Entropic Inequality Constraints from e-separation Relations in Directed Acyclic Graphs with Hidden Variables

Noam Finkelstein[1] **Beata Zjawin**[2, 3] **Elie Wolfe**[2] **Ilya Shpitser**[1] **Robert W. Spekkens**[2]

[1] Johns Hopkins University, Department of Computer Science, 3400 N Charles St, Baltimore, MD USA, 21218
[2] Perimeter Institute for Theoretical Physics, 31 Caroline St. N, Waterloo, Ontario, Canada, N2L 2Y5
[3] International Centre for Theory of Quantum Technologies, University of, Gdańsk, 80-308 Gdańsk, Poland

Abstract

Directed acyclic graphs (DAGs) with hidden variables are often used to characterize causal relations between variables in a system. When some variables are unobserved, DAGs imply a notoriously complicated set of constraints on the distribution of observed variables. In this work, we present entropic inequality constraints that are implied by e-separation relations in hidden variable DAGs with discrete observed variables. The constraints can intuitively be understood to follow from the fact that the capacity of variables along a causal pathway to convey information is restricted by their entropy; e.g. at the extreme case, a variable with entropy 0 can convey no information. We show how these constraints can be used to learn about the true causal model from an observed data distribution. In addition, we propose a measure of causal influence called the minimal mediary entropy, and demonstrate that it can augment traditional measures such as the average causal effect.

1 INTRODUCTION

A causal model of a system of random variables can be represented as a directed acyclic graph (DAG), where an edge from a node X to a node Y can be taken to mean that the random variable X is a direct cause of the random variable Y. Such causal models can be used to algorithmically deduce highly non-obvious properties of the system. For example, it is possible to deduce that the probability distribution of observed variables in the system, called the *observed data distribution*, must satisfy certain *constraints*.

When some variables in the system are unobserved, the constraints implied by the causal model are not well understood, and, for computational reasons, cannot be feasibly enumerated in full for arbitrary graphs. As a result, a number of methods have been developed for quickly providing a subset of these constraints [Wolfe et al., 2019, Kang and Tian, 2006, Poderini et al., 2019]. In this work, we contribute to this literature by describing entropic inequality constraints that hold whenever an e-separation relationship [Evans, 2012, Pienaar, 2017] is present in the graph.

The idea underlying these inequality constraints is that mutual information between two variables in a graphical model must be explained by variability of variables (termed bottleneck variables) that are between them along some path. Such paths need not be directed; a bottleneck variable may constitute the base of a fork structure or the mediary variable in a chain structure along the path. Each such path has a limited capacity for carrying information, which can be quantified in terms of the entropies of the bottleneck variables on that path. At the extreme case, if there is a bottleneck variable along a path with zero entropy, then subsequent variables on that path cannot learn about prior variables through the path, because the bottleneck variable will hold a fixed valueregardless of the values taken other any other variables, observed or unobserved. We will quantitatively relate the amount of information that can flow through a path to the entropies of its bottleneck variables below.

Constraints on the observed data distribution implied by a causal model have primarily been used to determine whether the observed data is *compatible* with a causal model, and to learn the true causal model directly from the observed data. Existing algorithms for learning causal models rely primarily on equality constraints. We suggest that incorporating our proposed inequality constraints, which can easily be read off a graphical model, can meaningfully improve these methods. In addition, we show how the entropy of latent variables can be linked to properties of the observed data distribution, yielding bounds on latent variable entropies or constraints on the observed data distribution.

We also demonstrate that our constraints can be used to bound an intuitive measure of the *strength* of a causal relationship between two variables, called the *Minimum Medi-*

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:1045–1055.

ary Entropy (MME). We show that the standard measure, called the *Average Causal Effect* (ACE), is not well suited to capturing the causal influence strength of a non-binary treatment on outcome, and can be misleading in some settings. For example, the ACE can be 0 even when treatment changes outcome for every subject in the population. The MME overcomes both of these issues, and can serve as an informative complement to the ACE.

The remainder of the paper is organized as follows. In Section 2, we discuss relevant material in causal inference and information theory. We present our constraints in Section 3, and several applications of the constraints in Section 4. Finally, a discussion of related work and directions for future study can be found in Section 5 and Section 6 respectively.

2 PRELIMINARIES

2.1 CAUSAL INFERENCE BACKGROUND

We begin by introducing key ideas from the literature on graphical causal models. Suppose we are interested in a system of related phenomena, each of which can be represented by a random variable. We denote observed variables in the system as $\mathbf{Y}$, unobserved variables as $\mathbf{U}$, and the full set of variables as $\mathbf{V} \equiv \mathbf{Y} \cup \mathbf{U}$.

We let $\mathcal{G}$ denote a DAG representing the system of interest. Each node in $\mathcal{G}$ corresponds to a variable in $\mathbf{V}$. The direct causes of each random variable V are defined to be its parents in $\mathcal{G}$, denoted $pa_{\mathcal{G}}(V)$. We adopt a nonparametric structural equations view of the DAG [Pearl, 2009, Richardson and Robins, 2013], under which the value of each variable V is a function of its direct causes and exogenous noise, denoted ϵ_V. The set of these structural equations is denoted $\mathcal{F} \equiv \{f_V(pa_{\mathcal{G}}(V), \epsilon_V) \mid V \in \mathbf{V}\}$. In most causal analyses, the exact form of these functions is unknown. Nevertheless, if the structure of causal dependencies in a system is known to be summarized by a graph $\mathcal{G}$, or, equivalently, to be described by some set of functions $\mathcal{F}$, then the distribution $P(\mathbf{V})$ is know to factorize as

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V \mid pa_{\mathcal{G}}(V)). \tag{1}$$

Equation (1) is the fundamental constraint that $\mathcal{G}$ places on the distribution $P(\mathbf{V})$ – if the equality holds, then the distribution is in the model; otherwise it is not. When all variables are observed, each term in the factorization is identifiable from observed data, and the constraint may easily be checked. When not all variables are observed, there is no known polynomial-time algorithm for expressing the constraints that the factorization of the full joint distribution places on the observed data distribution. In theory, necessary and sufficient conditions for the observed data distribution to be in the model can be obtained through the use of quantifier elimination algorithms [Geiger and Meek, 1999], but these have doubly exponential runtime and are prohibitively slow in practice.

We now review d-separation and e-separation, which are properties of the graph $\mathcal{G}$ that imply certain properties of distribution $P(\mathbf{V})$. We first introduce the notion of open and closed paths in conditional distributions. Triples in the graph of the form $A \to C \to B$ and $A \leftarrow C \to B$ are said to be open if we do not condition on C, and closed if we do condition on C. Triples of the form $A \to C \leftarrow B$, in which C is called a collider, are closed if we do not condition on C or any of its descendants, and open if we do. A path is said to be open under a conditioning set $\mathbf{C}$ if all contiguous triples along that path are open under that conditioning set.

Definition 1 (d-separation)**.** *Let* $\mathbf{A}$*,* $\mathbf{B}$ *and* $\mathbf{C}$ *be sets of variables in a DAG.* $\mathbf{A}$ *and* $\mathbf{B}$ *are said to be* d*-separated by* $\mathbf{C}$ *if all paths between* $\mathbf{A}$ *and* $\mathbf{B}$ *are closed after conditioning on* $\mathbf{C}$*. This is denoted* $(\mathbf{A} \perp_d \mathbf{B} \mid \mathbf{C})$.

It is a well-known consequence of Equation (1) that any d-separation relation $(\mathbf{A} \perp_d \mathbf{B} \mid \mathbf{C})$ in $\mathcal{G}$ implies the corresponding conditional independence relation $\mathbf{A} \perp \mathbf{B} \mid \mathbf{C}$ in the distribution $P(\mathbf{V})$. Conditional independence constraints of this form are about sub-populations in which the variables in $\mathbf{C}$ take the same value for all subjects. We can only evaluate whether these constraints hold when all variables in $\mathbf{C}$ are observed; otherwise there is no way to identify the relevant sub-populations. For that reason, it is impossible to determine whether conditional independences implied by $\mathcal{G}$ hold if they have hidden variables in their conditioning sets, leading to the need for other mechanisms to test implications of these independencies.

To describe e-separation, we first introduce the idea that a node can be *deleted* from a graph by removing the node and all of its incoming and outgoing edges. e-separation can then be defined as follows.

Definition 2 (e-separation)**.** *Let* $\mathbf{A}$*,* $\mathbf{B}$*,* $\mathbf{C}$ *and* $\mathbf{D}$ *be sets of variables in a DAG.* $\mathbf{A}$ *and* $\mathbf{B}$ *are said to be* e*-separated by* $\mathbf{C}$ *after deletion of* $\mathbf{D}$ *if* $(\mathbf{A} \perp_d \mathbf{B} \mid \mathbf{C})$ *after deletion of every variable in* $\mathbf{D}$*. This is denoted* $(\mathbf{A} \perp_e \mathbf{B} \mid \mathbf{C}$ *upon* $\neg\mathbf{D})$.

Conditioning on $\mathbf{C}$ may close some paths between $\mathbf{A}$ and $\mathbf{B}$, and open others. In the context of e-separation, the set $\mathbf{D}$, which we refer to as a *bottleneck* for $\mathbf{A}$ and $\mathbf{B}$ conditional on $\mathbf{C}$, is any set that includes at least one variable from each path between $\mathbf{A}$ and $\mathbf{B}$ that is open after conditioning on $\mathbf{C}$. If no subset of $\mathbf{D}$ is a bottleneck, then $\mathbf{D}$ is called a *minimal* bottleneck. This terminology reflects the fact that, conditional on $\mathbf{C}$, all information shared between $\mathbf{A}$ and $\mathbf{B}$ – that is, transferred from one to the other or transferred to each from a common source – must flow through $\mathbf{D}$.

It has been shown that every e-separation relationship among observed variables in a graph $\mathcal{G}$ corresponds to a constraint on the observed data distribution $P(\mathbf{Y})$ [Evans,

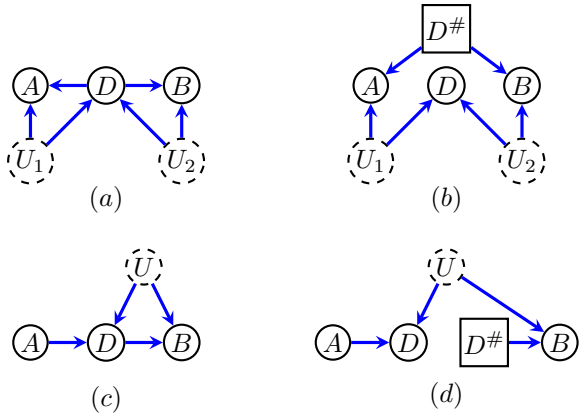

Figure 1: The Unrelated Confounders graph (a), and a split node model for it (b), as well as the Instrumental graph (c), and its split node model (d).

2012]. However, this result is not constructive, in the sense that it does not provide a strategy for deriving such constraints for a given e-separation relationship. The inequality constraints we provide in Section 3 partially fulfill this role; they provide explicit constraints that hold everywhere in the model whenever an e-separation relationship obtains in a graph.

2.1.1 Node Splitting

We will see that e-separation is related to the idea of splitting nodes in a graph. We define a node-splitting operation as follows. Given a graph $\mathcal{G}$ and a vertex D in the graph, the node splitting operation returns a new graph $\mathcal{G}^\#$ in which D is split into two vertices. One of the vertices is still called D, and it maintains all edges directed into D in the original graph $\mathcal{G}$, but none of its outgoing edges. This vertex keeps the name D because it will have the same distribution as D in the original graph, as all of its causal parents remain the same. The second random variable is labeled $D^\#$, and it inherits all of the edges outgoing from D in the original graph, but none of its incoming edges. Examples of the node splitting operation are illustrated in Fig. 1.

By a result in [Evans, 2012], $(\mathbf{A} \perp_e \mathbf{B} \mid \mathbf{C} \text{ upon } \neg\mathbf{D})$ in $\mathcal{G}$ if and only if $(\mathbf{A} \perp_d \mathbf{B} \mid \mathbf{C}, \mathbf{D}^\#)$ in $\mathcal{G}^\#$. Note that the node splitting operation described here is closely related to the operation of node splitting in Single World Intervention Graphs in causal inference [Richardson and Robins, 2013].

2.2 ENTROPIES

In this section, we review standard concepts in information theory, which we will use to express our inequality constraints. We begin with the definitions of entropy and mutual information.

Definition 3. *The **entropy** of a random variable X is defined as $H(X) \equiv -\sum_{x\in\mathcal{X}} P(x)\log_2 P(x)$, with the joint entropy of X and Y defined analogously. The **mutual information** between X and Y is defined as $I(X:Y) \equiv H(X) + H(Y) - H(X,Y)$.*

The entropy of a random variable can be thought of as the level of uncertainty one has about its value. Entropy is maximized by a uniform distribution over the domain of a random variable, as there is no reason to think any one value is more probable than another, and minimized by a point distribution, in which there is no uncertainty.

The mutual information between X and Y can be thought of as the amount of certainty we gain about the value of one, on average, if we learn the value of the other. It is maximized when one of X and Y is a deterministic function of the other, and is minimized when they are independent.

The entropy $H(X \mid Y{=}y)$ of X conditional on a specific value of $Y{=}y$ is obtained by replacing the distribution $P(X)$ in Definition 3 with $P(X \mid Y{=}y)$. The **conditional entropy** of X given Y, denoted $H(X \mid Y)$, is defined as the expected value of $H(X \mid Y{=}y)$. Conditional mutual information is analogously defined.

3 e-SEPARATION CONSTRAINTS

We have already described the intuition behind our constraints, which can be roughly summarized by the observation that the statistical dependence between random variables must be limited by the total amount of information that can flow through any bottleneck between them. We now describe how the tools introduced in Section 2 help us formalize this intuition.

First, we describe why e-separation helps formalize the idea of blocking "all paths" between two sets of variables. Consider the instrumental variable graph, depicted in Fig. 1(c). A and B are only d-separated by the set $\{D, U\}$, where U is unobserved. Consequently, they are not d-separated by any set consisting entirely of observed variables. They are, however, e-separated after deletion of the observed variable D. This tells us that all paths between A and B are through D, and we can take advantage of observed properties of D to bound the dependence between them even when nothing is known about the unobserved variable U. A similar story can be told about the Unrelated Confounders scenario depicted in Fig. 1(a).

When all variables are observed, e-separation does not imply any constraints that are not implied by d-separation, which follows from the fact that d-separation implies all constraints in such cases [Pearl, 1988]. However, as illustrated by the examples in Figs. 1(a) and 1(c), e-separation allows us to identify bottlenecks consisting entirely of observed variables between A and B even when paths between A

and B cannot be closed by *any* manner of conditioning on observed variables. To show how e-separation lead to entropic constraints, we will make use of Theorem 4.2 in [Evans, 2012], reframed as follows.

Theorem 4. *([Evans, 2012] Theorem 4.2) Suppose* $(\mathbf{A} \perp_e \mathbf{B} \mid \mathbf{C}$ *upon* $\neg\mathbf{D})$ *in* $\mathcal{G}$*, and that no variable in* $\mathbf{C}$ *is a descendant of any in* $\mathbf{D}$*. Then there exists a distribution* P^* *over* $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{D}^{\#}$ *such that*

$$\begin{aligned} &P(\mathbf{A}{=}\mathbf{a}, \mathbf{B}{=}\mathbf{b}, \mathbf{D}{=}\mathbf{d} \mid \mathbf{C}{=}\mathbf{c}) \\ &= P^*(\mathbf{A}{=}\mathbf{a}, \mathbf{B}{=}\mathbf{b}, \mathbf{D}{=}\mathbf{d} \mid \mathbf{C}{=}\mathbf{c}, \mathbf{D}^{\#}{=}\mathbf{d}) \end{aligned} \tag{2}$$

with $\mathbf{A} \perp \mathbf{B} \mid \mathbf{C}, \mathbf{D}^{\#}$ *in* P^**. If furthermore no variable in* $\mathbf{A}$ *is a descendant of any in* $\mathbf{D}$*, then there exists a distribution* P^* *such that* $P(\mathbf{B}{=}\mathbf{b}, \mathbf{D}{=}\mathbf{d} \mid \mathbf{A}{=}\mathbf{a}, \mathbf{C}{=}\mathbf{c}) = P^*(\mathbf{B}{=}\mathbf{b}, \mathbf{D}{=}\mathbf{d} \mid \mathbf{A}{=}\mathbf{a}, \mathbf{C}{=}\mathbf{c}, \mathbf{D}^{\#}{=}\mathbf{d})$ *with* $\mathbf{A} \perp \mathbf{B} \mid \mathbf{C}, \mathbf{D}^{\#}$ *in* P^**.* [1]

We provide the following intuition for this theorem. Our graph $\mathcal{G}$ represents the causal relationships within a system of random variables in the real world. The graph $\mathcal{G}^{\#}$ represents an alternative world in which the causal effects of $\mathbf{D}$ are "spoofed" by some random variable $\mathbf{D}^{\#}$. That is, children of $\mathbf{D}$ in $\mathcal{G}$, which should be functions of $\mathbf{D}$, are instead fooled into being functions of $\mathbf{D}^{\#}$.

In the alternative world represented by $\mathcal{G}^{\#}$, we suppose that the functional form f_V of a variable V in terms of its parents stays the same for all variables that are shared between graphs. This means that all non-descendants of $\mathbf{D}$ have the same joint distribution in our world and in the alternative world, as neither their parents nor the functions defining them in terms of their parents have changed. By contrast, descendants of $\mathbf{D}$ in $\mathcal{G}$ will have a different distribution in the alternative world, as their distributions are now functions of the distribution of $\mathbf{D}^{\#}$, which may be different from that of $\mathbf{D}$, and is unknown.

Now, suppose we condition on a particular value of $\mathbf{D}^{\#}{=}\mathbf{d}$ in $\mathcal{G}^{\#}$. Then, because the functional form of the causal mechanisms is shared across worlds, the descendants of $\mathbf{D}$ in $\mathcal{G}$ have the same distribution as they have when $\mathbf{D}{=}\mathbf{d}$ in the observed world. In addition, all of the non-descendants of $\mathbf{D}^{\#}$ are marginally independent from $\mathbf{D}^{\#}$, because it has no ancestors so all connecting paths must be collider paths. Therefore, both its non-descendants and its descendants have the same joint distribution they would have had when $\mathbf{D}{=}\mathbf{d}$ in the original graph. The results in the theorem then follow when we note that $\mathbf{C}$, and optionally $\mathbf{A}$, are non-descendants of $\mathbf{D}$, and that the relevant independence properties hold in the world of $\mathcal{G}^{\#}$.

In general, we cannot know what this P^* distribution is, because we never get to observe this alternate world. But when we condition on $\mathbf{D}^{\#}$, we are removing precisely the randomness we do not know about, yielding a distribution that we do know about. The fact that P^* agrees with P on a subset of their domains, and that it contains known independences, is sufficient to derive informative constraints, as seen below.

[1] In causal inference problems, a distribution P^* that satisfies the relevant conditions for this result may be constructed from counterfactual random variables $\mathbf{A}(\mathbf{d}), \mathbf{B}(\mathbf{d}), \mathbf{D}(\mathbf{d})$ and $\mathbf{C}(\mathbf{d})$.

3.1 ENTROPIC CONSTRAINTS FROM e-SEPARATION

We now show how the notion of e-separation permits the formulation of entropic inequality constraints. In these constraints, we use mutual information to represent dependence between sets of variables, and entropy to measure the information-carrying capacity of paths connecting them.

Theorem 5. *(Proof in Supplementary Materials.) Suppose observed variables are discrete. If* $(\mathbf{A} \perp_e \mathbf{B} \mid \mathbf{C}$ *upon* $\neg\mathbf{D})$ *and no element of* $\mathbf{C}$ *is a descendant of any in* $\mathbf{D}$*, then for any value* $\mathbf{c}$ *in the domain of* $\mathbf{C}$*, the following constraints hold:*

$$I(\mathbf{A} : \mathbf{B} \mid \mathbf{C}{=}\mathbf{c}, \mathbf{D}) \leq H(\mathbf{D} \mid \mathbf{C}{=}\mathbf{c}), \tag{3a}$$
$$I(\mathbf{A} : \mathbf{B} \mid \mathbf{C}, \mathbf{D}) \leq H(\mathbf{D} \mid \mathbf{C}). \tag{3b}$$

If in addition, no element of $\mathbf{A}$ *is a descendant of any in* $\mathbf{D}$*, then for any value* $\mathbf{c}$ *in the domain of* $\mathbf{C}$*, the following stronger constraints hold:*

$$I(\mathbf{A} : \mathbf{B}, \mathbf{D} \mid \mathbf{C}{=}\mathbf{c}) \leq H(\mathbf{D} \mid \mathbf{C}{=}\mathbf{c}), \tag{4a}$$
$$I(\mathbf{A} : \mathbf{B}, \mathbf{D} \mid \mathbf{C}) \leq H(\mathbf{D} \mid \mathbf{C}). \tag{4b}$$

This theorem potentially allows us to efficiently discover many entropic inequalities implied by any given graph, such as those implied by Fig. 2. In some cases, as in Fig. 2(a), the theorem recovers *all* Shannon-type entropic inequality constraints implied by the graph [Chaves et al., 2014, Chaves et al., 2014, Weilenmann and Colbeck, 2017]. In other cases, as in Fig. 2(b), the graph implies a Shannon-type entropic inequality constraint beyond what Theorem 5 can recover, per a result in [Weilenmann and Colbeck, 2020]. Indeed, entropic inequality constraints can be implied by graphs not exhibiting e-separation relations whatsoever, such as the triangle scenario [Steudel and Ay, 2015, Chaves et al., 2014].

The linear quantifier elimination of [Chaves et al., 2014, Chaves et al., 2014, Weilenmann and Colbeck, 2017] will always discover all the entropic inequalities which can be inferred from Theorem 5. However, the quantifier elimination method is computationally expensive, and is essentially intractable for graphs involving more than six or seven variables (observed and latent combined). Theorem 5, by contrast, provides an approach that is computationally tractable, but is capable of discovering fewer entropic constraints.

We describe the strategy used to obtain the results in Theorem 5 at a high level. First, we express some function of

the observed data distribution $g(P)$ as a sum over the domain of $\mathbf{D}$ of some non-negative function $f(P^*)$, such that $g(P) = \sum_{\mathbf{d}\in\mathcal{D}} f(P^*)$. We then find a pair of functions f_A and f_D, such that $f(P^*) \leq f_A(P^*) - f_D(P^*)$, with $f_A(P^*) = f_A(P)$ and $f_D \geq 0$. Finding this pair of functions is the only non-prescriptive part of this proof strategy. For the resulting constraint to be non-trivial, the decomposition must make use of the independence properties of P^*. We substitute these functions into $g(P) = \sum_{\mathbf{d}\in\mathcal{D}} f(P^*)$ to obtain the inequality $g(P) \leq \sum_{\mathbf{d}\in\mathcal{D}} f_A(P^*) - f_D(P^*)$. Because f_D is non-negative, it can be dropped from the expression to yield $g(P) \leq \sum_{\mathbf{d}\in\mathcal{D}} f_A(P^*)$. By construction, $f_A(P^*) = f_A(P)$, which can be used to produce an inequality constraint between two functions of the observed data distribution: $g(P) \leq \sum_{\mathbf{d}\in\mathcal{D}} f_A(P)$.

Our proof of Theorem 5, which is deferred to the appendix, demonstrates how f_A and f_D may be found for any function f that obeys subadditivity and the additive chain rule, such as the Shannon entropy, as used in the proposition.

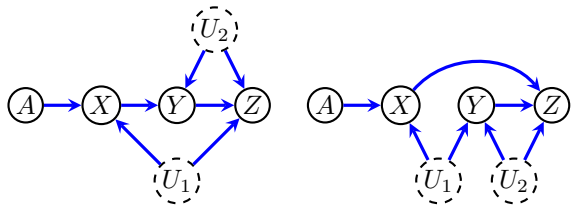

Figure 2: For graph (a), Theorem 5 implies the entropic inequality constrains $I(A:XYZ) \leq H(X)$ and $I(A:YZ) \leq H(Y)$. For graph (b), Theorem 5 implies $I(A:XYZ) \leq H(X)$ and $I(A:YZ|X) \leq H(Y|X)$. Note, however, that the entropic quantifier elimination method of Chaves et al. [2014] as applied by Weilenmann and Colbeck [2020], finds that the former inequality for graph (b) can be strengthened into $I(A:XYZ) \leq H(X|Y)$.

3.2 RELATING e-SEPARATION TO EQUALITY CONSTRAINTS

We have seen that d-separation and e-separation relations imply constraints on the observed data distribution. Verma and Pearl [1990] discuss equality constraints for latent variable models that apply in identified post-intervention distributions. Such equality constraints are sometimes called Verma constraints. A general description of the class of these constraints implied by a hidden variable DAG model, as well as discussion of properties of these constraints is given in [Richardson et al., 2017]. In this section, we examine the relationship between the e-separation-based constraints of Theorem 5 and the d-separation-based conditional independence and Verma constraints.

First, we observe that the presence of d-separation relations and Verma constraints in a graphical model imply the presence of an e-separation relation.

Proposition 6. *(Proof in Supplementary Materials.) If* $\mathbf{A}$ *is* d*-separated from* $\mathbf{B}$ *by* $\{\mathbf{C}, \mathbf{D}\}$*, then* $\mathbf{A}$ *is also* e*-separated from* $\mathbf{B}$ *by* $\mathbf{C}$ *upon deleting* $\mathbf{D}$*.*

This demonstrates that for any d-separation relation in the graph, it is possible to obtain an entropic constraint corresponding to any minimal bottleneck $\mathbf{D}$ through an e-separation relation. More precisely, when $\mathbf{A}$ is d-separated from $\mathbf{B}$ by $\{\mathbf{C}, \mathbf{D}\}$, then by Proposition 6, it is also the case that $\mathbf{A}$ is e-separated from $\mathbf{B}$ given $\mathbf{C}$ upon deleting $\mathbf{D}$, and therefore Theorem 5 can be applied to obtain entropic constraints. Note, however, that these are necessarily weaker than the entropic constraint $I(\mathbf{A}:\mathbf{B} \mid \mathbf{C}, \mathbf{D}) = 0$, which follows from the d-separation relation itself.

In summary, every d-separation relation in the graph is an instance of e-separation, but not vice-versa. When an instance of e-separation is also an instance of d-separation, then all the inequality constraints implied by e-separation are rendered defunct by the stronger equality constraints implied by d-separation.

We now show that a similar pattern of deprecating inequalities by equalities occurs in the presence of Verma constraints when certain counterfactual interventions are identifiable.

Proposition 7. *Consider a graph* $\mathcal{G}$ *which exhibits the* e*-separation relation* $(\mathbf{A} \perp_e \mathbf{B} \mid \mathbf{C}$ *upon* $\neg\mathbf{D})$ *and where no element of* $\mathbf{C}$ *is a descendant of any in* $\mathbf{D}$*. If the counterfactual distribution* $P(\mathbf{A}(\mathbf{D}{=}\mathbf{d}), \mathbf{B}(\mathbf{D}{=}\mathbf{d}), \mathbf{D}(\mathbf{D}{=}\mathbf{d}) \mid \mathbf{C})$ *is identifiable*[2] *then the inequalities of Theorem 5 are logically implied whenever the stronger equality constraints*

$$I\left(\mathbf{A}(\mathbf{D}{=}\mathbf{d}) : \mathbf{B}(\mathbf{D}{=}\mathbf{d}) \mid \mathbf{C}\right) = 0 \tag{5}$$

are satisfied for all values of $\mathbf{d}$*. Note that Equation* (5) *is satisfied if and only if the* margin *of the identified counterfactual distribution factorizes, i.e., when*

$$\begin{aligned} &P(\mathbf{A}(\mathbf{D}{=}\mathbf{d}), \mathbf{B}(\mathbf{D}{=}\mathbf{d}) \mid \mathbf{C}) \\ &\equiv \textstyle\sum_{\mathbf{d}'} P(\mathbf{A}(\mathbf{D}{=}\mathbf{d}), \mathbf{B}(\mathbf{D}{=}\mathbf{d}), \mathbf{D}(\mathbf{D}{=}\mathbf{d}){=}\mathbf{d}' \mid \mathbf{C}) \\ &\quad \textit{exhibits} \quad \mathbf{A}(\mathbf{D}{=}\mathbf{d}) \perp \mathbf{B}(\mathbf{D}{=}\mathbf{d}) \mid \mathbf{C}. \end{aligned} \tag{6}$$

The proof directly follows from that of Theorem 5. In proving Theorem 5, we derive entropic inequalities by relating the entropies pertaining to $P(\mathbf{A}, \mathbf{B}, \mathbf{D} \mid \mathbf{C})$ to entropies pertaining to the P^* distribution posited by Theorem 4. That is, Theorem 5 is an entropic consequence of Theorem 4. If

[2]The counterfactual distribution in this theorem allows intervened-on variables and outcomes to intersect. See [Shpitser et al., 2021] for a complete identification algorithm for counterfactual distributions of this type.

the conditions of Proposition 7 are satisfied, then the conditions of Theorem 4 are also automatically satisfied since one can then *explicitly* reconstruct

$$P^*(\mathbf{A}, \mathbf{B}, \mathbf{D}{=}\mathbf{d} \mid \mathbf{C}, \mathbf{D}^{\#}{=}\mathbf{d}^{\#}) \\ = P(\mathbf{A}(\mathbf{D}{=}\mathbf{d}^{\#}), \mathbf{B}(\mathbf{D}{=}\mathbf{d}^{\#}), \mathbf{D}(\mathbf{D}{=}\mathbf{d}^{\#}){=}\mathbf{d} \mid \mathbf{C}). \quad (7)$$

There is no opportunity to violate the entropic inequalities of Theorem 5 once the observational data has been confirmed as consistent with Theorem 4. In other words, in order to violate the inequalities of Theorem 5 it must be the case that no P^* consistent with Theorem 4 can be constructed, but this contradicts the explicit recipe of Equation (7).

See [Verma and Pearl, 1990, Tian and Pearl, 2002, Richardson et al., 2017] for details on how to derive the form of the equality constraints summarized by Equation (6). We note here that $P(\mathbf{A}(\mathbf{D}{=}\mathbf{d}), \mathbf{B}(\mathbf{D}{=}\mathbf{d}), \mathbf{D}(\mathbf{D}{=}\mathbf{d}){=}\mathbf{d} \mid \mathbf{C})$ is certainly identifiable if $\mathbf{D}$ is not a member of the same *district* ([Richardson et al., 2017]) as any element in $\{\mathbf{A}, \mathbf{B}\}$ within the subgraph of $\mathcal{G}$ over $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ and their ancestors. We also note that the identifiability of *merely* $P(\mathbf{A}(\mathbf{D}{=}\mathbf{d}), \mathbf{B}(\mathbf{D}{=}\mathbf{d}) \mid \mathbf{C})$ but not of $P(\mathbf{A}(\mathbf{D}{=}\mathbf{d}), \mathbf{B}(\mathbf{D}{=}\mathbf{d}), \mathbf{D}(\mathbf{D} = \mathbf{d}){=}\mathbf{d} \mid \mathbf{C})$ negates the implication from Equation (6) to Theorem 5. In Appendix A, we provide an example of a graph in which $P(\mathbf{A}(\mathbf{D}{=}\mathbf{d}), \mathbf{B}(\mathbf{D}{=}\mathbf{d}) \mid \mathbf{C})$ is identified, but the entropic constraints of Theorem 5 remain relevant. In addition, we demonstrate that the application of the entropic constraints to identified counterfactual distributions can also result in inequality constraints on the observed data distribution.

3.3 CONSTRAINTS AND BOUNDS INVOLVING LATENT VARIABLES

In this section, we consider d-separation relations with hidden variables in the conditioning set. Because we cannot condition on hidden variables, there is no way to check whether the corresponding independence constraints hold in the full data distribution. However, if we have access to auxiliary information about these hidden variables – such as information about their entropy or their cardinality – it is possible to obtain inequality constraints on the observed data distribution.

Proposition 8. *(Proof in Supplementary Materials.) If* $(\mathbf{A} \perp_d \mathbf{B} \mid \mathbf{C}, \mathbf{U})$*, then the entropy of* $\mathbf{U}$ *may be lower-bounded,* $H(\mathbf{U}) \geq H(\mathbf{U} \mid \mathbf{C}) \geq I(\mathbf{A} : \mathbf{B} \mid \mathbf{C})$.

In many scenarios, we may have more information about the *cardinality* of a hidden variable than its entropy. We take the cardinality of a set of variables to be the product of the cardinalities of the variables in the set. An upper bound on the cardinality of $\mathbf{U}$ entails an upper bound on its entropy. As observed above, the entropy of a random variable is maximized when it takes a uniform distribution. If we let $|\mathbf{U}|$ denote the cardinality of $\mathbf{U}$, and recall that the entropy of a uniformly distributed variable with cardinality m is simply $\log_2(m)$, then $\log_2 |\mathbf{U}| \geq H(\mathbf{U})$. The next corollary then follows immediately from Proposition 8.

Corollary 8.1. *If* $(\mathbf{A} \perp_d \mathbf{B} \mid \mathbf{C}, \mathbf{U})$*, then the cardinality of* $\mathbf{U}$ *may be lower-bounded,* $|\mathbf{U}| \geq 2^{I(\mathbf{A}:\mathbf{B}|\mathbf{C})}$.

Finally, we note that both of these inequalities can also be used if we *do not* know anything about the properties of $\mathbf{U}$, but would like to infer lower bounds for its entropy and cardinality from the observed data. In Section 4.2, we will explore a scenario in genetics in which these bounds and constraints may be of use.

Remark 9. *Constraints given in Proposition 8 and Corollary 8.1 are stronger than can be obtained from the* e*-separation relation* $(\mathbf{A} \perp_e \mathbf{B} \mid \mathbf{C}$ *upon* $\neg\mathbf{U})$ *on its own.*

To demonstrate Remark 9, we consider a set of structural equations consistent with Fig. 1(a). Suppose that D takes the value 0 when $U_1 \neq U_2$, and the value 1 otherwise, and that A and B take the value 0 if D is 0, and values equal to U_1 and U_2 respectively if D is 1. It follows that A and B are always equal, and therefore $I(A : B) = H(A)$. Now, suppose that U_1 and U_2 only take values not equal to 0, and that there are at least two values that each takes with nonzero probability. It immediately follows that $H(D) < H(A)$, and therefore that $H(D) < I(A : B)$, as D and A by construction take the value 0 with the same probability, but there is strictly more entropy in the remainder of A's distribution because D is binary and A takes at least two other values with nonzero probability.

4 APPLICATIONS

In this section, we explore several applications of the constraints developed above. In Sections 4.1 and 4.2, we show how our results can be used to learn about causal models from observational data. In Section 4.3, we further leverage the importance of the entropy of variables along a causal pathway to posit a new measure of causal strength, and observe that this measure can be bounded by an application of Theorem 5.

4.1 CAUSAL DISCOVERY

In this section, we present an example in which two hidden variable DAGs with the same equality constraints present different entropic inequality constraints. The ability to distinguish between models that share equality constraints has the potential to advance the field of causal discovery, in which causal DAGs are learned directly from the observed data. Causal discovery algorithms for learning hidden variable DAGs currently do so using only equality constraints.

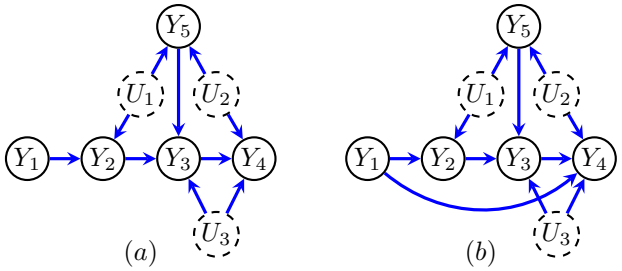

Figure 3: Two hidden variable DAGs that share equality constraints over observed variables, but (a) contains e-separation relations that are not in (b).

Our approach may be useful as a post-processing addition to such methods, whereby any graph found to satisfy the equality constraints in the observed data is tested against the entropic inequality constraints implied by e-separation relations in the model.

The hidden variable DAGs in Fig. 3, adapted from Appendix B in [Bhattacharya et al., 2020], share the same conditional independence constraints: $Y_1 \perp Y_3 \mid Y_2Y_5$ and $Y_1 \perp Y_5$, but exhibit different e-separation relations.

In Fig. 3(a), $(Y_1 \perp_e Y_3Y_4 \mid Y_2$ upon $\neg Y_5)$, $(Y_1Y_2 \perp_e Y_4 \mid$ upon $\neg Y_3)$, and $(Y_2 \perp_e Y_4 \mid Y_1$ upon $\neg Y_3)$. Applying Theorem 5 in each case, we obtain the three inequality constraints $I(Y_1 : Y_3Y_4Y_5 \mid Y_2) \leq H(Y_5 \mid Y_2)$, $I(Y_2 : Y_3Y_4 \mid Y_1) \leq H(Y_3 \mid Y_1)$, $I(Y_1Y_2 : Y_3Y_4) \leq H(Y_3)$.

In Fig. 3(b), we have added an edge, which removes some e-separation relations. We are left with $(Y_1 \perp_e Y_3 \mid Y_2$ upon $\neg Y_5)$, and $(Y_2 \perp_e Y_4 \mid Y_1$ upon $\neg Y_3)$. We can again apply Theorem 5 in each case, yielding the inequality constraints $I(Y_1 : Y_3Y_5 \mid Y_2) \leq H(Y_5 \mid Y_2)$ and $I(Y_2 : Y_3Y_4 \mid Y_1) \leq H(Y_3 \mid Y_1)$. The second of these constraints is shared by the graph in Fig. 3(a), and the first is strictly weaker than a constraint in Fig. 3(a).

Models similar to those shown in Fig. 3 sometimes arise in time-series data, where the variables in the chain represent observations taken at consecutive time steps. In such models, it is often assumed that treatments no longer have a direct effect on outcomes after a certain number of time steps. Here, that assumption is encoded in the lack of a direct edge from Y_1 to Y_4 in Fig. 3(a). We have shown above that this kind of assumption can be falsified even when it does not imply any additional equality constraints, as is often the case. In particular, if the stronger constraints implied by Fig. 3(a) are violated, but the weaker constraints of Fig. 3(b) are not, then the assumption is falsified.

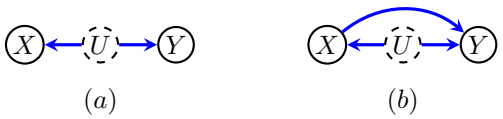

Figure 4: Identifying direct causal influence in the presence of a confounder with limited cardinality.

4.2 CAUSAL DISCOVERY IN THE PRESENCE OF LATENT VARIABLES

In this section, we consider a very simple possible application of the constraints and bounds relating to entropies of unobserved variables in genetics. Consider a causal hypothesis wherein the presence or absence of an unobserved gene influences two aspects of an organism's phenotype. Suppose that due to genetic sequencing studies, the number of variants of the gene in the population – i.e. the cardinality of the corresponding random variable – is known. Two possible hypotheses regarding the causal structure are depicted in Fig. 4, where U represents the gene and X and Y are the phenotype aspects. In Fig. 4(a), one presumes no causal influence of X on Y, whereas in Fig. 4(b), direct causal influence is allowed. In the former case, knowledge of the number of variants of the gene constrains the mutual information between the phenotypes, while in the latter case it is not constrained.

Thus, for certain types of statistical dependencies between X and Y, one can rule out the hypothesis of Fig. 4(a). For example, suppose we know the cardinality of U to be 3. Corollary 8.1 then implies the constraint that the mutual information between X and Y cannot exceed $\log_2(3) \approx 1.584$. Suppose further that we observe the distribution depicted in Table 1. The mutual information between X and Y in this distribution is ≈ 1.594. Because this mutual information violates the constraint implied by the model in Fig. 4(a), we know this model cannot be correct, and conclude that Fig. 4(b) is correct. More generally, strong statistical dependence between high cardinality variables cannot be explained by a low cardinality common cause and requires a direct influence between them.

		Y			
		0	1	2	3
X	0	0.002	0.001	0.400	0.001
	1	0.003	0.005	0.005	0.066
	2	0.224	0.003	0.003	0.001
	3	0.002	0.281	0.001	0.002

Table 1: An example joint distribution over two variables X and Y, each with cardinality 4.

Conversely, suppose Fig. 4(a) is known to be correct, and that there is no direct causal influence between the two aspects of phenotype. If the cardinality of U is not known, it can be bounded from below directly from observed data,

according to Corollary 8.1. In this case, the lower bound would be $2^{I(X:Y)} \approx 2^{1.594} \approx 3.018$. It follows that U must have a cardinality of 4 or above in this setting. The ability to extract such information from observational data may be useful in making substantive scientific decisions, or in guiding future sequencing studies.

In many applied data analyses, different variables may be observed for different subjects, i.e., data on some variables is "missing" for some subjects. A recent line of work has focused on properties of missing data models that can be represented as DAGs [Mohan et al., 2013]. Although the bounds and constraints above have been developed in the context of fully unobserved variables, they can also be used in missing data DAG models, for variables that are not observed for all subjects.

4.3 QUANTIFYING CAUSAL INFLUENCE

The traditional approach to measuring the strength of a causal relationship is by contrasting how different an outcome would be, on average, under two different treatments. Formally, if X is a cause of Y, the ACE is defined as $E[Y(X = x) - Y(X = x')]$. While the ACE is a very useful construct, we suggest that it has two important shortcomings, and present an alternative measure of causal strength called the *Minimal Mediary Entropy* or MME. The MME is based on the idea – explored throughout this work – that the entropy of variables along a causal pathway provide insight into the amount of information that can travel along that pathway.

In a scenario where treatment can be discerned to always cause outcome, we might expect the ACE, as a meaure of causal influence, to be large. The example below shows that this is not necessarily the case.

Example 1. Consider a randomized binary treatment X and a ternary outcome Y, with $P(Y{=}0 \mid X{=}0) = P(Y{=}2 \mid X{=}0) = 0.5$, and $P(Y{=}1 \mid X{=}1) = 1$. In this setting, $\mathsf{ACE} = 0$, even though treatment affects outcome for every subject in the population.

In less extreme settings, the ACE may be very small even when treatment affects outcome for almost every subject in the population, or very large, even when very few subjects have an outcome that is affected by treatment.

The ACE is likewise not always well suited to measuring the strength of a causal relationship when the treatment variable is non-binary. In such situations, no one causal contrast represents the causal influence, and the number of possible contrasts grows combinatorially in the cardinality of treatment. We now define the MME and discuss how it can overcome these issues.

Definition 10 (Minimal Mediary Entropy (MME))**.** *Given a DAG $\mathcal{G}$ containing a directed edge $X \to Y$, let $\mathcal{G}'_{X \to W \to Y}$ denote the graph constructed by substituting the single edge $X \to Y$ in $\mathcal{G}$ with the set of four edges $\{X \to W \to Y, W \leftarrow U_{WY} \to Y\}$, introducing auxilliary latent variables W and U_{WY}.* [3] *We then define $\mathsf{MME}_{X \to Y}$ as the smallest entropy $H(W)$ over all structural equations models reproducing the observed data distribution over $\mathcal{G}'_{X \to W \to Y}$ in which W has finite cardinality.*

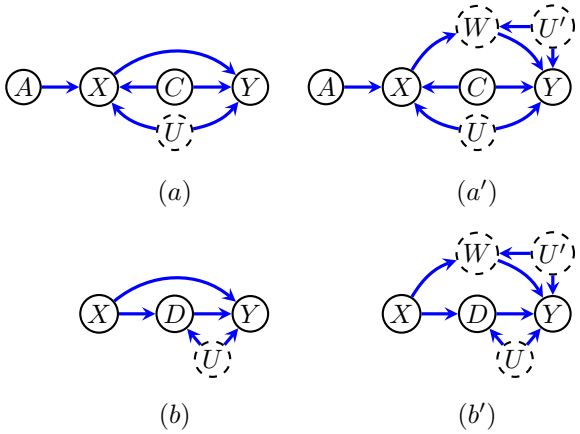

Figure 5: Modifying DAGs (a) and (b) by inserting a latent mediary W between X and Y yields DAGs (a') and (b') respectively. Note that in (a), even though X and Y are latent confounded, corollary 10.1 gives $\mathsf{MME}_{X \to Y} \geq I(A : Y|C{=}c)$ for any c by exploiting the fact that $A \in an(X)$. Also note that in (b), even though X affects Y both directly and indirectly though D, corollary 10.1 gives $\mathsf{MME}_{X \to Y} \geq I(X : Y|D) - H(D)$ for the direct effect.

Fig. 5 illustrates the process of edge substitution. Essentially, the edge $X \to Y$ in $\mathcal{G}$ is interrupted to pass through W in $\mathcal{G}'_{X \to W \to Y}$, such that the auxiliary latent variable W *fully mediates* the direct effect of X on Y. Note that caveat that MME is defined in terms of minimizing the entropy of W over *finite cardinality* W capable of reproducing the observed statistics. If W were allowed to be a continuously valued variable, then the observed data distribution would always be reproducible with arbitrarily small $H(W)$, due to the total lack of restriction in the instrumental model with a continuous mediary [Gunsilius, 2019].

With the presumption of finite cardinality W by fiat, however, we are are in a position to exploit Theorem 5 in order to practically lower bound the MME.

Corollary 10.1. *(Proof in Supplementary Materials.) Suppose that graphical construction $\mathcal{G}'_{X \to W \to Y}$ exhibits the e-separation relation $(\mathbf{A} \perp_e \mathbf{B} \mid \mathbf{C}$ upon $\neg\{\mathbf{D}, W\})$ for some $\mathbf{A} \subset \{X\} \cup an(X)$ and for some $\mathbf{B} \subset \{Y\} \cup desc(Y)$, where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are nonoverlapping subsets ($\mathbf{C}$ and*

[3] If a latent confounder were added between X and W, then although W would still mediate $X \to Y$, X and Y would share a source of unobserved confounding, altering the causal model.

$\mathbf{D}$ *possibly empty) of the observed variables in* $\mathcal{G}$*, and all the variables within* $\mathbf{D}$ *are discrete. Then*

$$\begin{aligned}
&\mathsf{MME}_{X\to Y} \\
&\quad \geq \max_{\mathbf{c}} \; I(\mathbf{A} : \mathbf{B} \mid \mathbf{C}{=}\mathbf{c}, \mathbf{D}) - H(\mathbf{D} \mid \mathbf{C}{=}\mathbf{c}) \qquad (8a) \\
&\quad \geq I(\mathbf{A} : \mathbf{B} \mid \mathbf{C}, \mathbf{D}) - H(\mathbf{D} \mid \mathbf{C}) \,. \qquad (8b)
\end{aligned}$$

Suppose that P_0 is a distribution in the model of the extended graph $\mathcal{G}'_{X\to W\to Y}$, such that P_0 marginalizes to the observed data distribution. Then the entropy $H(W)$ in P_0 is necessarily an upper bound on $\mathsf{MME}_{X\to Y}$, i.e. we have found a W with entropy $H(W)$ that fully mediates the causal influence of X on Y. Since W could always reproduce the observed data by simply copying the values of X, we have a trivial upper bound of $\mathsf{MME}_{X\to Y} \leq H(X)$. [4] This upper bound can typically be improved by even partially exploring the space of the distributions in $\mathcal{G}'_{X\to W\to Y}$.

Consider the simple model $X \to Y$ with $|X| = 3$ and $|Y| = 3$, and the observed data distribution $P(X{=}x, Y{=}y) = \begin{cases} \frac{5}{27} \text{ if } x = y \\ \frac{2}{27} \text{ if } x \neq y \end{cases}$. Our corollary gives us a lower bound on $\mathsf{MME}_{X\to Y} \geq I(X : Y) \approx 0.150$, contrasted with the trivial upper bound $\mathsf{MME}_{X\to Y} \leq H(X) \approx 1.585$. We can improve the trivial upper bounding by noting that this distribution can be reproduced by the following functional relationships $\begin{cases} W{=}0 \text{ and } Y{=}U_{WY} & \text{when } U_{WY}{=}X \\ W{=}1 \text{ and } Y{=}\text{uniformly random} & \text{when } U_{WY}{\neq}X \end{cases}$ and taking U_{WY} to be a uniform random distribution with cardinality three. In this model for $\mathcal{G}'_{X\to W\to Y}$ we obtain $P(W{=}0) = \frac{1}{3}$, corresponding to $\mathsf{MME}_{X\to Y} \leq H(W) \approx 0.918$.

5 RELATED WORK

This work builds most directly on [Evans, 2012], in which e-separation was introduced and Theorem 4 was derived, both of which are essential to our results. It follows in the tradition of a line of literature that aims to derive symbolic expressions of restrictions on the observed data distribution implied by a causal model with latent variables, including [Tian and Pearl, 2002, Balke and Pearl, 1993, Kang and Tian, 2006]. Entropic constraints were previously considered in [Chaves et al., 2014, Chaves et al., 2014, Weilenmann and Colbeck, 2017]. The entropic constraint for the instrumental scenario appears as Equation (5) in [Chaves et al., 2014], see also Appendix E of [Henson et al., 2014]. Our work is also closely related to work in the literature on information theory on how much information can pass through channels of varying types [Gamal and Kim, 2011]. Our proposed measure of causal strength, the MME, is motivated by weaknesses in standard causal strength measures (e.g. ACE), which was previously discussed in [Janzing et al., 2013].

Our results are also related to the causal discovery literature, which seeks to find the causal structures compatible with an observed data distribution [Spirtes et al., 2000]. The inequality constraints posed above can be used to check the outputs of existing causal discovery algorithms [Strobl et al., 2018, Spirtes et al., 2000, Bernstein et al., 2020].

[4] Consider example 1 which has $\mathsf{ACE}_{X\to Y} = 0$. That example has the feature that $I(X : Y) = H(X)$. Accordingly the lower bound of $\mathsf{MME}_{X\to Y} \geq I(X : Y)$ is evidently tight, given the trivial upper bounder $\mathsf{MME}_{X\to Y} \leq H(X)$.

6 CONCLUSION

In this work, we present inequality constraints implied by e-separation relations in hidden variable DAGs. We have shown that these constraints can be used for a number of purposes, including adjudicating between causal models, bounding the cardinalities of latent variables, and measuring the strength of a causal relationship. e-separation relations can be read directly off a hidden variable DAG, leading to constraints that can be easily obtained.

This work opens up two avenues for future work. The first is that our constraints demonstrate a practical use of e-separation relations, and should motivate the study of fast algorithms for enumerating all such relations in hidden variable DAGs. The second is that the constraints suggest that equality-constraint-based causal discovery algorithms can be improved; understanding how the inequality constraints can best be used to this end will take careful study.

Acknowledgements

This research was supported by Perimeter Institute for Theoretical Physics. Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Colleges and Universities. The first author was supported in part by the Mathematical Institute for Data Science (MINDS) research fellowship. The third author was supported in part by grants ONR N00014-18-1-2760, NSF CAREER 1942239, NSF 1939675, and R01 AI127271-01A1. We thank Murat Kocaoglu for helpful discussion about the MME.

References

Alexander Balke and Judea Pearl. Nonparametric Bounds on Causal Effects from Partial Compliance Data. *J. Am. Stat. Ass.*, 1993. URL `https://escholarship.org/uc/item/2rn0420q`.

Daniel Bernstein, Basil Saeed, Chandler Squires, and Car-

oline Uhler. Ordering-Based Causal Structure Learning in the Presence of Latent Variables. In *Proc. 23rd Int. Conf. Artificial Intelligence and Statistics*, volume 108, pages 4098–4108. PMLR, 2020. URL https://arxiv.org/abs/1910.09014.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable Causal Discovery Under Unmeasured Confounding, 2020.

R. Chaves, L. Luft, T. O. Maciel, D. Gross, D. Janzing, and B. Schölkopf. Inferring latent structures via information inequalities. In *Proc. 30th Conf. UAI*, pages 112–121. AUAI, 2014. URL http://arxiv.org/abs/1407.2256.

Rafael Chaves, Lukas Luft, and David Gross. Causal structures from entropic information: geometry and novel scenarios. *New J. Phys.*, 16(4):043001, 2014. doi: 10.1088/1367-2630/16/4/043001.

Robin J Evans. Graphical methods for inequality constraints in marginalized DAGs. In *Proc. 2012 IEEE Intern. Work. MLSP*, pages 1–6. IEEE, 2012. doi: 10.1109/MLSP.2012.6349796. URL https://arxiv.org/abs/1209.2978.

Abbas El Gamal and Young-Han Kim. *Network Information Theory*. Cambridge University Press, 2011. doi: 10.1017/CBO9781139030687.

Dan Geiger and Chris Meek. Quantifier elimination for statistical problems. In *Proc. 15th Conf. UAI*, pages 226–235. Morgan Kaufmann, 1999.

Florian Gunsilius. A path-sampling method to partially identify causal effects in instrumental variable models, 2019. URL https://arxiv.org/abs/1910.09502. Working paper.

Joe Henson, Raymond Lal, and Matthew F. Pusey. Theory-independent limits on correlations from generalized Bayesian networks. *New J. Phys.*, 16(11):113043, 2014. doi: 10.1088/1367-2630/16/11/113043.

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, Bernhard Schölkopf, et al. Quantifying causal influences. *Ann. Stat.*, 41(5):2324–2358, 2013. doi: 10.1214/13-AOS1145.

Changsung Kang and Jin Tian. Inequality Constraints in Causal Models with Hidden Variables. In *Proc. 22nd Conf. UAI*, pages 233–240. AUAI, 2006. URL https://arxiv.org/abs/1206.6829.

Karthika Mohan, Judea Pearl, and Jin Tian. Graphical Models for Inference with Missing Data. In *Advances in Neural Information Processing Systems*, pages 1277–1285. Curran Associates, Inc., 2013.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988. ISBN 1558604790. doi: 10.1016/B978-0-08-051489-5.50009-6.

Judea Pearl. Causal inference in statistics: An overview, 2009.

Jacques Pienaar. Which causal structures might support a quantum–classical gap? *New J. Phys.*, 19(4):043021, 2017. doi: 10.1088/1367-2630/aa673e.

Davide Poderini, Rafael Chaves, Iris Agresti, Gonzalo Carvacho, and Fabio Sciarrino. Exclusivity graph approach to Instrumental inequalities. In *Proc. 35th Conf. UAI*. AUAI, 2019. URL https://arxiv.org/abs/1909.09120.

Thomas S. Richardson and James M. Robins. *Single World Intervention Graphs*. Now Publishers Inc, 2013. ISBN 9781601988102. URL https://www.csss.washington.edu/Papers/wp128.pdf.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov Properties for Acyclic Directed Mixed Graphs, 2017. Working paper.

Ilya Shpitser, Thomas S. Richardson, and James M. Robins. Chapter 41: Multivariate counterfactual systems and causal graphical models. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. ACM Books, 2021.

Peter L. Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000. doi: 10.1007/978-1-4612-2748-9.

Bastian Steudel and Nihat Ay. Information-theoretic inference of common ancestors. *Entropy*, 17(4):2304–2327, 2015. ISSN 1099-4300. doi: 10.3390/e17042304.

Eric V. Strobl, Shyam Visweswaran, and Peter L. Spirtes. Fast causal inference with non-random missingness by test-wise deletion. *Int. J. Data Science and Analytics*, 6 (1):47–62, 2018. doi: 10.1007/s41060-017-0094-6.

Jin Tian and Judea Pearl. On the Testable Implications of Causal Models with Hidden Variables. In *Proc. 18th Conf. UAI*. AUAI, 2002. URL https://arxiv.org/abs/1301.0608.

T. Verma and J. Pearl. Equivalence and synthesis of causal models, 1990.

Mirjam Weilenmann and Roger Colbeck. Analysing causal structures with entropy. *Proc. Roy. Soc. A*, 473(2207):20170483, 2017. doi: 10.1098/rspa.2017.0483.

Mirjam Weilenmann and Roger Colbeck. Analysing causal structures in generalised probabilistic theories. *Quantum*, 4:236, 2020. doi: 10.22331/q-2020-02-27-236.

Elie Wolfe, Robert W. Spekkens, and Tobias Fritz. The Inflation Technique for Causal Inference with Latent Variables. *J. Caus. Inf.*, 7(2), 2019. doi: 10.1515/jci-2017-0020.