 1, \\
&b_{iq} P(x_q) = P(y_i, x_q) \ \forall i, \\
&0 \leq b_{ij} \leq 1 \ \forall i, j, \\
&\sum_{i,j} b_{ij} P(x_j) \log \left(\frac{b_{ij}}{\sum_k b_{ik} P(x_k)} \right) = I(Y_x; X) \leq \theta.
\end{aligned}
$$

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

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

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

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

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.

B. 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 1. 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. □

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

D. 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 1, 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 $j \neq q$. □

E. 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 1, 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 $P(v_0|u_1) = 1$,

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

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

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

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 (1), 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 (2), 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 (1) 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

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

Finally, we have

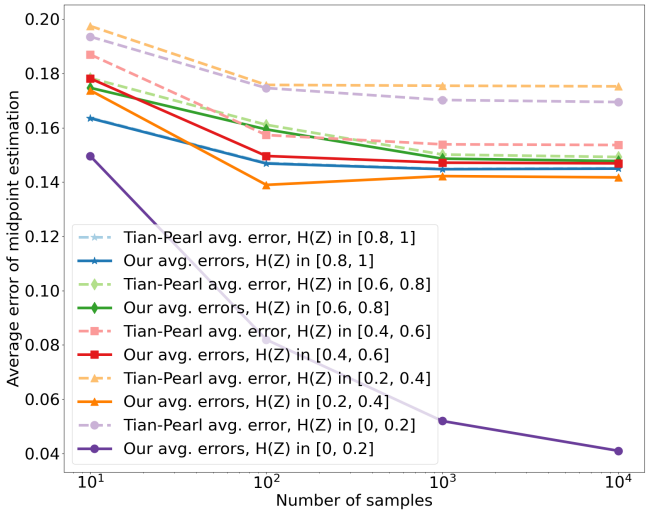
$$
\begin{aligned}
I(Y_x; X) &= H(Y_x) - H(Y_x|X) \\
&= H_b(\alpha_p P(x)) + (1 - \alpha_p P(x))\, H\left(\frac{\alpha_0}{1-\alpha_p}, \ldots, \frac{\alpha_{p-1}}{1-\alpha_p}, \frac{\alpha_{p+1}}{1-\alpha_p}, \ldots, \frac{\alpha_m}{1-\alpha_p}\right) \\
&\quad - P(x) H_b(\alpha_p) + (1 - \alpha_p P(x))\,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

This equals to Equation (2). 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. □

F. Finite Sample Experiment

We experiment with finite samples from simulated data. We test sampling distribution for $(|X| = 2, |Y| = 2)$. We generate 1000 distributions similar to the previous section. We draw samples from each and compute the empirical distribution. The entropy of the confounder groups the distributions. We compute the upper and lower bound for each estimated distribution using our method and take the midpoint of bounds as an estimation. Then we calculate the average error within each group with the ground truth of the causal effect. The results are shown in Figure 4. For confounders with entropy smaller than 0.2, the average error drops rapidly as the number of samples increases. For confounders with entropy smaller than 1, our method has a smaller average error than bounds without entropy constraints.



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

Figure 4. The average error of midpoint estimation with finite samples

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 $\mathbf{v} = \frac{1}{\mathbf{T}}[\mathbf{1}, \frac{\mathbf{1}}{\mathbf{2}}, \ldots, \frac{\mathbf{1}}{\mathbf{n}}]$, 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 $\mathbf{u} = \frac{\mathbf{1}}{\mathbf{T}}[\mathbf{1}, \frac{\mathbf{1}}{\mathbf{2}}, \ldots, \frac{\mathbf{1}}{\mathbf{m}}]$ 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. Convergence of causal effect

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 a uniform distribution. 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 result is shown in Figure 5. 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 $P(x)$.

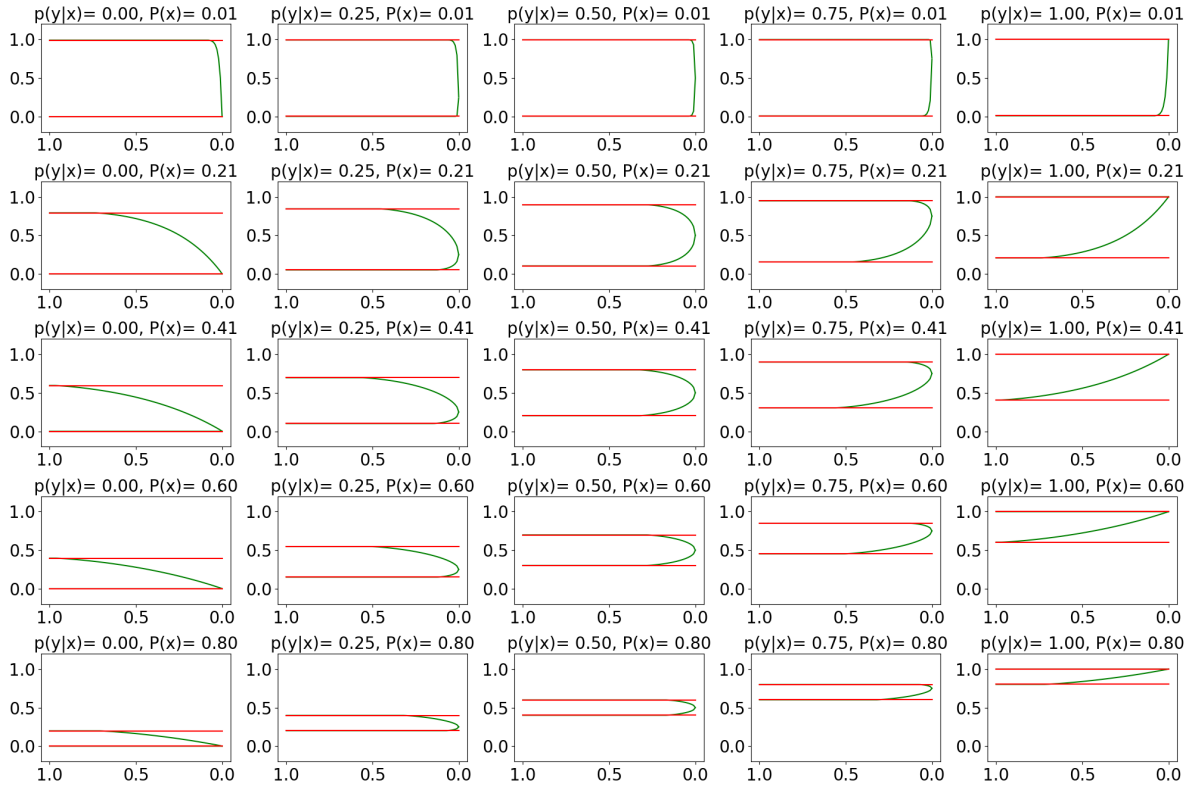


Figure 5. Bounds of the causal effect. The x-axis represents the entropy constraint, and the y-axis represents the causal effect $P(y|do(x))$. For each row $P(y|x)$ increases as $P(x)$ is fixed; $P(x)$ increases from top to bottom. The gap between the upper and lower bound decreases monotonically as $P(x)$ increases. The entropy threshold is high when $P(x)$ is close to 0.5 and $P(y|x)$ is close to 1 or 0.

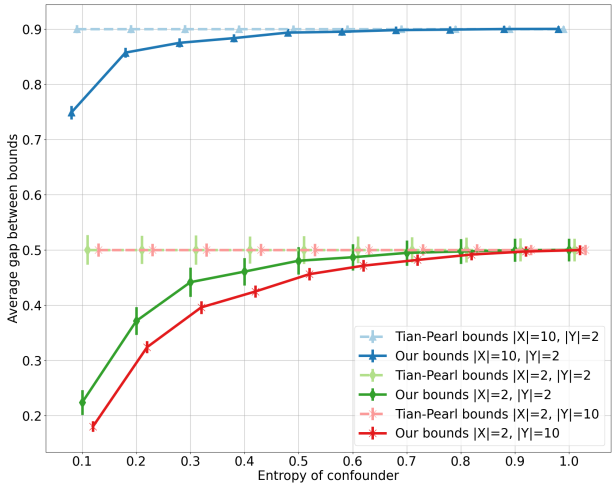
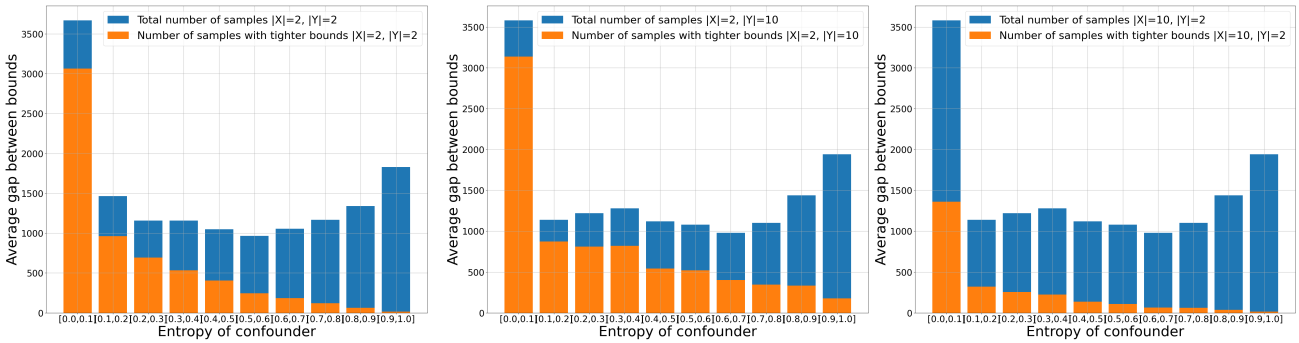


Figure 6. The average gap between bounds



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

Figure 7. The number of samples with tighter bounds. The blue bars represent the total number of distributions in each group and orange bar shows the number of distributions with tighter bound.