# Conditional Common Entropy for Instrumental Variable Testing and Partial Identification

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

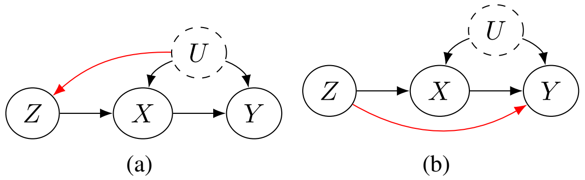
## Abstract

Instrumental variables (IVs) are widely used for estimating causal effects. There are two main challenges when using instrumental variables. First of all, using IV without additional assumptions such as linearity, the causal effect may still not be identifiable. Second, when selecting an IV, the validity of the selected IV is typically not testable since the causal graph is not identifiable from observational data. In this paper, we propose a method for bounding the causal effect with instrumental variables under weak confounding. In addition, we present a novel criterion to falsify the IV with side information about the confounder. We demonstrate the utility of the proposed method with simulated and real-world datasets.

## 1. Introduction

Instrumental variable is a popular approach for estimating causal effect in various domains, such as education (Card, 1993), economy (Rosenzweig & Wolpin, 2000), public health (Hirano et al., 2000), public policy (Abadie, 2003) and, marketing (Blundell et al., 2012). The earliest known work related to instrumental variables was published by Virtue (1929). The goal was to study butter's price elasticity of supply. The local rainfall condition was chosen as the instrumental variable because it affects the butter supply through grass and milk production while not directly related to the demand and price of butter.

Another common scenario for instrumental variables involves natural experiments when a randomized control trial is infeasible. An example of such a scenario is imperfect compliance in the randomized experiment (Balke & Pearl, 1994; Imbens & Angrist, 1994). Specifically, in a randomized controlled trial for the treatment effects, there may be instances where patients do not adhere to their assigned treatments perfectly, influenced by factors that also affect the outcome. In this case, the assigned treatment serves as the instrumental variable for estimating treatment effects. In epidemiology, Mendelian Randomization (MR) uses genetic variants to identify the causal effects between some risk factors and disease. Similarly, the public or market policy can often be used to study the causal effect (Leigh & Schembri, 2004; Davies et al., 2018). Next, we describe some challenges of applying IV to estimate the causal effect.



(a)

(b)

*Figure 1.* (a) shows an invalid IV graph where $Z$ is affected by the unobserved confounder. (b) shows an invalid IV graph, with direct effect from $Z$ to $Y$.

Formally, a variable $Z$ is said to be an instrumental variable if it satisfies three assumptions. (1) $Z$ is not independent of treatment variable $X$ (relevence), (2) $Z$ affects outcome variable $Y$ only through $X$ (exclusion), and (3) $Z$ is independent of the unobserved confounder (exchangeability). The last two assumptions are in general untestable since there is no conditional independence condition between $Z$ and $Y$ in the observational data. Pearl (1995) introduced a testable condition for IV graphs: instrumental inequality, a necessary condition for the instrumental variable. Most of the existing methods provide testable conditions to reject invalid IV. Selecting suitable instrumental variables remains a challenging problem. In this paper, we introduce the conditional common entropy and apply it to establish a new testable condition for valid instrumental variables under the assumption that unobserved confounders are weak.

Another challenge of using instrumental variables (IV) lies in addressing non-identifiable causal queries. One type of approach using regression by making assumptions about the underlying generating model (Bowden & Turkington, 1990; Hartford et al., 2017; Singh et al., 2019; Puli & Ran-

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

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

ganath, 2020; Muandet et al., 2020; Xu et al., 2020; Wang et al., 2021; Frauen & Feuerriegel, 2022; Ailer et al., 2023). Another type of approach estimates the upper and lower bounds of causal effect using observational data, which is also known as partial identification. Alternative definitions of exclusion and exchangeability have been used in existing works. (Swanson et al., 2018) thoroughly reviews various definitions used in partial identification. In this paper, we focus on the least stringent assumption set, i.e., marginal stochastic exclusion and marginal exchangeability, and introduce a novel approach for estimating bounds of causal effect under weak confounder assumption.

An interesting question is how to use invalid instrumental variables in causal inference. Most existing works focus on synthesis IV with multiple candidates or consistent estimators given the majority of the instruments are valid (Windmeijer et al., 2018). Not many works study the partial identification of causal effects with weakly invalid instrumental variables. We provide a brief review of the related works in Section 5. The work that is most relevant to ours was proposed by Cinelli & Hazlett (2022), which introduced a sensitivity analysis framework that quantifies the degree of the IV assumption violation in terms of partial $R^2$. With a similar spirit, we propose a framework to quantify the degree of IV violation with information-theoretic quantity and incorporate it with our approach for estimating bounds of causal effect with weakly invalid instruments.

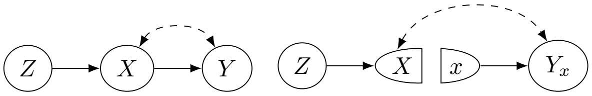
The main contributions of our paper can be summarized as follows:

- We introduce the graph-specific conditional common entropy to quantify the strength of unobserved confounder (Section 3) and provide an algorithm to approximate conditional common entropy for variables in high-dimension.
- We propose a method for bounding causal effects with instrumental variables under weak confounding assumptions. For the invalid instrumental variable, we quantify the strength of violation in terms of conditional common entropy. Our approach can incorporate the strength of IV assumption violation as a sensitivity parameter and get tight bounds when the violation is weak.
- Under the weak confounding assumption, we propose conditional common entropy as a new testable criterion for valid instrumental variables. We show that when the entropy of the unobserved confounder is upper bounded, we can reject invalid instrumental variables effectively. We proposed a heuristic approach for selecting IV from a set of covariates when we cannot make the weak confounding assumption. We demonstrate the effectiveness of the proposed bounds and IV selection method with synthetic and real-world data.

## 2. Backgrounds

**Notations** Throughout this paper, we denote random variables with uppercase letters, e.g., $X, Y, Z$, and their corresponding states are represented by lowercase letters $x_i$, $y_i$, and $z_i$. The cardinality of the variable is denoted as $|X|$. We use $P(y, x)$ as the abbreviation for $P(Y = y, X = x)$. The uppercase letter with lowercase subscript denotes an interventional distribution, e.g., $P(Y_x = y)$ is defined as the probability of observing $y$ under the intervention on $x$.

**Single World Intervention Graph** Richardson & Robins (2013) introduced Single World Intervention Graph (SWIG), a graphical representation that establishes a connection between interventional and counterfactual distributions with the DAG. The interventional distribution $Y_x$ is represented by a node in the SWIG, and the treatment variable is split into the observed variable $X$ and intervention target $X = x$. An example of SWIG corresponding to the IV graph is shown in Figure 2(b).



(a) IV Graph (b) Corresponding SWIG

*Figure 2.* The IV graph and the corresponding Single world intervention graph.

**Common Entropy** Various approaches for measuring common information between two variables have been studied, such as the mutual information, Gács and Körner's common randomness (Gács et al., 1973), and Wyner common information (Wyner, 1975). The exact common information was proposed by Kumar et al. (2014) to measure the common part of two random variables. Unlike mutual information, the exact common information measures the entropy of the simplest variable that explains the dependency between variables $X$ and $Y$. This has also been referred to as the common entropy.

Formally, the common entropy is defined as

$$\begin{aligned}\text{CE}(X;Y) := \min_{q(x,y,w)} &\ H(W)\\ \text{s.t. } &I(X;Y|W) = 0;\\ &q(x,y,w) \text{ compatible with the obs.}\end{aligned}$$

$q(x, y, w)$ compatible with observed distribution means

$$\sum_w q(x, y, w) = p(x, y), \forall x, y;$$
$$0 \leq q(x, y, w) \leq 1, \forall x, y, w$$
$$\sum_{x,y,w} q(x, y, w) = 1.$$

For random variables $X, Y$ that are generated by a common source $U$, we have the inequality $I(X;Y) \leq CE(X;Y) \leq H(U)$ by the data processing inequality.

The common entropy has been used in the entropic causal inference framework by Kocaoglu et al. (2020) to learn the causal graph. In this paper, we extend this idea to finding the minimum entropy variable when conditioned on another set of variables and demonstrate the application of conditional common entropy under the following assumption.

**Assumption 2.1** (Weak Confoundedness)**.** Consider a causal model with a set of endogenous variables $\mathcal{V}$ and exogenous variables $\mathcal{U}$. For any latent confounder $U \in \mathcal{U}$, we have $H(U) \leq \theta$.

In practice, the above assumption might be obtained by expert knowledge of the partial information of the latent confounder, such as marginal distribution or cardinality. For example, if we know the observed variables $\mathcal{V}$ describe most of the variables in the system except for some protected attribute $S$, which we cannot measure from individuals. The strength of the confounder can be bounded by $H(S)$ or $\log_2(|S|)$.

**Partial Identification of Causal Effect** The average causal effect is defined by the expectation of difference between the outcome from the treated and nontreated group,

$$\text{ACE}(X \to Y) = \mathbb{E}[Y_{x_1} - Y_{x_0}].$$

Robins (1989) and Manski (1990) derive bounds of average causal effect in the IV graph, which is known as the "natural bounds". These bounds are sharp under the following assumptions of IV (Swanson et al., 2018).

**Assumption 2.2** (Marginal stochastic exclusion)**.**

$$E[Y_{z,x}] = E[Y_{z',x}], \forall z, z', x$$

**Assumption 2.3** (Marginal exchangeability of counterfactuals)**.**

$$Z \perp\!\!\!\perp Y_{z,x}, \forall z, x$$

This set assumption requires the average directed effects of instrument $Z$ on the outcome $Y$ to be zero at the population level when the treatment $X$ is holding constant.

Recently, some research works have derived tighter bounds with some side information. For example, Li et al. (2023) derives a closed-form expression for tighter bounds given the marginal distribution of confounders. Jiang et al. (2023) proposed the entropic partial identification to obtain the tighter bounds of causal effect given the entropy of confounders. In this paper, we extend the entropic partial identification to the IV setting.

In the next section, we first introduce the conditional common entropy and show the application of common entropy in IV verification and partial identification under Assumption 2.1.

## 3. Conditional Common Entropy

### 3.1. Definition and Properties

**Definition 3.1.** Conditional common entropy of two random variables $Z, Y$ given $X$ with the joint probability distribution $P(X, Y, Z)$ is defined as follows:

$$\text{CCE}(Z;Y|X) := \min_{q(x,y,z,w)} H(W) \tag{1}$$
$$\text{s.t. } I(Z;Y|X,W) = 0;$$
$$q(x,y,z,w) \text{ compatible with the obs}$$

We first show some general properties of conditional common entropy.

**Lemma 3.2** (Bounded by conditional mutual information)**.** *For a pair of random variables $Z, Y$ and a set of variables $\mathbf{X}$, the following inequality holds:*

$$CCE(Z;Y|\mathbf{X}) \geq I(Z;Y|\mathbf{X}) \tag{2}$$

Note that the proposition above also holds when $X$ is an empty set, which shows the common entropy between $Z$ and $Y$ is bounded by the mutual information between them.

An interesting question about conditional common entropy is how it compares to the common entropy. For example, how does $\text{CE}(Z;Y)$ compares to $\text{CCE}(Z;Y|X)$. The following lemma characterizes this relationship.

**Lemma 3.3.** *For a pair of random variables $Z, Y$ and two sets of variables $\mathbf{X}, \mathbf{U}$, the following inequality holds:*

$$CCE(Z;Y|\mathbf{X}) \leq CCE(Z;Y|\mathbf{X},\mathbf{U}) + H(\mathbf{U}) \tag{3}$$

The inequality simplifies to $\text{CE}(Z;Y) \leq \text{CCE}(Z;Y|U) + H(U)$ when $\mathbf{X}$ is an empty set, which provides a clue to the question above.

By the Lemma 3.3, we can relate the conditional common entropy with the latent variable in the causal graph.

### 3.2. Conditional Common Entropy in Causal Graph

**Corollary 3.4.** *For a pair of random variables $Z, Y$ and two sets of variables $\mathbf{X}, \mathbf{U}$, if $Z \perp\!\!\!\perp Y|\mathbf{X}, \mathbf{U}$, then we have*

$$CCE(Z; Y|\mathbf{X}) \leq H(\mathbf{U})$$

This directly follows from Lemma 3.3, since we have $\text{CCE}(Z; Y|\mathbf{X}, \mathbf{U}) = 0$ from the conditional independence.

In the IV graph, the entropy of latent confounders is lower bounded by $\text{CCE}(Z; Y|X)$. This bound is not tight since it does not enforce the independence between variable $Z$ and $W$ which attains the conditional common entropy. Therefore, the derived distribution $P(X, Y, Z, W)$ does not necessarily satisfy the causal Markov assumption. Next, we define the graph-specific conditional common entropy.

**Definition 3.5** (Graph-specific CCE)**.** Let $P(\mathbf{V})$ be the joint distribution over variables $\mathbf{V}$ and Markov relative to the graph $\mathcal{G}$. Let $Z, Y \in \mathbf{V}$ and a set of variables $\mathbf{X} \subset \mathbf{V}$ satisfies $(Z \not\perp\!\!\!\perp_d Y|\mathbf{X})$ in $\mathcal{G}$. Let $(v \leftrightarrow v')$ or $(v \rightarrow v')$ be an edge in $\mathcal{G}$ such that $(Z \perp\!\!\!\perp_d Y|\mathbf{X})$ upon its deletion. Define the **graph-specific conditional common entropy** $\text{CCE}_{\mathcal{G},(v\rightarrow v')}$ to be the minimum entropy of $W$ for some $P(\mathbf{V} \cup \{W\})$ that compatible with the graph $\mathcal{G}'$, where $\mathcal{G}'$ is the graph that replace $(v \rightarrow v')$ with $(v \rightarrow W \rightarrow v')$. Similarly define $\text{CCE}_{\mathcal{G},(v\leftrightarrow v')}$ for $(v \leftrightarrow v')$.

The above definition ensures that the distribution $P(\mathbf{V} \cup \{W\})$ which attains the graph-specific conditional common entropy satisfies the Markov condition regarding $\mathcal{G}'$. For the IV graph $\mathcal{G}$ with observed variables $(X, Y, Z)$ as shown in Figure 2(a), $\mathcal{G}'$ is the graph that includes the latent confounder and marginalizes to $\mathcal{G}$. For the marginalized graphs, Evans (2012) derives the inequality constraints on the observed variables. The Definition 3.5 provides the constraint on the complexity of the latent variable.

In the rest of this paper, We use $\text{CCE}_{\mathcal{IV}}(Z; Y|X)$ as abbreviation for $\text{CCE}_{\mathcal{IV},(X\leftrightarrow Y)}(Z; Y|X)$ in the IV graph (Figure 2(a)). Similarly, we use $\text{CCE}_{\mathcal{D}}(Z; Y|X, U)$ to denote $\text{CCE}_{\mathcal{D},(Z\rightarrow Y)}(Z; Y|X, U)$ in the invalid IV graph (Figure 1(b)).

**Theorem 3.6.** *Given variables $X, Y, Z$ in a causal graph $\mathcal{G}$ with distribution $P(Z, X, Y)$, and latent confounder $U$. If we have $Z \perp\!\!\!\perp Y|X, U$, then the following inequality holds*

$$CCE(Z; Y|X) \leq CCE_{\mathcal{G},\leftrightarrow}(Z; Y|X) \leq H(U) \qquad (4)$$

The above theorem enables us to derive a testable condition for the valid instrumental variable under Assumption 2.1, which we discuss in Section 4.2.

The $\text{CCE}_{\mathcal{D}}(Z; Y|X, U)$ quantifies strength of the edge $Z \rightarrow Y$ in the graph $\mathcal{D}$ (Figure 1(a)). Although this cannot be computed from the observational data if $U$ is a latent variable, we can incorporate it as a sensitivity parameter in our partial identification algorithm with the IV.

### 3.3. Approximating Conditional Common Entropy

In this section, we first show that the conditional common entropy $\text{CCE}(Z; Y|X)$ can be computed in terms of the common entropy.

**Proposition 3.7.** *Given $P(X, Y, Z)$ with $|X| = n$. Let $W_i$ be the random variable that attains the common entropy for $P(Z, Y|x_i)$. The conditional common entropy can be computed by $H(W) = \sum_i H(W'_i)P(x_i)$ where $W'_i$ is some permutation of $W_i$.*

Proposition 3.7 shows that the conditional common entropy can be computed from the variables that attain the common entropy of conditional distributions. The exact conditional common entropy can be computed for binary $Z, Y$ since there exists a closed-form solution for the common entropy.

Computing the exact value of common entropy is a challenging task since it involves solving a non-convex optimization problem. Another problem is how to apply the additional conditional independence constraint for graph-specific CCE. To address these issues, we propose an iterative algorithm inspired by (Kocaoglu et al., 2020) to approximate the graph-specific conditional common entropy [1].

First, we introduce the relaxed objective function by incorporating the conditional independence constraint as a regularization term in the loss function.

$$\mathcal{L} = I(Z; Y|X, W) + (\beta_0 + \beta_1)H(W) - \beta_1 H(W|Z) \quad (5)$$

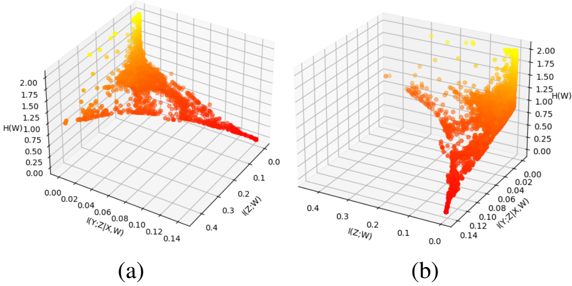
The loss function consists of three terms: $I(Z; Y|X, W)$, $\beta_0 H(W)$, and $\beta_1 I(Z; W)$. The first two terms correspond to finding the minimum entropy variable that separates $Z$ and $Y$. The third term $\beta_1 I(Z; W) = \beta_1(H(W) - H(W|Z))$ corresponds to the additional independence constraint for $W$ in the IV graph. The relaxation of the loss function allows us to search for the latent variable with the following IV LatentSearch algorithm.

The algorithm takes the joint distribution $P(X, Y, Z)$, a random initialization of $q(W|X, Y, Z)$, the number of search iterations, and a pair of parameters $\beta_0, \beta_1$. At each iteration $i$, we obtain the conditional distributions $q_i(W|\cdot)$ and update the joint $q_{i+1}(X, Y, Z, W)$. The updated terms are found through the partial derivative of the loss function in Equation (5). We also show that the output from Algorithm 1 is also a stationary point of Equation (5), and therefore a necessary condition for the original problem in Equation (1) to be optimal.

[1] Our code is available at https://github.com/ziwei-jiang/Conditional-Common-Entropy

**Algorithm 1** IV LatentSearch

```
Input: Joint distribution $P(X,Y,Z)$; Number of iterations $N$; initialization $q(W|X,Y,Z)$; $\beta_0, \beta_1 \geq 0$.
for $i \leftarrow 1$ to $N$ do
    Form the joint:
    $q_i(X,Y,Z,W) \leftarrow q_i(W|X,Y,Z)P(X,Y,Z)$.
    Get posteriors:
    $q_i(W) \leftarrow \sum_{x,y,z} q_i(X,Y,Z,W)$
    $q_i(W|X) \leftarrow \frac{\sum_{y,z} q_i(X,Y,Z,W)}{\sum_{y,z,w} q_i(X,Y,Z,W)}$
    $q_i(W|X,Z) \leftarrow \frac{\sum_{y} q_i(X,Y,Z,W)}{\sum_{y,w} q_i(X,Y,Z,W)}$
    $q_i(W|X,Y) \leftarrow \frac{\sum_{z} q_i(X,Y,Z,W)}{\sum_{z,w} q_i(X,Y,Z,W)}$
    Update:
    $q_{i+1}(X,Y,Z,W) \leftarrow \frac{q_i(W|X,Z)q_i(W|X,Y)q_i(U)^{\beta_0+\beta_1}}{f(X,Y,Z)q_i(W|X)q(W|Z)^{\beta_1}}$
    where $f(X,Y,Z) = \sum_u \frac{q_i(W|X,Z)q_i(W|X,Y)q_i(U)^{\beta_0+\beta_1}}{q_i(W|X)q(W|Z)^{\beta_1}}$
end for
Return: $q_N(W|X,Y,Z)P(X,Y,Z)$
```



(a) (b)

*Figure 3.* After running Algorithm 1 with $\beta_0$ and $\beta_1$ range from 0.5 to 0, the conditional common entropy can be estimated as the minimum value of $H(W)$ when $I(Z;W)$ and $I(Y;Z|X,W)$ are sufficiently close to zero.

**Theorem 3.8.** *The output from Algorithm 1 after convergence is also a stationary point of Equation (5).*

As the value of $\beta_0, \beta_1$ decreases, greater emphasis is placed on conditional independence constraints. Therefore, the mutual information $I(Z;W)$ and $I(Z;Y|X,W)$ decreases as $H(W)$ increases. An example of the plot of these three terms is in Figure 3.

# 4. Application of Conditional Common Entropy in IV graph

## 4.1. Bounding Causal Effect with IV under weak confounding

We first provide a motivational example of the problem.

Suppose we want to test the effect of a new drug that developed in the lab. Among 1,000 patients, 512 were randomly selected to take the new drug, and the placebo was assigned to the rest. The selected patients were asked to take the drug every day at home, but only 282 of them properly followed the instructions. Of the 230 patients who did not properly follow the instructions, 99 did not recover after the trial, and of the patients who followed the instructions, 196 recovered. Of the 488 patients who were not assigned treatment, 341 of them did not recover. The data is summarized in the Table 1.

| | Assigned | Not Assigned |
|---|---|---|
| Took Drug | 196 out of 282 are recovered | 0 |
| Did Not Take Drug | 131 out of 230 are recovered | 147 out of 488 are recovered |

*Table 1.* Drug Effect Example

If the study was conducted with double-blind trials with the placebo assigned to the non-treatment groups, the drug assignment could be taken as an instrumental variable for the treatment effect, which is bounded by the natural bounds. However, if we cannot rule out the possibility that the drug assignment may have a placebo effect, the variable $Z$ may become an invalid instrumental variable as shown in Figure 1(b). In such cases, the bounds of causal effect can only be reliably estimated by Tian-Pearl bounds:

$$\begin{aligned}
0.278 &\leq P(y_1|do(x_0)) \leq 0.560 \\
0.196 &\leq P(y_1|do(x_1)) \leq 0.914 \\
-0.364 &\leq \text{ACE}(X \rightarrow Y) \leq 0.636
\end{aligned}$$

This result is not informative since the average causal effect could be either positive or negative. It would be helpful to know how the bounds are affected by the strength of the direct effect from $Z$ to $Y$, e.g., the placebo effect.

We can use $\phi := \text{CCE}_{\mathcal{D}}(Z;Y|X,U)$ as a sensitivity parameter to quantify the strength of direct effect from $Z$ to $Y$. As depicted by Figure 4, $\phi = 0$ if and only if there is no directed or bidirected path between $Z$ and $Y$. Moreover, we have the following result to connect this parameter to the interventional distribution.

**Theorem 4.1.** *Given variables $Z, X, Y, U$ in an invalid IV graph as shown in Figure 1(a) with distribution $P(Z,X,Y,U)$, we have the following inequality*

$$I(Y_x;Z) \leq \mathit{CCE}_{\mathcal{D}}(Z;Y|X,U).$$

We propose the following method to estimate the bounds of causal effect under Assumption 2.1 and take consideration of possible IV violation by using $\text{CCE}_{\mathcal{D}}(Z;Y|X,U)$ as a sensitivity parameter.

**Theorem 4.2.** *Under Assumption 2.1, for variables $(X,Y,Z)$ with $|X| = n, |Y| = m$, and $|Z| = l$ and the compatible joint distribution $P(X,Y,Z)$, assumes $\mathit{CCE}_{\mathcal{IV}}(Z;Y|X,U) = \phi$. The causal effect of $x_t$ on $y_o$*

*is bounded by* $LB \leq P(y_o|do(x_t)) \leq UB$*, where*

$$LB/UB = \min / \max \left( \sum_{jl} b_{ojl} P(z_l) \right)$$

*subject to*

$$P(z_k) b_{itk} = P(y_i, x_t, z_k) \ \forall i, k; \sum_{ij} b_{ijk} = 1 \ \forall k$$

$$\sum_i b_{ijk} = P(x_j|z_k) \ \forall j, k; 0 \leq b_{ijk} \leq 1 \ \forall i, j, k$$

$$\sum_{ijk} \log \left( \frac{b_{ijk} P(z_k)}{(\sum_{j'k'} b_{ij'k'} P(z_{k'}))(\sum_{i'} b_{i'jk} P(z_k))} \right)$$
$$b_{ijk} P(z_k) \leq \theta + \phi.$$

*If the Assumption 2.3 holds, we can replace the last inequality constraint with the following two:*

$$\sum_{ij} b_{ijk} \log \left( \frac{b_{ijk}}{(\sum_{j'} b_{ij'k})(\sum_{i'} b_{i'jk})} \right) \leq \theta \ \forall k,$$

$$\sum_{ijk} b_{ijk} P(z_k) \log \left( \frac{\sum_{j'} b_{ij'k}}{(\sum_{j'k} b_{ij'k} P(z_k))} \right) \leq \phi.$$

Theorem 4.2 is a relaxation of the natural bounds by relaxing the conditional independence constraint to conditional mutual information constraint on the interventional distributions. Our method incorporates different cases of IV violation with the mutual information constraint as shown in Figure 4.

A lower bound of sensitivity parameter directly follows from Theorem 3.6 and Lemma 3.3.

**Corollary 4.3.** *Under the setting of Theorem 4.2, the sensitivity parameter is lower bounded by*

$$\phi \geq CCE(Z; Y|X) - \theta$$

Apply the Theorem 4.2 to the drug effect example, we obtain the plot shown in Figure 5, as the strength of the edge $\phi \to 0$, our bounds converge to the natural bounds.

### 4.2. Instrumental Variable Verification

A testable condition for a variable $Z$ to be a valid instrument follows directly from Theorem 3.6,

**Corollary 4.4.** *Under the Assumption 2.1, a covariate* $Z$ *is a valid instrumental variable only if*

$$CCE_{\mathcal{IV}}(Z; Y|X) \leq \theta.$$

Similar to many existing IV verification approaches, the test with conditional common entropy can only falsify the invalid instrument since it is a necessary condition for IV. In Section 6, we show that our proposed method effectively rejects invalid instrumental variables.

## 5. Related Work

**Bounding the Causal Effect** Given a causal graph with latent confounders, the causal effect might not be uniquely identifiable. The instrumental variables can be used to derive tight bounds of causal effects in those cases. Balke & Pearl (1997) discussed bounding causal effects for discrete treatment and outcome with instrumental variables and the canonical partition method. Richardson & Robins (2014) obtained sharp bounds on the average causal effect under the assumption $X \perp\!\!\!\perp Y_x, Y_{x'}$. Zhang et al. (2022) introduce the canonical SCM to derive bounds of counterfactual queries with discrete variables. Duarte et al. (2023) propose an automated method to determine the feasible region of the causal effect that can be applied to any causal graph with discrete variables. More recently, this method has been extended to continuous variables setting (Kilbertus et al., 2020; Zhang & Bareinboim, 2021; Hu et al., 2021; Padh et al., 2023).

On the other hand, when the causal graph is unknown, a Partial Ancestral Graph (PAG) can be learned with causal discovery algorithms (Spirtes et al., 2001). A PAG describes the Markov equivalence class of the Maximal Ancestral Graph (MAG) that represents the projection of the true DAG on the observed variables. Since multiple DAGs correspond to a PAG, the causal effect may not be uniquely identifiable. Malinsky & Spirtes (2016) presents a method to determine possible causal effects by enumerating the MAG in the Markov equivalence class. Wang et al. (2023) proposed an insightful method with super-exponentially less complexity that outputs the same set of possible causal effects without enumerating causal graphs.

**Instrumental Variables Veritification** Pearl (1995) derived the instrumental inequality to falsify the instrumental variables and conjectured there's no testable condition when $X$ is continuous. This has been shown by Bonet (2001) and Gunsilius (2021). Furthermore, Bonet (2001) showed that instrumental inequality is not sufficient when the instrument has a cardinality of 3. Richardson & Robins (2010) derived geometric characterization of instrumental inequality. Evans (2012) generalized instrumental inequality with e-separation. Wang et al. (2017) proposed simple statistical tests to validate the binary instrumental variable model. Sharma (2018) proposed a necessary and probably sufficient test for instrumental variables. Kédagni & Mourifié (2020) generalized instrumental inequality to discrete treatment unrestricted outcome and instrument. Xie et al. (2022) derived a necessary condition for a variable to be a valid instrument for the linear non-Gaussian acyclic causal model.

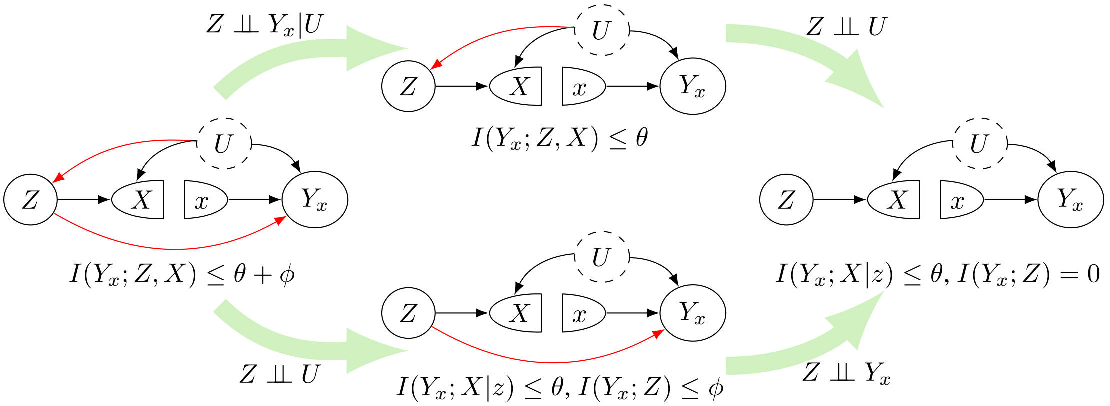

$Z \perp\!\!\!\perp Y_x | U$
$U$
$Z$
$X$
$x$
$Y_x$
$I(Y_x; Z, X) \leq \theta$
$Z \perp\!\!\!\perp U$
$U$
$Z$
$X$
$x$
$Y_x$
$U$
$Z$
$X$
$x$
$Y_x$
$I(Y_x; Z, X) \leq \theta + \phi$
$I(Y_x; X|z) \leq \theta, I(Y_x; Z) = 0$
$U$
$Z$
$X$
$x$
$Y_x$
$Z \perp\!\!\!\perp U$
$I(Y_x; X|z) \leq \theta, I(Y_x; Z) \leq \phi$
$Z \perp\!\!\!\perp Y_x$

*Figure 4.* Single world intervention graphs corresponding to violations of IV assumptions. The left graph with $Z$ violates both Assumption 2.2 and Assumption 2.3. The top graph with $Z$ violates Assumption 2.3. The bottom graph with $Z$ violates Assumption 2.2. The constraints are more stringent from left to right.

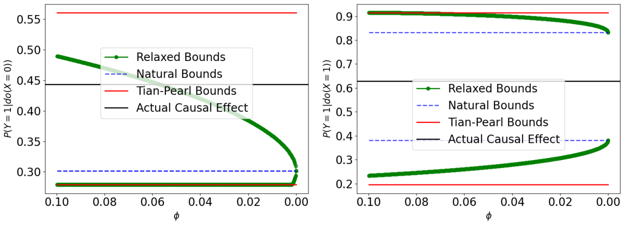


*Figure 5.* Bounds of the drug effect example in Table 1. The green curve shows the bounds as $\phi$ goes to zero. Since the IV is invalid, the natural bounds overestimate the average causal effect.

**Incorporate Invalid IV** The invalid instrumental variable may cause a larger bias than the correlated noise. Bound et al. (1995) shows that the bias of causal effect estimation could be exacerbated if the instrumental variable is weakly associated with treatment. In epidemiology, the allele scores can be used as instrumental variables (Burgess & Thompson, 2013) for Mendelian randomization. The allele score summarizes multiple genetic variants that are associated with risk factors. It requires each variant used to compute the allele score to be a valid instrumental variable. Kuang et al. (2020) relaxed this assumption and proposed a new method to synthesize summarized IV that can handle invalid IV candidates. Hartford et al. (2021) proposed a machine learning-based method of instrumental variable estimation with multiple candidate IV when the majority of them are valid through an ensemble model of IV estimators. Cinelli & Hazlett (2022) developed a sensitivity analysis framework using omitted variable bias to handle the violation of exclusion restriction and exchangeability assumption. They provided bounds on the bias if the maximum explanatory power of omitted variables is not stronger than a multiple of the explanatory power of observed variables.

**Information Theoretic Causal Inference** Many works in causality have been done with the information-theoretic approach. Relative entropy (Janzing et al., 2013) and directed information (Etesami & Kiyavash, 2014; Quinn et al., 2015) have been used to study the strength of the causal effect. Researchers have used entropy (Kocaoglu et al., 2017) and minimum description length (Budhathoki & Vreeken, 2018) for learning causal structure. A closely related work by Finkelstein et al. (2021), introduced entropic inequality constraints that are implied by e-separation relations in hidden variable DAGs. They derive a measure of causal influence called minimal media entropy that can be used to measure the strength of an edge. The author also provides a lower bound for the latent variable in the IV graph in terms of mutual information. By Lemma 3.2, their lower bound is smaller than the conditional common entropy. Therefore, our result offers a tighter lower bound for the minimum entropy confounder in the IV graph.

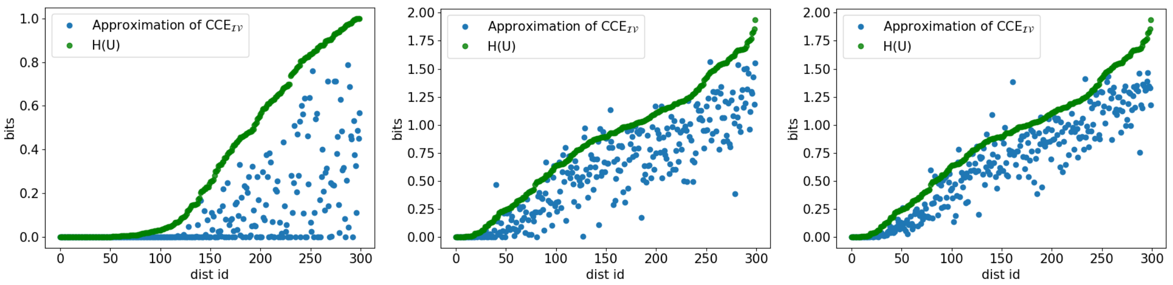
## 6. Experiment

In this section, we first demonstrate the proposed method with simulated data and then provide some case studies with real-world data with instrumental variables.

### 6.1. Examing Conditional Common Entropy with Synthetic Data

We sample the conditional probability distributions according to the factorization in the IV graph. We discuss more details of the experiment setting in Appendix A.

The results are shown in Figure 6. In most cases, our algorithm outputs a good approximation of the $\text{CCE}_{\mathcal{IV}}$.



(a) $|X| = |Y| = |Z| = 2$ (b) $|X| = 2|, |Y| = |Z| = 4$ (c) $|X| = |Y| = |Z| = 4$

*Figure 6.* Plots for comparing conditional common entropy with entropy of latent confounder in IV. Note that the value of conditional common entropy does not monotonically change with the entropy of the latent confounder. This is because the connection induced by the large entropy latent confounder could be small.

### 6.2. Testing Instrumental Variables with Conditional Common Entropy

In this section, we demonstrate the application of conditional common entropy in the IV setting. Similar to the other IV verification method, we can only falsify the invalid instrument. We simulate data with a similar setting as the previous section but from an invalid IV graph as shown in Figure 1(b). For each sample distribution, we use Algorithm 1 to approximate the conditional common entropy and compare it with the entropy of the confounder. Similar to the instrumental inequality, which rejects the distributions that cannot generated by the IV graph, our approach utilizes the side information, the entropy of the unobserved confounder, to check if a distribution cannot be generated by an IV graph confounded by a small entropy variable. The results are shown in Figure 11.

When the side information is not available, i.e., when Assumption 2.1 does not hold, we show that the conditional common entropy can still provide a signal for choosing the instrumental variable.

We generate distributions with a procedure as described in Appendix A. Given two candidate IVs: $Z$ and $V$, where $Z$ is a valid instrument and $V$ has a direct effect on $Y$. We compares the value of CCE$_{\mathcal{IV}}(Z; Y|X, V)$ and CCE$_{\mathcal{IV}}(V; Y|X, Z)$. As shown in Figure 7, for high-dimension variables, the valid IV almost always attains smaller CCE$_{\mathcal{IV}}$ compared to the invalid IV. Intuitively, this is because when computing the CCE$_{\mathcal{IV}}$ with valid IV $Z$, all other paths between $Z$ and $Y$ are blocked and thus it is easier to separate those two variables. We demonstrate our method for selecting IV from more than two candidates in Appendix J.

### 6.3. Bounding the Causal Effect with IV and Weakly Invalid IV

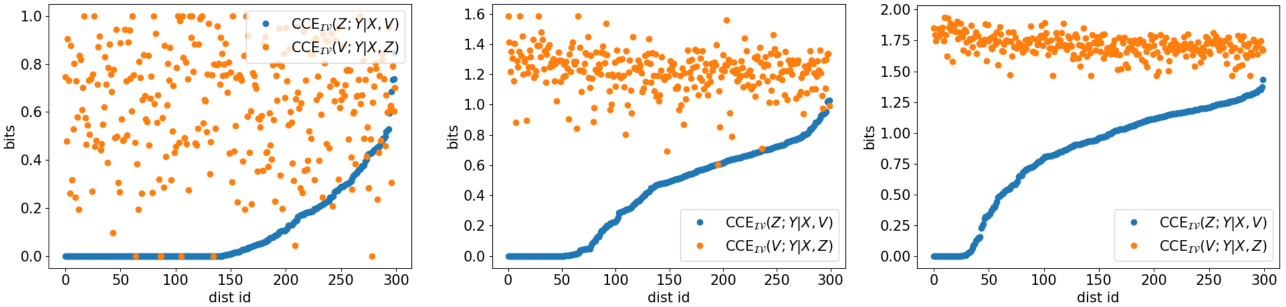
Next, we demonstrate the partial identification of causal effects with both valid and weakly invalid instrumental variables. The data are generated as described in Appendix A.

To visualize the results, we measure the gaps between bounds and group the samples by the entropy of confounders. We plot the average value of the gaps within each group. As shown in Figure 10, for the invalid IV graph $\mathcal{D}$, we can apply our algorithm to get tighter bounds, while the natural bounds may not be valid. For the IV graph, we get bounds in general better than natural bounds when the entropy of the confounder is small.

### 6.4. Case Study: Lung Cancer Dataset

In this section, we demonstrate our result in a more realistic setting with a synthetic dataset introduced by Lauritzen & Spiegelhalter (1988). We take a subset of variables and treat others as unobserved variables. Given we have three variables in the dataset: "Shortness-of-breath (dyspnoea)", "bronchitis", and "smoking". The goal is to study the causal effect of bronchitis on dyspnoea. Suppose lung cancer is the only unobserved variable that might correlated with both bronchitis and dyspnoea. From the marginal distribution, we know it has a small entropy $H(U) = 0.31$. Denote $X =$ bronchitis , $Y =$ dyspnoea, $Z =$ Smoking, and $U =$ Lung cancer. We want to use smoking as an instrument because it does not have a direct effect on dyspnea. However, we are unsure if it correlated with other variables in the confounding path, such as lung cancer. In this case, we can use conditional common entropy to check the validity of the candidate instrumental variable.

By Proposition 3.7, the conditional common entropy can be computed exactly. We find the common entropy value, CCE $= 0.104$. In this case, the conditional common entropy



(a) $|X| = |Y| = |Z| = 2$ (b) $|X| = |Y| = |Z| = 3$ (c) $|X| = |Y| = |Z| = 4$

*Figure 7.* Select an IV from two covariates. For each sample distribution, we plot the conditional common entropy of the outcome $Y$, valid instrument $Z$ (blue), and invalid instrument $V$ (orange). The experimental results show that for high dimensional variables, invalid instruments almost always yield higher conditional common entropy.

is smaller than the entropy $H(U)$, and we cannot reject the invalid instrument with CCE. Apply Algorithm 1, we can find the graph-specific conditional common entropy $CCE_{\mathcal{IV}} = 0.393$, which allows us to reject the invalid IV.

Since we know $\theta = 0.31$ and $\phi = 0$, we can apply $Theorem$ 4.2 to get bounds of causal effect. With the entropy constraint, we obtain the average causal effect bounded by $[-0.078, 0.828]$. Comparing the bounds without entropy constraint $[-0.272, 0.828]$, our lower bound is close to zero, which suggests that a negative causal effect between the two variables is unlikely. Note that the bounds we obtained are the same as using entropy constraint (Jiang et al., 2023). However, in this example, the relationship between smoking and other variables is unknown. Simply ignoring the variable smoking and using $\theta$ might not give the correct result.

### 6.5. Case Study: PimaIndiansDiabetes Dataset

We provide another example with the PimaIndiansDiabetes dataset (Smith & Dickson, 1988) for studying the causal effect of glucose levels on blood pressure. The dataset contains 768 entries of measurements of various health conditions such as glucose, insulin, and BMI. Due to the dataset size and the high dimensionality of the variables, we convert the data to binary variables. For glucose, we group the data to samples with a threshold of 125 mg/dL, which is a threshold of abnormal results in fasting blood glucose tests. Similarly, we binarize the blood pressure with a threshold of 80. Then, for insulin and BMI, we find 85 and 30 as suitable thresholds that maintain relatively high mutual information of joint distribution.

In this example, we do not have access to the entropy of confounders. Therefore, we can compare the $\text{CCE}_{\mathcal{IV}}$ of the two candidates: insulin and BMI. We compute the $\text{CCE}_{\mathcal{IV}}$ for each candidate as described in Section 6.2. We find $\text{CCE}_{\mathcal{IV}}$ for insulin and BMI are $0.66$ and $0.91$ respectively. The result suggests that insulin is likely to be an instrument variable for the causal effect of glucose on blood pressure.

In this example, we discretize the continuous variables with some critical threshold, e.g., with a glucose level above 125 mg/dL as an indicator of diabetes. Our method is demonstrated in this example which represents a higher level of abstraction compared to the original problem.

## 7. Discussion

In this paper, we propose conditional common entropy to quantify the strength of the latent confounder and the strength of a path. We provide an algorithm for approximating conditional common entropy.

The proposed method in this paper relies on the weak confounding assumption, which requires additional knowledge of the latent confounder. Our method is sound for any valid upper bound of entropy. In practice, our method could be used with expert knowledge or other sources of information. For example, in the case study described by Pearl et al. (2016) about "Exercise" and "Cholesterol level" with the "Age" as a confounder. If age information is not collected during the survey, the joint distribution of $P(Exercise, Cholesterol, Age)$ is not available and the causal effect might not be identifiable. In this case, it is cumbersome to recollect the data, but the marginal distribution over the appropriate population might be easier to obtain. In the case when some sensitive attributes such as race, ethnicity, or financial status, are not available per individual. One may apply our method with other data sources to obtain a marginal distribution or only the cardinality of these variables.

## Acknowledgements

We thank the reviewers for their valuable comments and constructive feedback. This research has been supported in part by NSF CAREER 2239375, IIS 2348717, Amazon Research Award, and Adobe Research.

## Impact Statement

This paper presents work whose goal is to advance the field of causal inference. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## References

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

Acemoglu, D., Johnson, S., and Robinson, J. A. The colonial origins of comparative development: An empirical investigation. *American economic review*, 91(5):1369–1401, 2001.

Ailer, E., Hartford, J., and Kilbertus, N. Sequential underspecified instrument selection for cause-effect estimation. In *International Conference on Machine Learning*, pp. 408–420. PMLR, 2023.

Angrist, J. D. and Krueger, A. B. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

Balke, A. and Pearl, J. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pp. 46–54. Elsevier, 1994.

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.

Blundell, R., Horowitz, J. L., and Parey, M. Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation. *Quantitative Economics*, 3(1):29–51, 2012.

Bonet, B. Instrumentality tests revisited. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 48–55, 2001.

Bound, J., Jaeger, D. A., and Baker, R. M. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450, 1995.

Bowden, R. J. and Turkington, D. A. Instrumental variables. *Cambridge Books*, 1990.

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

Burgess, S. and Thompson, S. G. Use of allele scores as instrumental variables for mendelian randomization. *International journal of epidemiology*, 42(4):1134–1144, 2013.

Card, D. Using geographic variation in college proximity to estimate the return to schooling, 1993.

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

Cinelli, C. and Hazlett, C. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Available at SSRN 4217915*, 2022.

Davies, N. M., Dickson, M., Davey Smith, G., Van Den Berg, G. J., and Windmeijer, F. The causal effects of education on health outcomes in the uk biobank. *Nature human behaviour*, 2(2):117–125, 2018.

Duarte, G., Finkelstein, N., Knox, D., Mummolo, J., and Shpitser, I. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, pp. 1–16, 2023.

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

Evans, R. J. Graphical methods for inequality constraints in marginalized dags. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE, 2012.

Finkelstein, N., Zjawin, B., Wolfe, E., Shpitser, I., and Spekkens, R. W. Entropic inequality constraints from e-separation relations in directed acyclic graphs with hidden variables. In *Uncertainty in Artificial Intelligence*, pp. 1045–1055. PMLR, 2021.

Frauen, D. and Feuerriegel, S. Estimating individual treatment effects under unobserved confounding using binary instruments. *arXiv preprint arXiv:2208.08544*, 2022.

Gács, P., Korner, J., et al. Common information is far less than mutual information. *Problems of Control and Information Theory*, 2:149–162, 1973.

Gunsilius, F. F. Nontestability of instrument validity under continuous treatments. *Biometrika*, 108(4):989–995, 2021.

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

Hartford, J. S., Veitch, V., Sridhar, D., and Leyton-Brown, K. Valid causal inference with (some) invalid instruments. In *International Conference on Machine Learning*, pp. 4096–4106. PMLR, 2021.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X.-H. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, 2000.

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.

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

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Schölkopf, B. Quantifying causal influences. *THE ANNALS of STATISTICS*, pp. 2324–2358, 2013.

Jiang, Z., Wei, L., and Kocaoglu, M. Approximate causal effect identification under weak confounding. In *International Conference on Machine Learning*, pp. 15125–15143. PMLR, 2023.

Kédagni, D. and Mourifié, I. Generalized instrumental inequalities: testing the instrumental variable independence assumption. *Biometrika*, 107(3):661–675, 2020.

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., Vishwanath, S., and Hassibi, B. Entropic causal inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Kocaoglu, M., Shakkottai, S., Dimakis, A. G., Caramanis, C., and Vishwanath, S. Applications of common entropy for causal inference. *Advances in neural information processing systems*, 33:17514–17525, 2020.

Kuang, Z., Sala, F., Sohoni, N., Wu, S., Córdova-Palomera, A., Dunnmon, J., Priest, J., and Ré, C. Ivy: Instrumental variable synthesis for causal inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 398–410. PMLR, 2020.

Kumar, G. R., Li, C. T., and El Gamal, A. Exact common information. In *2014 IEEE International Symposium on Information Theory*, pp. 161–165. IEEE, 2014.

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.

Leigh, J. P. and Schembri, M. Instrumental variables technique: cigarette price provided better estimate of effects of smoking on sf-12. *Journal of clinical epidemiology*, 57(3):284–293, 2004.

Li, A., Mueller, S., and Pearl, J. Epsilon-identifiability of causal quantities. *arXiv preprint arXiv:2301.12022*, 2023.

Malinsky, D. and Spirtes, P. Estimating causal effects with ancestral graph markov models. In *Conference on Probabilistic Graphical Models*, pp. 299–309. PMLR, 2016.

Manski, C. F. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33:2710–2721, 2020.

Padh, K., Zeitler, J., Watson, D., Kusner, M., Silva, R., and Kilbertus, N. Stochastic causal programming for bounding treatment effects. In *Conference on Causal Learning and Reasoning*, pp. 142–176. PMLR, 2023.

Pearl, J. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 435–443, 1995.

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

Puli, A. and Ranganath, R. General control functions for causal effect estimation from ivs. *Advances in neural information processing systems*, 33:8440–8451, 2020.

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. Analysis of the binary instrumental variable model, 2010.

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.

Richardson, T. S. and Robins, J. M. Ace bounds; sems with equilibrium conditions. *Statistical Science*, 29(3): 363–366, 2014.

Robins, J. M. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pp. 113–159, 1989.

Rosenzweig, M. R. and Wolpin, K. I. Natural "natural experiments" in economics. *Journal of Economic Literature*, 38(4):827–874, 2000.

Sharma, A. Necessary and probably sufficient test for finding valid instrumental variables. *arXiv preprint arXiv:1812.01412*, 2018.

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

Smith, J. W. and Dickson, W. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, pp. 261. American Medical Informatics Association, 1988.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, prediction, and search.* MIT press, 2001.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., and Richardson, T. S. Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947, 2018.

Virtue, G. The tariff on animal and vegetable oils, 1929.

Wang, L., Robins, J. M., and Richardson, T. S. On falsification of the binary instrumental variable model. *Biometrika*, 104(1):229–236, 2017.

Wang, T.-Z., Qin, T., and Zhou, Z.-H. Estimating possible causal effects with latent variables via adjustment. In *International Conference on Machine Learning*, pp. 36308–36335. PMLR, 2023.

Wang, Z., Zhou, Y., Ren, T., and Zhu, J. Scalable quasi-bayesian inference for instrumental variable regression. *Advances in Neural Information Processing Systems*, 34: 10469–10482, 2021.

Windmeijer, F., Farbmacher, H., Davies, N., and Smith, G. D. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 2018.

Wyner, A. The common information of two dependent random variables. *IEEE Transactions on Information Theory*, 21(2):163–179, 1975.

Xie, F., He, Y., Geng, Z., Chen, Z., Hou, R., and Zhang, K. Testability of instrumental variables in linear non-gaussian acyclic causal models. *Entropy*, 24(4):512, 2022.

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. 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. Data Generation in the Synthetic Experiments

For the synthetic experiment, we sample the data according to the graph Figure 2(a) and Figure 1(b). We follow the procedure described by Chickering & Meek (2012) to sample the conditional distributions.

For the IV graph, we sample $P(Z) \sim$ Dir$(1)$ and $P(U) \sim$ Dir$(0.1)$. Similarly, for each $z$ and $u$, we sample $P(X|z, u) \sim$ Dir$(1)$. For the variable $Y$, we sample $P(Y|x, u) \sim$ Dir$(\mathbf{V_i})$ where $\mathbf{V_i}$ is a rolling shifting vector as described by Chickering & Meek (2012). For the Invalid IV graph, the data generating process is similar to the previous case, except the conditional distribution $P(Y|x, u, z) \sim$ Dir$(\mathbf{V_i})$. To control the degree of IV violation, we replace the edge $Z \rightarrow Y$ with a Markov chain $Z - M - Y$ where $M$ is a small entropy variable sampled from Dirichlet.

In the IV selection experiment, we sample $P(Z), P(V) \sim$ Dir$(1)$ and $P(U) \sim$ Dir$(0.1)$. Then $P(X|z, v, u) \sim$ Dir$(1)$ and $P(Y|u, v, x) \sim$ Dir$(\mathbf{V_i})$. In each case, we first form the joint distribution according to the DAG, then marginalize over the latent variable $U$.

The algorithm converges around 200 iterations. To approximate the CCE, we iteratively search with 100 values of $\beta_0 \in [0, 1]$ and $\beta_1 \in [0, 0.5]$. The result is shown in Figure 8. Then we take the CCE as the minimum entropy $H(W)$ such that both $I(Y; Z|X, W)$ and $I(Z; W)$ are smaller than the threshold $1e - 5$.

## B. Proof of Lemma 3.2

Let the variable $W$ note the variable achieves conditional common entropy with joint distribution $P(X, Y, Z, W)$. By the definition of conditional common entropy, when condition on $X = x$, the variables $Z \rightarrow W \rightarrow Y$ form a Markov chain. By the data processing inequality, we have

$$I(Z; Y|x) \leq H(W|x), \forall x.$$

Take the expectation of $X$ on both sides we get

$$I(Z; Y|X) \leq H(W|X) \leq H(W)$$

## C. Proof of Lemma 3.3

Let $Z, Y$ denote a pair of variables and $\mathbf{X}, \mathbf{U}$ be two sets of variables. Let $W_1$ be the variable that attains CCE$(Z; Y|\mathbf{X})$, i.e. $Z \perp\!\!\!\perp Y|\mathbf{X}, W_1$. Let $W_2$ be the variable that attains CCE$(Z; Y|\mathbf{X}, \mathbf{U})$, i.e. $Z \perp\!\!\!\perp Y|\mathbf{X}, \mathbf{U}, W_2$.

Suppose for the sake of contradiction that $H(W_1) > H(W_2) + H(\mathbf{U})$. Let $W_3$ be the Cartesian of the set $\{W_2\} \cup \mathbf{U}$. Then $H(W_3) = H(W_2, \mathbf{U}) \leq H(W_2) + H(\mathbf{U})$. By the construction, we can find a distribution with $W_3$ such that $Z \perp\!\!\!\perp Y|\mathbf{X}, W_3$ and $H(W_3) < H(W_1)$ which contradicts that $W_1$ attains the conditional common entropy.

## D. Proof of Theorem 3.6

For $W$ that achieves graph-specific conditional common entropy, it also attains the conditional common entropy for the distribution $P(X, Y, Z)$. So the first inequality CCE$(Z; Y|X) \leq$ CCE$_{\mathcal{G}} \leq H(U)$ holds.

For a latent confounder $U$ that is represented by a directed edge in the graph $\mathcal{G}$. Since $U$ is a variable generated by the underlying causal model and the distribution $P(\mathbf{V} \cup \{U\}$ satisfies the Causal Markov condition. So, the distribution $P(\mathbf{V} \cup \{U\}$ satisfies all the independence constraints from the graph. Therefore we have $H(U) \geq$ CCE$_{\mathcal{G},\leftrightarrow}$.

## E. Proof of Proposition 3.7

Let $P(X, Y, Z)$ be the joint probability with discrete variables: $|Y| = m, |Z| = n$. For each $X = x$, let $W$ be the variable that attains common entropy for the conditional distribution with $\sum_w P(y|w, z, x)P(w|z, x) = P(x, y, z)$. Without loss of generality, assumes $W$ has $l$ states. Then $P(W|Z, x)$ is an $l$ by $n$ matrix, and $P(W|Z, x)P(Z|x) = P(W|x)$ attains the minimum entropy. Any row permutation $\hat{P}(W|Z, x)$ does not change the value of $H(W|x)$, and we have $\hat{P}(P|W, x)\hat{P}P(W|Z, x) = P(Y|Z, x)$ with $\hat{P}(Y|W, x)$.

Now we show that a $W$ constructed by combining permutation of the common entropy variable is the smaller entropy variable such that $(Z \perp\!\!\!\perp Y|X, W)$.

Suppose for the sake of contradiction that there exists another variable $U$ that attains smaller entropy and $(Z \perp\!\!\!\perp Y|X, U)$. Then it holds that $(Z \perp\!\!\!\perp Y|U, x)\forall x$. Clearly, this implies at least for some $x$, we have $H(U|x) < H(W|x)$, which contradicts that $W$ attains minimum entropy for $x$.

## F. Proof of Theorem 3.8

To show that Algorithm 1 converges to a stationary point, we first write the loss functions that incorporate the constraints.

$$
\begin{aligned}
\mathcal{L}(q(W|x,y,z)) &= I(Z;Y|X,U) + \beta_0 H(W) + \beta_1 I(Z;W) + \sum_{xyz} \delta_{xyz} \left( \sum_w q(w|x,y,z) - 1 \right) \\
&= I(Z;Y|X,W) + \beta_0 H(W) + \beta_1 H(W) - \beta_1 H(W|Z) \sum_{xyz} \delta_{xyz} \left( \sum_w q(w|x,y,z) - 1 \right) \\
&= \sum_{xyzw} P(x,y,z) P(w|x,y,z) \log \frac{P(w|x,y,z)P(w|x)}{P(w|x,z)P(w|x,y) + I(Y;Z|X) + (\beta_0+\beta_1)H(W) - \beta_1 H(W|Z)}
\end{aligned}
$$

The stationary point is where the derivative of the loss function equals zero. The partial derivative with respect to the parameter is given as

$$
\begin{aligned}
\frac{\partial P(w|x)}{\partial P(w|x,y,z)} &= P(y,z|x) \\
\frac{\partial P(w|x,z)}{\partial P(w|x,y,z)} &= P(y|x,z) \\
\frac{\partial P(w|x,y)}{\partial P(w|x,y,z)} &= P(z|x,y) \\
\frac{\partial P(w)}{\partial P(w|x,y,z)} &= P(x,y,z)
\end{aligned}
$$

So the partial derivative with respect to the loss function $\mathcal{L}$ is given by

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial P(w|x,y,z)} &= P(x,y,z) \left[ \log P(w|x,y,z) + 1) + (\log P(w|x) + 1) - (\log P(w|x,z) + 1) - (\log P(w|x,y) + 1) \right] \\
&\quad - \beta_0 P(x,y,z)(\log P(w) + 1) + \beta_1 P(x,y,z)(\log P(w|z) + 1) + P(x,y,z)\delta \\
&= P(x,y,z) \left[ \log \frac{P(w|x,y,z)P(w|x)P(w|z)^{\beta_1}}{P(w|x,z)P(w|x,y)P(w)^{\beta_0+\beta_1}} - \beta_0 + \beta_1 + \delta_{xyz} \right].
\end{aligned}
$$

Settting $\frac{\partial \mathcal{L}}{\partial P(w|x,y,z)} = 0$, we get

$$
\begin{aligned}
\frac{1}{2^{\delta_{xyz} - \beta_0 + \beta_1}} &= \frac{P(w|x,y,z)P(w|x)P(w|z)^{\beta_1}}{P(w|x,z)P(w|x,y)P(w)^{\beta_0+\beta_1}} \\
P(w|x,y,z) &= \frac{P(w|x,z)P(w|x,y)P(w)^{\beta_0+\beta_1}}{P(w|x)P(w|z)^{\beta_1}} \frac{1}{2^{\delta_{xyz} - \beta_0 + \beta_1}}
\end{aligned}
$$

So, any stationary point should satisfy the equation above. In addition, we want to enforce $\sum_w P(w|x,y,z) = 1$.

$$\sum_{xyz} \frac{P(w|x,z)P(w|x,y)P(w)^{\beta_0+\beta_1}}{P(w|x)P(w|z)^{\beta_1}} \frac{1}{2^{\delta_{xyz}-\beta_0+\beta_1}} = 1$$

$$\frac{1}{2^{\delta_{xyz}-\beta_0+\beta_1}} \sum_{xyz} \frac{P(w|x,z)P(w|x,y)P(w)^{\beta_0+\beta_1}}{P(w|x)P(w|z)^{\beta_1}} = 1$$

$$\frac{1}{2^{\delta_{xyz}-\beta_0+\beta_1}} = \frac{1}{\sum_{xyz} \frac{P(w|x,z)P(w|x,y)P(w)^{\beta_0+\beta_1}}{P(w|x)P(w|z)^{\beta_1}}}$$

So, the normalization condition is satisfied if

$$P(w|x,y,z) = F(x,y,z) \frac{P(w|x,z)P(w|x,y)P(w)^{\beta_0+\beta_1}}{P(w|x)P(w|z)^{\beta_1}}$$

where $F(x,y,z) = \frac{1}{\sum_{xyz} \frac{P(w|x,z)P(w|x,y)P(w)^{\beta_0+\beta_1}}{P(w|x)P(w|z)^{\beta_1}}}$.

This is the same term in the Algorithm 1, which means a stationary point of Algorithm 1 is equivalence to the stationary point for Equation (5).

## G. Proof of Theorem 4.1

For an invalid IV graph $\mathcal{G}$ with a direct edge from $Z$ to $Y$ (as shown in Figure 1(b)), Let $W$ denote the variable that attains the graph-specific conditional common entropy $CCE_{\mathcal{G}}$ with joint distribution $\hat{P}(W,X,Y,Z,U)$. We use $\hat{P}$ to denote the distribution output from the algorithm. By the definition of graph-specific conditional common entropy, the joint distribution should satisfy (1) $\sum_w \hat{P}(w,x,y,z,u) = \hat{P}(x,y,z,u) \forall x,y,z,u$ and (2) the conditional independence constraints, including $(Z \perp\!\!\!\perp Y|X,W,U)$, $(U \perp\!\!\!\perp Z|W)$, and $(W \perp\!\!\!\perp X|Z,U)$.

We first defined the interventional distributions of $\hat{P}(W,X,Y,Z,U)$ as follows.

$$\hat{P}(Y_x = y) = \sum_{u,w,z} \hat{P}(y|x,u,w)\hat{P}(w,u)$$

$$\hat{P}(Y_x = y|z) = \sum_{u,w} \hat{P}(y|x,u,w,z)\hat{P}(w,u|z)$$

$$\hat{P}(Y_x = y|z,w) = \sum_{u} \hat{P}(y|x,u,w,z)\hat{P}(u|z,w)$$

$$\hat{P}(Y_x = y|w) = \sum_{u} \hat{P}(y|x,u,w)\hat{P}(u|w)$$

Then, we show that the first two distributions are the same as the interventional distribution from the original distribution.

$$\begin{aligned}
P(Y_x = y) &= \sum_{u,z} P(y|x,u,z)P(u,z) && \text{(backdoor adjustment)} \\
&= \sum_{u,z} \left( \sum_{w} \hat{P}(y|x,u,w)\hat{P}(w|x,u,z) \right) P(u,z) && (Z \perp\!\!\!\perp Y | X, W, U) \\
&= \sum_{u,z} \left( \sum_{w} \hat{P}(y|x,u,w)\hat{P}(w|u,z) \right) P(u,z) && (W \perp\!\!\!\perp X | Z, U) \\
&= \sum_{u,z,w} \hat{P}(y|x,u,w)\hat{P}(w,u,z) \\
&= \sum_{u,w} \hat{P}(y|x,u,w)\hat{P}(w,u) && \text{(marginalize out } z\text{)} \\
&= \hat{P}(Y_x = y)
\end{aligned}$$

$$\begin{aligned}
P(Y_x = y|z) &= \sum_{u} P(y|x,u,z)P(u|z) && \text{(backdoor adjustment)} \\
&= \sum_{u} \left( \sum_{w} \hat{P}(y|x,u,w,z)\hat{P}(w|x,u,z) \right) P(u|z) \\
&= \sum_{u} \left( \sum_{w} \hat{P}(y|x,u,w,z)\hat{P}(w|u,z) \right) P(u|z) && (W \perp\!\!\!\perp X | Z, U) \\
&= \sum_{u,w} \hat{P}(y|x,u,w,z)\hat{P}(w,u|z) \\
&= \hat{P}(Y_x = y|z)
\end{aligned}$$

For the last two, it is clear that it is compatible with the original interventional distribution after marginalizing out $w$.

$$\begin{aligned}
&\sum_{w} \hat{P}(Y_x = y|z,w)\hat{P}(w|z) \\
&= \sum_{w,u} \hat{P}(y|x,u,w,z)\hat{P}(u|z,w)\hat{P}(w|z) \\
&= \sum_{w,u} \hat{P}(y|x,u,w,z)\hat{P}(u,w|z) \\
&= \hat{P}(Y_x = y|z) = P(Y_x = y|z)
\end{aligned}$$

$$\begin{aligned}
&\sum_{w} \hat{P}(Y_x = y|w)\hat{P}(w) \\
&= \sum_{w,u} \hat{P}(y|x,u,w)\hat{P}(u|w)\hat{P}(w) \\
&= \sum_{w,u} \hat{P}(y|x,u,w)\hat{P}(u,w) \\
&= \hat{P}(Y_x = y) = P(Y_x = y)
\end{aligned}$$

Next, we want to show that in the distribution of $\hat{P}(W, X, Y, Z, U)$, the variables $(Z, W, Y_x)$ forms a Markov chain.

$$\begin{aligned}
P(Y_x|w,z) &= \sum_u \hat{P}(y|x,u,w,z)\hat{P}(u|z,w) \\
&= \sum_u \hat{P}(y|x,u,w)\hat{P}(u|z,w) && (Z \perp\!\!\!\perp Y|X,W,U) \\
&= \sum_u \hat{P}(y|x,u,w)\hat{P}(u|w) && (U \perp\!\!\!\perp Z|W) \\
&= \hat{P}(Y_x = y|w)
\end{aligned}$$

Since the interventional distributions are equal, we have $I(Y_x;Z) = \hat{I}(Y_x;Z)$. Then by the data processing inequality, we have $I(Y_x;Z) = \hat{I}(Y_x;Z) \leq H(W)$.

## H. Proof of Theorem 4.2

To show the LB and UB are bounds of the causal effect, we need to show that in the IV and invalid IV graphs, the causal effect lies in the feasible region of the optimization problem.

Let $P(Y_{x_t}, X|Z)$ be the counterfactual distribution when intervening on $x_t$. Then the causal effect $P(y_o|do(x_t)) = \sum_{jk} P(Y_{x_t} = y_o, x_j|z_k)$ and the follow equalities hold.

$$\begin{aligned}
&b_{itk}P(z_k) = P(Y_{x_t} = y_i, x_t, z_k) = P(Y_{x_t} = y_i, x_t|z_k)P(z_k) \\
&\sum_{ij} b_{ijk} = 1 = \sum_{ij} P(Y_{x_t} = y_i, x_j|Z) \\
&\sum_{i} b_{ijk} = P(x_j|z_k) = \sum_{i} P(Y_{x_t} = i, x_j|z_k)
\end{aligned}$$

So, there exists some $b_{ijk}$ that satisfies the equality constraints for the causal effect.

Then, we only have to show that for each graph, there is a solution under our mutual information constraints.

First, consider case 1 if a variable $Z$ violates both Assumption 2.2 and Assumption 2.3. Let $W$ be the variable that achieves the conditional common entropy $\text{CCE}_{\mathcal{G}}$. By the definition, we have $Z \perp\!\!\!\perp Y|X,U,W$. Similar to the proof of Theorem 4.1, we have that

$$I(Y_x;Z) \leq \phi. \tag{6}$$

Furthermore, since $Z$ and $U$ blocks all the backdoor path from $X$ to $Y$, we have $X \perp\!\!\!\perp Y_x\,|Z,U$ and by the data processing inequality,

$$I(Y_x;X|Z) \leq H(U|Z) \leq \theta \tag{7}$$

So we have $I(Y_x;Z,X) = \sum_{ijk} b_{ijk}P(z_k)\log\left(\frac{b_{ijk}P(z_k)}{(\sum_{j',k'} b_{ij'k'}P(z_k))(\sum_{i'} b_{i'jk}P(z_k))}\right) \leq \theta + \phi$ in this case.

The same inequality holds for case 2, where the Assumption 2.2 holds. In that case, we have $\phi = 0$ since there is no direct path between $Z$ and $Y$. So we have $I(Y_x;Z,X) \leq \theta$

For case 3, where the Assumption 2.3 holds, we have $Z \perp\!\!\!\perp U$. So the inequality in Equation (7) becomes

$$I(Y_x;X|z) \leq H(U|z) = H(U) \leq \theta \quad \forall z.$$

If in addition, the Assumption 2.2 hold, we have $0 \leq I(Y_x;Z) \leq \phi = 0$. So, we have shown in each case of the IV graph that there exists a solution to our optimization problem in the feasible space, so the bounds are valid.

## I. Additional Case Study Experiment

To examine the performance of our IV selection method in real-world data, we take the two datasets that have been widely used for IV literature and select IV using our method.

For the Return of Schooling dataset (Angrist & Krueger, 1991), researchers interested in the causal effect of years of schooling $(X)$ on the wage level $(Y)$. We select the IV from two candidates: quarter of birth $(Z)$ and year of birth $(V)$. We discretize the variables and estimate the graph-specific conditional common entropy $\text{CCE}_{\mathcal{IV}}(Z;Y|X,V)$ and $\text{CCE}_{\mathcal{IV}}(V;Y|X,Z)$.

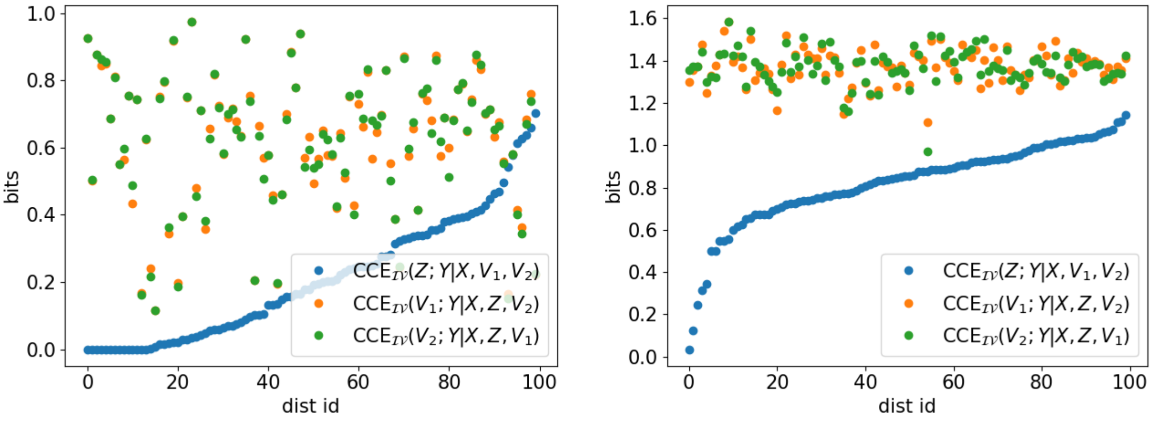
Similarly, for the Colonial Origins of Economic Development dataset (Acemoglu et al., 2001), for the causal effect of colonial institutions $(X)$ to the GDP level $(Y)$, we compare two IV candidates: mortality $(Z)$ and latitude $(V)$. The result is summarized in Table 2. For both datasets, our results conform with the studies in the IV literature. We apply a rather trivial procedure to discretize the data. In practice, one could use more sophisticated methods such as entropy-based discretization to maximize the mutual information between discretized variables.

| | Return of Scholing | Colonial Origins |
|---|---|---|
| X | Years of Schooling | Proxy of Colonial Institutions |
| Y | Wage Level | GDP level |
| Z | Quarter of Birth | Latitude |
| $\text{CCE}_{\mathcal{IV}}(Z;Y\|X,V)$ | 0.820 | 1.318 |
| $\text{CCE}_{\mathcal{IV}}(V;Y\|X,Z)$ | 2.945 | 2.297 |

*Table 2.* Drug Effect Example

## J. Selecting IV From More Candidates

To examine our method of selecting IV in a more general setting, we consider two cases: one variable is valid IV among three candidates, and two variables are valid IV among three candidates. We generate the simulated data with a similar procedure as described in Appendix A. For the experiments with one valid IV $(Z)$ and two invalid IV $(V_1, V_2)$, we estimate the $\text{CCE}_{\mathcal{IV}}(Z;Y|X,V_1,V_2)$, $\text{CCE}_{\mathcal{IV}}(V_1;Y|X,Z,V_2)$, and $\text{CCE}_{\mathcal{IV}}(V_2;Y|X,Z,V_1)$. The results are shown in Figure 8.



(a) $|Z| = |V_1| = |V_2| = 2$

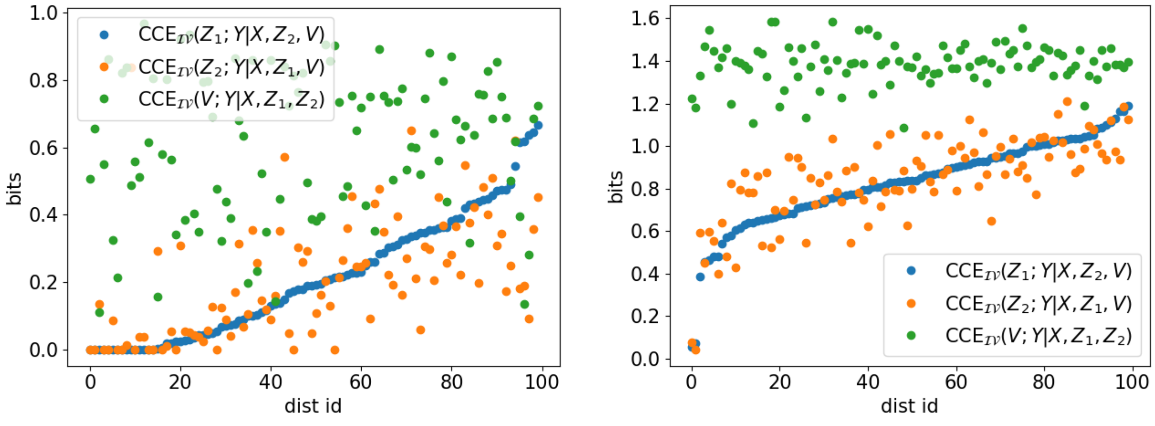
(b) $|Z| = |V_1| = |V_2| = 3$

*Figure 8.* Selecting one IV from three

For the experiments with two valid IV $(Z_1, Z_2)$ and one invalid IV $(V)$, we estimate the $\text{CCE}_{\mathcal{IV}}(Z_1;Y|X,Z_2,V)$, $\text{CCE}_{\mathcal{IV}}(Z_2;Y|X,Z_1,V)$, and $\text{CCE}_{\mathcal{IV}}(V;Y|X,Z_1,Z_2)$. The results are shown in Figure 9.

In both experiments, the results show that our method identifies the valid IV with high probability. In general, this method can be used as a confidence score when comparing IV candidates.

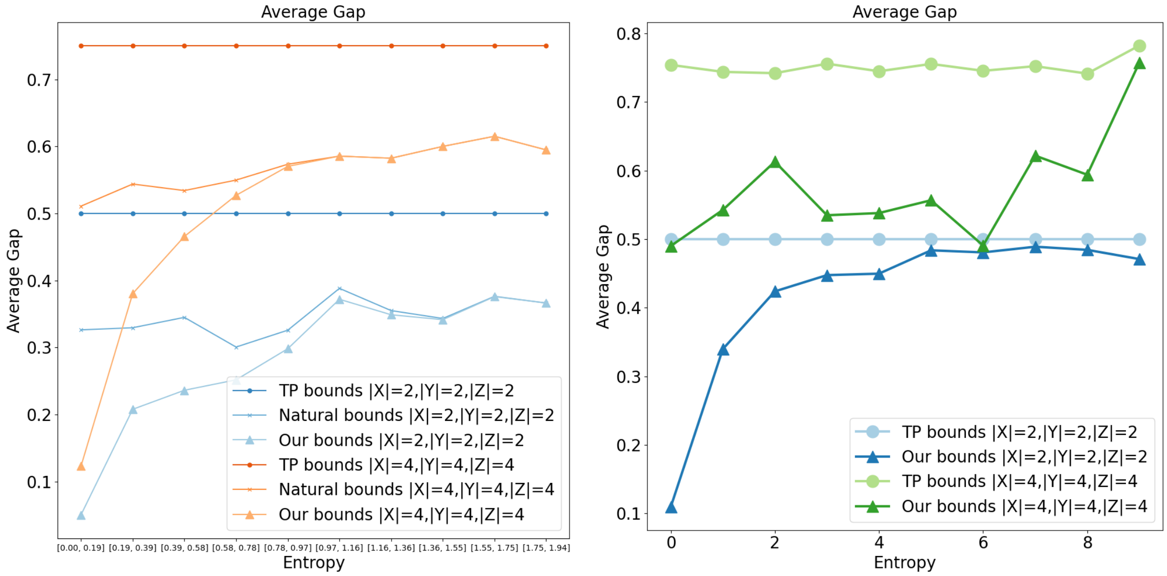
## K. Additional Results



(a) $|Z| = |V_1| = |V_2| = 2$

(b) $|Z| = |V_1| = |V_2| = 3$

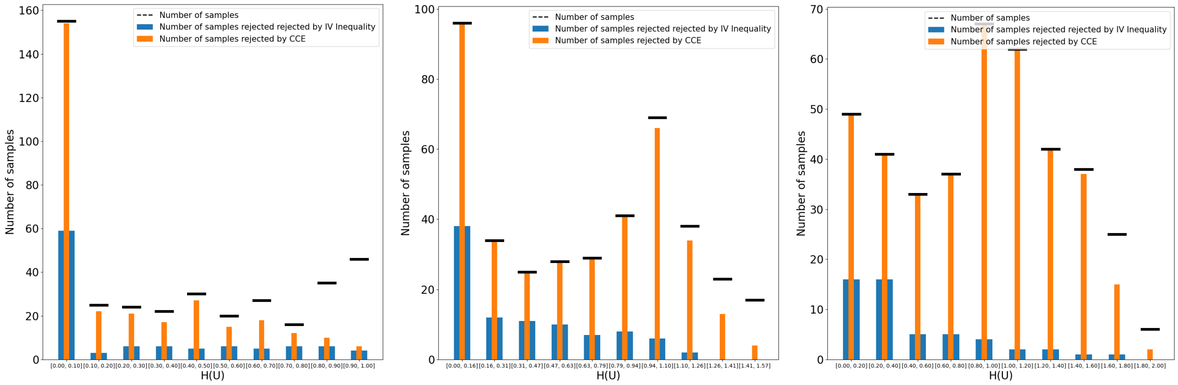
*Figure 9.* Selecting two IVs from three



(a) IV graph $\mathcal{IV}$

(b) Invalid IV graph $\mathcal{D}$

*Figure 10.* Average gaps between bounds.



(a) $|X| = |Y| = |Z| = 2$ (b) $|X| = |Y| = |Z| = 3$ (c) $|X| = |Y| = |Z| = 4$

*Figure 11.* Invalid IV rejection with CCE