

# Entropic Inequality Constraints from $e$ -separation Relations in Directed Acyclic Graphs with Hidden Variables Supplementary Material

Noam Finkelstein<sup>1</sup>

Beata Zjawin<sup>2,3</sup>

Elie Wolfe<sup>2</sup>

Ilya Shpitser<sup>1</sup>

Robert W. Spekkens<sup>2</sup>

<sup>1</sup> Johns Hopkins University, Department of Computer Science, 3400 N Charles St, Baltimore, MD USA, 21218

<sup>2</sup> Perimeter Institute for Theoretical Physics, 31 Caroline St. N, Waterloo, Ontario, Canada, N2L 2Y5

<sup>3</sup> International Centre for Theory of Quantum Technologies, University of, Gdańsk, 80-308 Gdańsk, Poland

## A COMPARING ENTROPIC INEQUALITIES TO GENERALIZED INDEPENDENCE RELATIONS

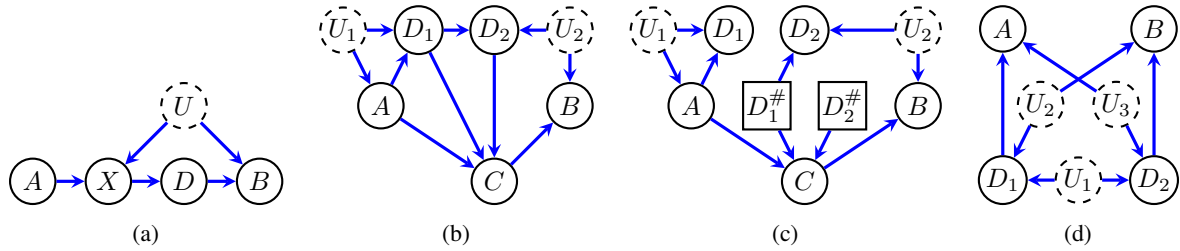


Figure S1: In graphs (a) and (b), the entropic inequality constraints are logically implied by equality constraints. Graphs (b) and (c) demonstrate that for a set of variables  $\mathbf{D}$ , the counterfactual random variable  $\mathbf{D}(\mathbf{D} = \mathbf{d})$  is not necessarily equal to the factual  $\mathbf{D}$ . Graph (d) provides an example where the entropic inequality constraints remain relevant even though the counterfactual distribution after intervention on an  $e$ -separating set over the remaining variables is identified.

In Proposition 7, we showed that for graphical models in which the counterfactual  $P(\mathbf{A}(\mathbf{D}=\mathbf{d}), \mathbf{B}(\mathbf{D}=\mathbf{d}), \mathbf{D}(\mathbf{D}=\mathbf{d})=\mathbf{d} \mid \mathbf{C})$  is identified, the entropic constraints of Theorem 5 are weaker than the corresponding Verma constraints. We now illustrate this point with a few examples. In Figs. S1(a) the counterfactual  $P(\mathbf{A}(\mathbf{D}=\mathbf{d}), \mathbf{B}(\mathbf{D}=\mathbf{d}), \mathbf{D}(\mathbf{D}=\mathbf{d})=\mathbf{d})$  is identified, and in S1(b) the counterfactual  $P(\mathbf{A}(\mathbf{D}=\mathbf{d}), \mathbf{B}(\mathbf{D}=\mathbf{d}), \mathbf{D}(\mathbf{D}=\mathbf{d})=\mathbf{d} \mid \mathbf{C})$  is identified. Accordingly, our entropic inequalities are implied by equality constraints, due to Proposition 7. The resulting inequality constraints therefore cannot provide any additional information about whether these causal structures are compatible with observed distributions.

By contrast, in Fig. S1(d) the counterfactual  $P(\mathbf{A}(\mathbf{D}=\mathbf{d}), \mathbf{B}(\mathbf{D}=\mathbf{d}), \mathbf{D}(\mathbf{D}=\mathbf{d})=\mathbf{d} \mid \mathbf{C})$  is not identified, even though  $P(\mathbf{A}(\mathbf{D}=\mathbf{d}), \mathbf{B}(\mathbf{D}=\mathbf{d}) \mid \mathbf{C})$  is. Although Fig. S1(d) implies no equality constraints [Evans and Richardson, 2019], we find that it *does* entail the entropic inequality constraint following from the  $e$ -separation relation  $(A \perp_e B \mid \text{upon } \neg\{D_1, D_2\})$ . It is therefore an example of a graph in which our inequality constraints are *not* made redundant by known equality constraints, despite the fact that intervention on  $\mathbf{D}$  is identified. This example is also an illustration of the fact that not every equality restriction featuring non-adjacent variables in an identifiable counterfactual distribution implies equality restrictions on the observed data distribution. However, some such non-adjacent variables may be involved in inequality restrictions.

The critical identifiability question for determining whether the entropic constraints are made redundant by equality constraints is  $P(\mathbf{A}(\mathbf{D}=\mathbf{d}), \mathbf{B}(\mathbf{D}=\mathbf{d}), \mathbf{D}(\mathbf{D}=\mathbf{d})=\mathbf{d} \mid \mathbf{C})$ . This distribution involves the counterfactual random variable  $\mathbf{D}(\mathbf{D}=\mathbf{d})$ . Note that although any *single* random variable under intervention on itself is equivalent to the random variable under no intervention, the same does not necessarily hold for *sets* of random variables. Figs. S1(b) and S1(c) demonstrate this point – because  $D_2$  is a descendant of  $D_1$ , after intervention on both,  $D_2$  no longer takes its natural value.

## B $e$ -SEPARATION IN IDENTIFIED COUNTERFACTUAL DISTRIBUTIONS

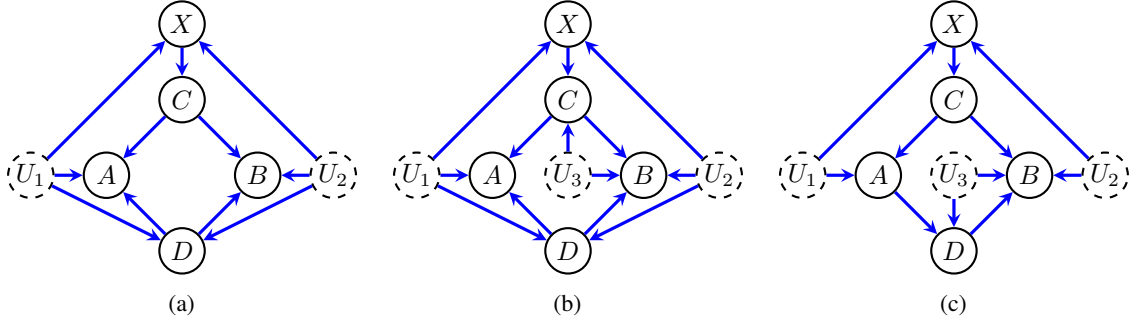


Figure S2: In all three graphs,  $A$  and  $B$  are  $e$ -separated by  $D$  after intervention on  $C$ . The counterfactual distribution over  $\{A, B, D\}$  after intervention on  $C$  is only identified in graphs (a) and (c), however.

A Single World Intervention Graph (SWIG) [Richardson and Robins, 2013], which represents the model after intervention on one or more random variables, can be obtained through a node-splitting operation as illustrated in Fig. S1(c). As described in Section 3.2,  $d$ -separation relations that appear under interventions with identified distributions can be used to derive equality constraints on the observed data distribution. In this section, we explore the significance of  $e$ -separation relations in identified counterfactual distributions.

We begin by noting that any  $e$ -separation relation that exists in a SWIG corresponds to an  $e$ -separation in the original DAG.

**Proposition 1.**  $(A \perp_e B \mid C \text{ upon } \neg D)$  after intervention on  $E$  only if  $(A \perp_e B \mid C \text{ upon } \neg\{D, E\})$ .

This proposition follows directly from the relationship between the fixing [Richardson et al., 2017] and deletion operations. In particular, fixing and deleting vertices induce the same graphical relationships among the remaining variables in the graph.

It may at first seem that this result indicates that  $e$ -separation relations in SWIGs cannot be used to derive inequality constraints on the observed data distribution that are not already implied by  $e$ -separation relations in the original model. However, entropic inequality constraints on counterfactual distributions have a different form than such constraints on the factual distribution. This is because entropies of counterfactual variables do not in general correspond to entropies of factual variables, so there is no way to express inequality constraints that follow from  $e$ -separation relations in SWIGs as entropic inequalities on the original distribution.

To illustrate this point, consider Fig. S2. In each graph,  $(A \perp_e B \mid \text{upon } \neg D)$  in the SWIG resulting from intervention on  $C$ . However, in Fig. S2(b), the distribution after intervention on  $C$  is not identified, whereas in Figs. S2(a) and S2(c) it is identified as  $P(A(c), B(c), D(c)) = \sum_x \frac{P(A, B, C=c, D, X=x)}{P(C=c|X=x)}$ . This means the entropic inequalities  $I(A(c) : B(c)) \leq H(D(c))$  on this counterfactual distribution (one for each level of  $C$ ) imply inequality constraints on the observed data distribution as well. These inequality constraints will be obtained in Figs. S2(a) and S2(c), but not in Fig. S2(b).

Moreover, these inequality constraints can be shown to be nontrivial. Since Figs. S2(a) and S2(b) share the same  $d$ -separation and  $e$ -separation relations it follows that any distributions compatible with Fig. S2(b) cannot be witnessed as incompatible with Fig. S2(a) using non-nested entropic equalities or inequalities. Consider the following structural equation model for Fig. S2(b): Let  $U_1, U_2$  and  $U_3$  be binary and uniformly distributed, and let  $X = U_2$ ,  $A = U_2 \oplus \epsilon_A$ ,  $C = X \oplus U_3$ ,  $B = C \oplus U_3 \oplus \epsilon_B$ , and  $D = \epsilon_D$  where “ $\oplus$ ” indicates addition mod 2 and where  $\epsilon_k$  is a random variable very heavily biased towards zero for  $k \in \{A, B, D\}$ . This establishes that  $C$  and  $X$  are uniformly distributed and statistically independent from each-other, and hence that  $P(A, B) = P(A(c=0), B(c=0))$ . This construction also gives  $A \oplus B = U_2 \oplus \epsilon_A \oplus C \oplus U_3 \oplus \epsilon_B = U_2 \oplus \epsilon_A \oplus X \oplus \epsilon_B = \epsilon_A \oplus \epsilon_B$  and hence  $A \approx B$ . This yields  $I(A(c=0) : B(c=0)) \approx H(A) \approx 1$  whereas  $H(D(c=0)) = H(D) \approx 0$ , strongly violating the entropic inequality  $I(A(c=0) : B(c=0)) \leq H(D(c=0))$  which applies only to Fig. S2(a).

## C PROOFS

### Proof of Theorem 5

Let  $\mathcal{G}^\#$  represent the graph in which every node in  $\mathbf{D}$  is split,  $P^*$  denote the distribution over variables in  $\mathcal{G}^\#$ . For notational convenience, we let  $P_{\mathbf{d}^\#}(\cdot | \cdot) = P^*(\cdot | \cdot, \mathbf{D}^\# = \mathbf{d}^\#)$ , and  $I_{\mathbf{d}^\#}$  and  $H_{\mathbf{d}^\#}$  be the mutual information and entropy in this distribution. Recall that by Theorem 4, if  $(\mathbf{A} \perp_e \mathbf{B} | \mathbf{C} \text{ upon } \neg \mathbf{D})$ , then:

- 4.i.  $I_{\mathbf{d}^\#}(\mathbf{A} : \mathbf{B} | \mathbf{C} = \mathbf{c}) = 0$ , and
- 4.ii.  $P(\mathbf{A}, \mathbf{B}, \mathbf{D} = \mathbf{d}^\# | \mathbf{C} = \mathbf{c}) = P_{\mathbf{d}^\#}(\mathbf{A}, \mathbf{B}, \mathbf{D} = \mathbf{d}^\# | \mathbf{C} = \mathbf{c})$ .

From the latter condition (4.ii.) we readily have that  $H(\cdot | \cdot, \mathbf{D} = \mathbf{d}^\#) = H_{\mathbf{d}^\#}(\cdot | \cdot, \mathbf{D} = \mathbf{d}^\#)$ .

It should also be clear that

$$H(\mathbf{X}) = H_{\mathbf{d}^\#}(\mathbf{X}) \quad \text{whenever } \mathbf{X} \text{ are among the nondescendants of } \mathbf{D}^\# \text{ in } \mathcal{G}^\#. \quad (\text{S1})$$

In our argument below,  $\mathbf{C}$  and  $\mathbf{D}$  are examples of such a set  $\mathbf{X}$ . If we view the distribution  $P_{\mathbf{d}^\#}$  from which  $H_{\mathbf{d}^\#}$  is derived as an interventional distribution, then the above identity follows from the exclusion restriction displayed graphically by rule 3 of the po-calculus [Malinsky et al., 2019].

It will prove extremely useful to show that  $H(\cdot | \cdot, \mathbf{D}) = H_{\mathbf{d}^\#}(\cdot | \cdot, \mathbf{D})$  can be seen to follow from  $H(\cdot | \cdot, \mathbf{D} = \mathbf{d}^\#) = H_{\mathbf{d}^\#}(\cdot | \cdot, \mathbf{D} = \mathbf{d}^\#)$  and  $P(\mathbf{D} = \mathbf{d}^\# | \cdot) = P_{\mathbf{d}^\#}(\mathbf{D} = \mathbf{d}^\# | \cdot)$  via

$$\begin{aligned} H(\cdot_{\text{pre}} | \cdot_{\text{post}}, \mathbf{D}) &= \sum_{\mathbf{d}} P(\mathbf{d} | \cdot_{\text{post}}) H(\cdot_{\text{pre}} | \cdot_{\text{post}}, \mathbf{d}) \\ &= \sum_{\mathbf{d}^\#} P(\mathbf{d}^\# | \cdot_{\text{post}}) H(\cdot_{\text{pre}} | \cdot_{\text{post}}, \mathbf{d}^\#) && \text{[summing over dummy index]} \\ &= \sum_{\mathbf{d}^\#} P_{\mathbf{d}^\#}(\mathbf{d}^\# | \cdot_{\text{post}}) H_{\mathbf{d}^\#}(\cdot_{\text{pre}} | \cdot_{\text{post}}, \mathbf{d}^\#) && \text{[applying Theorem 4]} \\ &= H_{\mathbf{d}^\#}(\cdot_{\text{pre}} | \cdot_{\text{post}}, \mathbf{D}) \end{aligned} \quad (\text{S2})$$

We will use conditions (S1) and (S2) to translate entropic constraints on  $P_{\mathbf{d}^\#}$  into entropic constraints on  $P$ . The two cases in Theorem 5 share the same implicit entropic constraints on  $P_{\mathbf{d}^\#}$ , but the implications on  $P$  are different: those scope of condition (S1)'s applicability increases under the promise that  $\mathbf{A}$  are nondescendants of  $\mathbf{D}$  in  $\mathcal{G}$ .

From this point on we will focus on deriving entropic inequality constraints on  $P_{\mathbf{d}^\#}$  such that *all the terms in the derived inequalities are translatable* according to conditions (S1) and (S2)<sup>1</sup>, because such constraints apply both to  $P_{\mathbf{d}^\#}$  and  $P$ . We hereafter denote  $\mathbf{C} = \mathbf{c}$  with simply  $\mathbf{c}$  for notational brevity. Firstly, consider the following entropic inequalities,

$$0 \leq I_{\mathbf{d}^\#}(\mathbf{A} : \mathbf{D} | \mathbf{c}) = H_{\mathbf{d}^\#}(\mathbf{A} | \mathbf{c}) - H_{\mathbf{d}^\#}(\mathbf{A} | \mathbf{c}, \mathbf{D}), \quad (\text{S3a})$$

$$0 \leq I_{\mathbf{d}^\#}(\mathbf{B} : \mathbf{D} | \mathbf{c}) = H_{\mathbf{d}^\#}(\mathbf{B} | \mathbf{c}) - H_{\mathbf{d}^\#}(\mathbf{B} | \mathbf{c}, \mathbf{D}), \quad (\text{S3b})$$

$$0 \leq H_{\mathbf{d}^\#}(\mathbf{D} | \mathbf{A}, \mathbf{B}, \mathbf{c}) = H_{\mathbf{d}^\#}(\mathbf{A}, \mathbf{B}, \mathbf{D} | \mathbf{c}) - H_{\mathbf{d}^\#}(\mathbf{A}, \mathbf{B} | \mathbf{c}), \quad (\text{S3c})$$

$$0 \leq -I_{\mathbf{d}^\#}(\mathbf{A} : \mathbf{B} | \mathbf{c}) = H_{\mathbf{d}^\#}(\mathbf{A}, \mathbf{B} | \mathbf{c}) - H_{\mathbf{d}^\#}(\mathbf{A} | \mathbf{c}) - H_{\mathbf{d}^\#}(\mathbf{B} | \mathbf{c}). \quad (\text{S3d})$$

The first two are subadditivity inequalities, which follow from the nonnegativity of conditional mutual information. The penultimate inequality follows from monotonicity (the fact that all conditional entropies are nonnegative). The final inequality is an expression of the fact that  $I_{\mathbf{d}^\#}(\mathbf{A} : \mathbf{B} | \mathbf{c}) = 0$  per (4.i.) above. Summing all four inequalities (S3) leads to the derived inequality

$$0 \leq H_{\mathbf{d}^\#}(\mathbf{A}, \mathbf{B}, \mathbf{D} | \mathbf{c}) - H_{\mathbf{d}^\#}(\mathbf{A} | \mathbf{c}, \mathbf{D}) - H_{\mathbf{d}^\#}(\mathbf{B} | \mathbf{c}, \mathbf{D}) \quad (\text{S4})$$

By applying conditions (S1) and (S2) to inequality (S4) we obtain

$$\begin{aligned} 0 &\leq H(\mathbf{A}, \mathbf{B}, \mathbf{D} | \mathbf{c}) - H(\mathbf{A} | \mathbf{c}, \mathbf{D}) - H(\mathbf{B} | \mathbf{c}, \mathbf{D}) \\ \text{i.e., } I(\mathbf{A} : \mathbf{B} | \mathbf{c}, \mathbf{D}) &\leq H(\mathbf{D} | \mathbf{c}). \end{aligned} \quad (\text{S5})$$

<sup>1</sup>Formally, the problem of inferring the implications of system of linear inequalities on a strict subset of their variables may be solved by means of Fourier-Motzkin elimination or related algorithms [Gläbke et al., 2018].

Now consider the case where we are further promised that  $\mathbf{A}$  are nondescendants of  $\mathbf{D}$  in  $\mathcal{G}$  and hence nondescendants of  $\mathbf{D}^\#$  in  $\mathcal{G}^\#$ . This means that in addition to the above results we also have that  $H_{\mathbf{d}^\#}(\mathbf{A} \mid \mathbf{c}) = H(\mathbf{A} \mid \mathbf{c})$ . We proceed as before, but instead of summing all four of the (S3) inequalities we only take the sum of the latter three. This yields

$$0 \leq H_{\mathbf{d}^\#}(\mathbf{A}, \mathbf{B}, \mathbf{D} \mid \mathbf{c}) - H_{\mathbf{d}^\#}(\mathbf{A} \mid \mathbf{c}) - H_{\mathbf{d}^\#}(\mathbf{B} \mid \mathbf{c}, \mathbf{D}) \quad (\text{S6})$$

By applying conditions (S1) and (S2) to inequality (S6) we obtain

$$\begin{aligned} 0 &\leq H(\mathbf{A}, \mathbf{B}, \mathbf{D} \mid \mathbf{c}) - H(\mathbf{A} \mid \mathbf{c}) - H(\mathbf{B} \mid \mathbf{c}, \mathbf{D}) \\ \text{i.e., } I(\mathbf{A} : \mathbf{B}, \mathbf{D} \mid \mathbf{c}) &\leq H(\mathbf{D} \mid \mathbf{c}). \end{aligned} \quad (\text{S7})$$

In both cases, the constraint is maintained after taking the expectation of both sides with respect to  $\mathbf{C}$ . Because each term in the expectation will satisfy the inequality, so will the sum.

Note that this proof technique can be adapted to derive stronger entropic inequalities for graphs which exhibit multiple different  $e$ -separation relations involving the same  $\mathbf{D}$  set. If  $(\mathbf{A}_1 \perp_e \mathbf{B}_1 \mid \mathbf{C}_1 \text{ upon } \neg \mathbf{D})$  and  $(\mathbf{A}_2 \perp_e \mathbf{B}_2 \mid \mathbf{C}_2 \text{ upon } \neg \mathbf{D})$  and so forth, then Theorem 4 still demands the existence of a *single*  $P_{\mathbf{d}^\#}$  whose various margins must now satisfy *multiple distinct* zero conditional mutual informational equalities. We can accommodate multiple entropic equality constraints on  $P_{\mathbf{d}^\#}$  just as easily as we can accommodate a single equality constraint: The translation between constraints on  $P_{\mathbf{d}^\#}$  and  $P_{\mathbf{d}^\#}$  will continue to be governed by conditions (S1) and (S2).

### Proof of Proposition 6

If conditioning on some variables  $\mathbf{D}$  is sufficient to close a path, then that path must go through  $\mathbf{D}$ , and therefore deletion of  $\mathbf{D}$  eliminates the path. By construction, the deletion operation can never open a path, unlike the conditioning operation. If  $(\mathbf{A} \perp_d \mathbf{B} \mid \mathbf{C}, \mathbf{D})$ , then all paths from  $\mathbf{A}$  to  $\mathbf{B}$  go through  $\mathbf{C}$  or  $\mathbf{D}$ , or through colliders that are not in  $\{\mathbf{C}, \mathbf{D}\}$ , nor have any descendants therein. It follows that  $(\mathbf{A} \perp_e \mathbf{B} \mid \mathbf{C} \text{ upon } \neg \mathbf{D})$ , as after deletion of  $\mathbf{D}$  all paths through  $\mathbf{C}$  remain blocked through conditioning, all paths through  $\mathbf{D}$  are eliminated, and all other paths remain blocked by colliders.

### Proof of Proposition 8

By the standard proof used for the data processing inequality (see [Cover and Thomas, 2012]),

$$I(\mathbf{A} : \mathbf{U} \mid \mathbf{C}) \geq I(\mathbf{A} : \mathbf{B} \mid \mathbf{C}) \quad \text{whenever } \mathbf{A} \perp \mathbf{B} \mid \{\mathbf{C}, \mathbf{U}\}. \quad (\text{S8})$$

The data processing inequality can be written

$$H(\mathbf{U} \mid \mathbf{C}) \geq H(\mathbf{A}, \mathbf{U} \mid \mathbf{C}) + H(\mathbf{B} \mid \mathbf{C}) - H(\mathbf{A}, \mathbf{B} \mid \mathbf{C}). \quad (\text{S9})$$

The result  $H(\mathbf{U} \mid \mathbf{C}) \geq I(\mathbf{A} : \mathbf{B} \mid \mathbf{C})$  then follows from the fact that the joint entropy of two random variables is greater than the entropy of either, i.e.,  $H(\mathbf{A}, \mathbf{U} \mid \mathbf{C}) \geq H(\mathbf{A} \mid \mathbf{C})$  and the definition  $I(\mathbf{A} : \mathbf{B} \mid \mathbf{C}) = H(\mathbf{A} \mid \mathbf{C}) + H(\mathbf{B} \mid \mathbf{C}) - H(\mathbf{A}, \mathbf{B} \mid \mathbf{C})$ . The result  $H(\mathbf{U}) \geq H(\mathbf{U} \mid \mathbf{C})$  follows from monotonicity, i.e., the fact that conditional entropy is never greater than marginal entropy.

### Proof of Corollary 10.1

*Proof.* A consequence of  $X \in pa(W)$ , is that  $\mathbf{A} \subset X \cup an(X)$  implies that no element of  $\mathbf{A}$  is a descendant of  $W$ . This allows to confirm the following sequence of inequalities,

$$\max_c I(\mathbf{A} : \mathbf{B} \mid \mathbf{C}=\mathbf{c}, \mathbf{D}) \quad (\text{S10a})$$

$$\leq \max_{\text{SEMs for } \mathcal{G}'} \max_c I(\mathbf{A} : \mathbf{B}, W \mid \mathbf{C}=\mathbf{c}, \mathbf{D}) \quad (\text{S10b})$$

$$\leq \max_{\text{SEMs for } \mathcal{G}'} \max_c H(\mathbf{D}, W \mid \mathbf{C}=\mathbf{c}) \quad (\text{S10c})$$

$$\leq \max_c \left( H(\mathbf{D} \mid \mathbf{C}=\mathbf{c}) + \max_{\text{SEMs for } \mathcal{G}'} H(W \mid \mathbf{C}=\mathbf{c}) \right) \quad (\text{S10d})$$

$$\leq \max_c H(\mathbf{D} \mid \mathbf{C}=\mathbf{c}) + \max_{\text{SEMs for } \mathcal{G}'} H(W) \quad (\text{S10e})$$

$$= \max_c H(\mathbf{D} \mid \mathbf{C}=\mathbf{c}) + \text{MME}_{X \rightarrow Y}$$

where all the steps above are consequences of subadditivity except for the step from Equation (S10b) to Equation (S10c), which is just the application of Equation (4a). Finally, Equation (8b) follows from Equation (8a) by convexity.  $\square$

## D RELATION BETWEEN COMMON ENTROPY AND MME

The MME bears some resemblance to a concept called *common entropy* [Kumar et al., 2014], which is defined for a distribution  $P(X, Y)$  as the smallest possible entropy of an unobserved variable  $W$  such that  $X \perp Y \mid W$ . Unlike the MME, the common entropy is a function only of the probability distribution  $P(X, Y)$ , and not of the graph  $\mathcal{G}$ . Any  $W$  that renders  $X$  and  $Y$  conditionally independent must also fully mediate the effect of  $X$  on  $Y$ , which at first glance might be taken to mean that the common entropy is an upper bound on the MME, because it implies that the MME can search over a larger set of distributions to obtain a low-entropy mediator. Indeed in the simple  $X \rightarrow Y$  model, it is the case that  $\text{MME}_{X \rightarrow Y}$  is bounded from above by the common entropy between  $X$  and  $Y$  for precisely this reason.

However, the common entropy is not an upper bound on the MME in general. To see this, consider the graph presented in Fig. 1(c). This model contains distributions in which  $A$  and  $B$  are highly correlated, but  $D$  and  $B$  are entirely uncorrelated. For such distributions, the common entropy of  $B$  and  $D$  would be 0, as they are already marginally independent. However, by Corollary 10.1, the MME would be bounded from below by  $I(A : B)$ , which can be larger than 0. The intuition for this phenomenon is that if the edge  $D \rightarrow B$  were missing,  $A$  and  $B$  would be marginally independent, so a high mutual information between them is evidence for the causal significance of the edge.

## References

- Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory, Second Edition*. Wiley Science, 2012. doi: 10.1002/047174882X.
- Robin J. Evans and Thomas S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 25(2):848 – 876, 2019. doi: 10.3150/17-BEJ1005.
- T Gläble, D Gross, and R Chaves. Computational tools for solving a marginal problem with applications in Bell non-locality and causal modeling. *J. Phys A*, 51(48):484002, November 2018. doi: 10.1088/1751-8121/aae754.
- Gowtham Ramani Kumar, Cheuk Ting Li, and Abbas El Gamal. Exact common information. In *2014 IEEE International Symposium on Information Theory*, pages 161–165, 2014. doi: 10.1109/ISIT.2014.6874815.
- Daniel Malinsky, Ilya Shpitser, and Thomas S. Richardson. A potential outcomes calculus for identifying conditional path-specific effects. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.
- 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.