Journal of Artificial Intelligence Research 82 (2025) 1-37 Submitted 05/2024; published 01/2025

Approximate Implication for Probabilistic Graphical Models

Batya Kenig BATYAK@TECHNION.AC.IL
Technion, Israel Institute of Technology
Haifa, Israel

Abstract

The graphical structure of Probabilistic Graphical Models (PGMs) represents the conditional independence (CI) relations that hold in the modeled distribution. Every *separator* in the graph represents a conditional independence relation in the distribution, making them the vehicle through which new conditional independence relations are inferred and verified. The notion of separation in graphs depends on whether the graph is directed (i.e., a *Bayesian Network*), or undirected (i.e., a *Markov Network*).

The premise of all current systems-of-inference for deriving CIs in PGMs, is that the set of CIs used for the construction of the PGM hold *exactly*. In practice, algorithms for extracting the structure of PGMs from data discover *approximate CIs* that do not hold exactly in the distribution. In this paper, we ask how the error in this set propagates to the inferred CIs read off the graphical structure. More precisely, what guarantee can we provide on the inferred CI when the set of CIs that entailed it hold only approximately? It has recently been shown that in the general case, no such guarantee can be provided.

In this work, we prove new negative and positive results concerning this problem. We prove that separators in undirected PGMs do not necessarily represent approximate CIs. In other words, no guarantee can be provided for CIs inferred from the structure of undirected graphs. We prove that such a guarantee exists for the set of CIs inferred in directed graphical models, making the *d-separation* algorithm a sound and complete system for inferring *approximate CIs*. We also establish improved approximation guarantees for independence relations derived from *marginal* and *saturated* CIs.

1. Introduction

Conditional independencies (CI) are assertions of the form $X \perp Y | Z$, stating that the random variables (RVs) X and Y are independent when conditioned on Z. The concept of conditional independence is at the core of Probabilistic Graphical Models (PGMs) that include Bayesian and Markov networks. The CI relations between the random variables enable the modular and low-dimensional representations of high-dimensional, multivariate distributions, and tame the complexity of inference and learning, which would otherwise be very inefficient (Koller & Friedman, 2009; Pearl, 1989).

The *implication problem* is the task of determining whether a set of CIs termed *antecedents* logically entail another CI, called the *consequent*, and it has received considerable attention from both the AI and Database communities (Pearl & Paz, 1986; Geiger, Verma, & Pearl, 1989; Geiger, Paz, & Pearl, 1991a; Sayrafi, Van Gucht, & Gyssens, 2008; Kenig & Suciu, 2020; Kenig, Mundra, Prasaad, Salimi, & Suciu, 2020). Known algorithms for deriving CIs from the topological structure of the graphical model are, in fact, an instance of implication. The Directed Acyclic Graph (DAG) structure of Bayesian Networks is generated based on a set of CIs termed the *recursive basis* (Geiger, Verma, & Pearl, 1990), and the d-separation algorithm is used to derive additional CIs, implied by this set. In undirected PGMs, also called Markov networks or Markov Random Fields (MRFs), every pair of non-adjacent vertices u and v signify that u and v are conditionally independent given the rest of the vertices in the graph. If the underlying distribution is strictly positive, then this set of CIs (i.e., corresponding to the pairs of non-adjacent vertices) imply a much larger set of CIs associated with the *separators* of the graph (Studený, 2018). A separator in an undirected graph $G(V, E)$ is a subset of vertices $C \subseteq V$ whose removal breaks the graph into two or more connected components. Specifically, if C is a separator in the undirected PGM $G(V, E)$, then the vertex set $V \backslash C$ can be partitioned into two disjoint sets $A, B \subseteq V$, where every path between a vertex $a \in A$ and $b \in B$ passes through a vertex in C. This partitioning corresponds to the conditional independence relation $A \perp B | C$.

The d-separation algorithm is a sound and complete method for deriving CIs in probability distributions represented by DAGs (Geiger et al., 1989, 1990). In undirected PGMs, graph-separation completely characterizes the conditional independence statements that can be derived from the conditional independence statements associated with the non-adjacent vertex pairs of the graph (Pearl, Geiger, & Verma, 1989; Geiger & Pearl, 1993; Studený, 2018). The foundation of deriving CIs in directed and undirected models is the *semigraphoid axioms* and the *graphoid axioms*, respectively (Dawid, 1979; Geiger, Paz, & Pearl, 1991b; Geiger & Pearl, 1993).

Current systems for inferring CIs, and the graphoid axioms in particular, assume that both antecedents and consequent hold *exactly*, hence we refer to these as an exact implication (EI). However, almost all known approaches for learning the structure of a PGM rely on CIs extracted from data, which hold to a large degree, but cannot be expected to hold exactly. Of these, structure-learning approaches based on information theory have been shown to be particularly successful, and thus widely used to infer networks in many fields (Cheng, Greiner, Kelly, Bell, & Liu, 2002; de Campos, 2006; Chen, Anantha, & Lin, 2008; Zhao, Zhou, Zhang, & Chen, 2016; Kenig et al., 2020).

In this paper, we drop the assumption that the CIs hold exactly, and consider the *relaxation problem*: if an exact implication holds, does an *approximate implication* hold too? That is, if the antecedents approximately hold in the distribution, does the consequent approximately hold as well? What guarantees can we give for the approximation? In other words, the relaxation problem asks whether, and under what conditions, we can convert an exact implication to an approximate one. When relaxation holds, then the error to the consequent can be bounded, and any system-of-inference for deriving exact implications (e.g., the semigraphoid axioms, d-separation, graph-separation), can be used to infer an approximate implication.

To study the relaxation problem we need to measure the degree of satisfaction of a CI. In line with previous work, we use Information Theory. This is the natural semantics for modeling CIs because $X \perp Y | Z$ if and only if $I(X; Y|Z) = 0$, where I is the conditional mutual information. A CI is called a *conditional* if it has the form $X \rightarrow Y$ and Y is a function of X. In this case, $X \rightarrow Y$ if and only if $h(Y|X) = 0$, where h is the conditional entropy. An exact implication (EI) $\sigma_1, \cdots, \sigma_k \Rightarrow \tau$ is an assertion of the form $(h(\sigma_1){=}0 \wedge \cdots \wedge h(\sigma_k){=}0) \Rightarrow h(\tau){=}0$, where $\tau, \sigma_1, \sigma_2, \ldots$ are either triples $(X; Y|Z)$ or pairs $(Y|X)$ representing CIs and conditionals respectively. If $\sigma = (X; Y|Z)$, then $h(\sigma) \overset{\text{def}}{=} I(X; Y|Z)$ is the mutual information measure; if $\sigma = (Y|X)$, then $h(\sigma) \overset{\text{def}}{=} h(Y|X)$ is the conditional entropy measure. An approximate implication (AI) is a linear inequality $h(\tau) \leq \lambda h(\Sigma)$, where $h(\Sigma) \stackrel{\text{def}}{=} \sum_{i=1}^{k} h(\sigma_i)$, and $\lambda \geq 0$ is the approximation factor. We say that a class of CIs λ-*relaxes* if every exact implication (EI) from the class can be transformed to an approximate implication (AI) with an approximation factor λ. We observe that an approximate implication always implies an exact implication because the mutual information $I(\cdot;\cdot|\cdot) \geq 0$ and conditional entropy $h(\cdot|\cdot)$ are nonnegative measures. Therefore, if $0 \leq h(\tau) \leq \lambda h(\Sigma)$ for some $\lambda \geq 0$, then $h(\Sigma) = 0 \Rightarrow h(\tau) = 0$.

Results. A conditional independence assertion $(A;B|C)$ is called *saturated* if it mentions all of the random variables in the joint distribution, and it is called *marginal* if $C = \emptyset$. The exact variant of implication was extensively studied (Geiger et al., 1989; Geiger & Pearl, 1993, 1988; Geiger et al., 1991a, 1990) (see below the related work). In this paper, we study approximate implication. Our results are summarized in Table 1.

We first consider exact implications $\Sigma \Rightarrow \tau$, where the set of antecedents Σ is comprised of saturated CIs and conditionals, and no assumption is made on the consequent CI τ. The syntactic fragment of exact implication from saturated and conditional antecedents was also studied in the database community (where saturated CIs are called *MVDs - Multivalued Dependencies*, and conditionals are called *FDs - Functional Dependencies*) (Beeri, Fagin, & Howard, 1977; Beeri, 1980). In an undirected PGM $G(V,E)$, every separator Z corresponds to a saturated CI statement $(X;Y|Z)$, where every path from a vertex $x \in X$ to a vertex $y \in Y$ passes through a vertex in Z. Hence, Σ can be viewed as a set of CIs that hold in a probability distribution represented by a Markov Network. We show that if τ can be derived from Σ by applying the *semigraphoid axioms*, then the implication relaxes. Specifically, if $\tau = (A;B|C)$, then $h(\tau) \leq \min\{|A|,|B|\}h(\Sigma)$ (i.e., where $|A|$ denotes the number of RVs in the set A). For n jointly-distributed random variables, this leads to a relaxation bound of $\min\{|A|,|B|\} \leq {}^{n}\!/_{2}$. In previous work, it was shown that $h(\tau) \leq |A| {\cdot} |B| {\cdot} h(\Sigma)$, leading to a relaxation bound of ${}^{n^2}\!/_{4}$ (Kenig & Suciu, 2022). This work (Theorem 4.3) tightens the relaxation bound by an order of magnitude.

We also prove a negative result. If the implication involves the application of the *intersection axiom* (i.e., that is one of the *graphoid axioms* (Pearl, 1989)), then no relaxation exists. We present a strictly positive probability distribution in which the intersection axiom does not relax. Consequently, no relaxation exists for implications derived using the intersection axiom, or more broadly, the graphoid axioms. Inferring CIs in Markov Networks relies on the intersection axiom (Pearl, 1989; Studený, 2018). This negative result essentially establishes that if the CI relations associated with the non-adjacent vertex pairs in the graph do not hold exactly, then no guarantee can be made regarding the CI relations associated with the separators of the graph.

We show that every conditional independence relation $(A;B|C)$ read off a Directed Acyclic Graph (DAG) by the d-separation algorithm (Geiger et al., 1989), admits a 1-approximation (Theorem 4.4). In other words, if Σ is the *recursive basis* of CIs used to build the Bayesian network (Geiger et al., 1989), then it is guaranteed that $I(A;B|C) \leq \sum_{i \in \Sigma} h(\sigma_i)$. Furthermore, we present a family of implications for which our 1-approximation is tight (i.e., $I(A;B|C) = \sum_{i \in \Sigma} h(\sigma_i)$). This result first appeared in Kenig (2021). In this paper, we simplify the proof, and relate it to the d-separation algorithm.

We prove that every CI $(A;B|C)$ implied by a set of marginal CIs admits a $\min\{|A|,|B|\}$-approximation. For n jointly-distributed random variables, this leads to a relaxation bound of $\min\{|A|,|B|\} \leq n/2$. In previous work, it was shown that $h(\tau) \leq |A|{\cdot}|B|{\cdot}h(\Sigma)$, leading to a relaxation bound of $n^2/4$ (Kenig, 2021). The relaxation bound established in this work (Theorem 4.5) is smaller by an order of magnitude.

Of independent interest is the technique used for proving the approximation guarantees. The *I-Measure* (Yeung, 1991) is a theory which establishes a one-to-one correspondence between information theoretic measures such as entropy and mutual information (defined in Section 2) and set theory.

Related Work. The AI community has extensively studied the exact implication problem for Conditional Independencies (CI). In a series of papers, Geiger et al. showed that the *semigraphoid axioms* (Pearl & Paz, 1986) are sound and complete for deriving CI statements that are implied by marginal CIs (Geiger & Pearl, 1993), and *recursive CIs* that are used in Bayesian networks (Geiger et al., 1990; Geiger & Pearl, 1988). In the same paper, they also showed that, when restricted to the set of strictly positive probability distributions, the *graphoid axioms* are sound and complete for deriving CI statements from saturated CIs (Geiger & Pearl, 1993). The completeness of d-separation follows from the fact that the set of CIs derived by d-separation is precisely the closure of the recursive basis under the semigraphoid axioms (Verma & Pearl, 1990). Studený proved that in the general case, when no assumptions are made on the antecendents, no finite axiomatization exists (Studený, 1990). That is, there does not exist a finite set of axioms (deductive rules) from which all general conditional independence implications can be deduced. Recently, Li (2023) has shown that the CI implication problem is, in general, undecidable.

The database community has studied the implication problem for integrity constraints (Armstrong & Delobel, 1980; Beeri et al., 1977; Kontinen, Link, & Väänänen, 2013; Maier, 1983), and showed that the implication problem is decidable and axiomatizable when the antecedents are Functional Dependencies (FDs) or *Multivalued Dependencies* (which correspond to saturated CIs (Lee, 1987; Kenig & Suciu, 2020)), and undecidable for *Embedded Multivalued Dependencies* (Herrmann, 1995).

The relaxation problem was first studied by Kenig and Suciu in the context of database dependencies (2020), where they showed that CIs derived from a set of saturated antecedents, admit an approximate implication. Importantly, they also showed that not all exact implications relax, and presented a family of 4-variable distributions along with an exact implication that does not admit an approximation (see Theorem 16 in Kenig and Suciu (2020)). Consequently, it is not straightforward that exact implication necessarily implies its approximation counterpart, and arriving at meaningful approximation guarantees requires making certain assumptions on the antecedents, consequent, derivation rules, or combination thereof.

Organization. We start in Section 2 with preliminaries. In Section 3 we describe the role exact implication has played in probabilistic graphical models, and introduce the notion of approximate implication. We formally state the results, and their practical implications in Section 4. We prove that the intersection axiom does not relax in Section 5. In Section 6 we establish, through a series of lemmas, properties of exact implication that will be used for proving our results. In Sections 7, 8, and 9, we prove the relaxation bounds for exact implications from the set of saturated CIs, the recursive basis, and marginal CIs respectively, (see Table 1). We conclude in Section 10.

Type of EI	Relaxation Bounds			
	General	Saturated+FDs $\Rightarrow$ any	Recursive Basis $\Rightarrow$ any	Marginals $\Rightarrow$ any
Semigraphoid	$(2^n)!$ (Kenig & Suciu, 2022)	$n/2$ (Thm. 4.3)	1 (Thm. 4.4)	$n/2$ (Thm. 4.5)
Graphoid	∞ (Thm. 4.1)			

Table 1: **Summary of results.** The relaxation bounds for the implication $\Sigma \Rightarrow \tau$ under various restrictions. (1) *General*; derivation rules are the semigraphoid axioms, and no restrictions are placed on Σ. (2) Σ is a set of saturated CIs and conditionals, and τ is any CI or conditional. (3) Σ is the *recursive basis* used to generate the Bayesian network, and τ is any CI. (4) Σ is a set of marginal CIs, and τ is any CI or conditional. (5) When the set of derivation rules include the *intersection axiom* (e.g., the *graphoid axioms*), then no finite relaxation bound exists.

2. Preliminaries

We denote by $[n]$ the set $\{1, 2, \ldots, n\}$. If $\Omega = \{X_1, \ldots, X_n\}$ denotes a set of variables and $U, V \subseteq \Omega$, then we abbreviate the union $U \cup V$ with UV.

2.1 Conditional Independence

Recall that two discrete random variables X, Y are called *independent* if $p(X = x, Y = y) = p(X = x) \cdot p(Y = y)$ for all outcomes x, y. We denote by $\mathcal{D}(X)$ the domain of the random variable X. Fix $\Omega = \{X_1, \ldots, X_n\}$, a set of n jointly distributed discrete random variables with finite domains $\mathcal{D}_1, \ldots, \mathcal{D}_n$, respectively; let p be the probability mass. For $\alpha \subseteq [n]$, denote by X_α the joint random variable $(X_i : i \in \alpha)$ with domain $\mathcal{D}_\alpha \stackrel{\text{def}}{=} \prod_{i \in \alpha} D_i$. We write $p \models X_\beta \perp X_\gamma | X_\alpha$ when X_β, X_γ are conditionally independent given X_α; in the special case that X_α functionally determines X_β, we write $p \models X_\alpha \rightarrow X_\beta$. We say that a set of random variables $\{X_1, \ldots, X_k\}$ are *mutually independent given Z* if $p(X_1 = x_1, \ldots, X_k = x_k | Z) = p(X_1 = x_1 | Z) \cdots p(X_k = x_k | Z)$. If $Z = \emptyset$, then we say that $\{X_1, \ldots, X_k\}$ are *mutually independent.*

An assertion $X \perp Y | Z$ is called a *Conditional Independence* statement, or a CI; this includes $Z \rightarrow Y$ as a special case (see Section 2.2). When $XYZ = \Omega$ we call it *saturated*, and when $Z = \emptyset$ we call it *marginal*. A set of CIs Σ *implies* a CI τ, in notation $\Sigma \Rightarrow \tau$, if every probability distribution that satisfies Σ also satisfies τ.

2.2 Background on Information Theory

We adopt required notation from the literature on information theory (Yeung, 2008). For $n > 0$, we identify the functions $2^{[n]} \rightarrow \mathbb{R}$ with the vectors in $\mathbb{R}^{2^n}$. All logarithms are taken in base 2.

Polymatroids. A function $h \in \mathbb{R}^{2^n}$ is called a *polymatroid* if $h(\emptyset) = 0$ and satisfies the following inequalities, called *Shannon inequalities*:

1. Monotonicity: $h(A) \leq h(B)$ for $A \subseteq B$.

2. Submodularity: $h(A \cup B) + h(A \cap B) \leq h(A) + h(B)$ for all $A, B \subseteq [n]$.

The set of polymatroids is denoted $\Gamma_n \subseteq \mathbb{R}^{2^n}$. For any polymatroid h and subsets $A, B, C, D \subseteq [n]$, we define[1]

$$h(B|A) \stackrel{\mathsf{def}}{=} h(AB) - h(A), \text{ and} \tag{1}$$

$$I_h(B;C|A) \stackrel{\mathsf{def}}{=} h(AB) + h(AC) - h(ABC) - h(A). \tag{2}$$

We denote $I_h(B;C|\emptyset)$ by $I_h(B;C)$. Then, $\forall h \in \Gamma_n$, $I_h(B;C|A) \geq 0$ by submodularity, and $h(B|A) \geq 0$ by monotonicity. When h is clear from the context we sometimes write $I(\cdot;\cdot|\cdot)$ instead of $I_h(\cdot;\cdot|\cdot)$. We say that A *functionally determines* B, in notation $A \rightarrow B$ if $h(B|A) = 0$. The *chain rule* is the identity:

$$I_h(B;CD|A) = I_h(B;C|A) + I_h(B;D|AC). \tag{3}$$

We call the triple $(B;C|A)$ *elemental* if $|B| = |C| = 1$; $h(B|A)$ is a special case of I_h, because $h(B|A) = I_h(B;B|A)$. By the chain rule, it follows that every CI $(B;C|A)$ can be written as a sum of at most $|B| \cdot |C| \leq {n^2}/{4}$ elemental CIs.

Entropy. If X is a random variable with a finite domain $\mathcal{D}$ and probability mass p, then $H(X)$ denotes its entropy

$$H(X) \stackrel{\mathsf{def}}{=} \sum_{x \in \mathcal{D}} p(x) \log \frac{1}{p(x)}. \tag{4}$$

For a set of jointly distributed random variables $\Omega = \{X_1, \ldots, X_n\}$ we define the function $h : 2^{[n]} \rightarrow \mathbb{R}$ as $h(\alpha) \stackrel{\mathsf{def}}{=} H(X_\alpha)$; h is called an *entropic function*, or, with some abuse, an *entropy*. It is easily verified that the entropy H satisfies the Shannon inequalities, and is thus a polymatroid. The quantities $h(B|A)$ and $I_h(B;C|A)$ are called the *conditional entropy* and *conditional mutual information* respectively. The conditional independence $p \models B \perp C \mid A$ holds if and only if $I_h(B;C|A) = 0$, and similarly $p \models A \rightarrow B$ if and only if $h(B|A) = 0$, where h is the entropic vector of p. Thus, entropy provides us with an alternative characterization of CIs.

The following identity holds for conditional entropy (Yeung, 2008):

$$h(X_1 \cdots X_k|Z) = h(X_1|Z) + h(X_2|X_1 Z) + \cdots + h(X_k|X_1 \cdots X_{k-1} Z). \tag{5}$$

If the RVs $X_1, \ldots, X_k$ are mutually independent given Z, then

$$h(X_1 \cdots X_k|Z) = h(X_1|Z) + h(X_2|Z) + \cdots + h(X_k|Z). \tag{6}$$

Axioms for Conditional Independence. A dependency model M is a subset of triplets $(X;Y|Z)$ for which the CI $X \perp Y|Z$ holds. For example, we say that M is a dependency model of a joint-distribution p if the CI $X \perp Y|Z$ holds in p for every triple $(X;Y|Z) \in M$. A *semigraphoid* is a dependency model that is closed under the following four axioms:

1. Symmetry: $X \perp Y|Z \Leftrightarrow Y \perp X|Z$.
2. Decomposition: $X \perp YW|Z \Rightarrow X \perp Y|Z$.

[1] Recall that AB denotes $A \cup B$.

3. Weak Union: $X \perp YW|Z \Rightarrow X \perp Y|WZ$.
4. Contraction: $X \perp Y|Z$ and $X \perp W|YZ \Rightarrow X \perp YW|Z$.

If, in addition, the dependency model is closed under the *intersection axiom*:

$$X \perp Y|ZW \text{ and } X \perp W|ZY \Rightarrow X \perp YW|Z, \tag{7}$$

then the dependency model is called a *graphoid*. The *closure* of a set of CIs Σ with respect to the semigraphoid (graphoid) axioms, is a set of CIs $\Sigma' \supseteq \Sigma$ that can be derived from Σ by repeated application of the semigraphoid (graphoid) axioms.

Let p be a joint probability distribution over the random variables $\Omega = \{X_1, \dots, X_n\}$, and let h be its entropy function. Since $p \models X \perp Y|Z$ iff $I_h(X;Y|Z) = 0$, and since $I_h(X;Y|Z) \geq 0$, it follows that the semigraphoid axioms are corollaries of the chain rule (see (3)). Since the chain rule, and the non-negativity of $I_f(X;Y|Z)$ hold for all polymatroids $f \in \Gamma_n$, it follows that the semigraphoid axioms hold for all polymatroids $f \in \Gamma_n$. Therefore, all polymatroids are semigraphoids. By applying the chain-rule, the semigraphoid axioms can be generalized to the following information inequalities:

1. Symmetry: $I_h(X;Y|Z) = I_h(Y;X|Z)$.
2. Decomposition: $I_h(X;YW|Z) \geq I_h(X;Y|Z)$.
3. Weak Union: $I_h(X;YW|Z) \geq I_h(X;Y|WZ)$.
4. Contraction: $I_h(X;Y|Z) + I_h(X;W|YZ) = I_h(X;YW|Z)$.

This is not the case for the intersection axiom (see (7)), which holds only for a strict subset of the probability distributions.A sufficient condition for the intersection axiom to hold is that the distribution is strictly positive. In Section 5, we show that contrary to the semigraphoid axioms, the intersection axiom does not correspond to any information inequality.

2.2.1 Representing Information Measures

In what follows we consider two information measures, CIs represented by triples $(A;B|C)$ whose information measure is $I(A;B|C)$, and conditionals $A \rightarrow B$ whose information measure is $h(B|A)$. In this section, we show that we can make certain assumptions about the sets A, B, and C. First, we may assume that for every triple $(A;B|C)$ it holds that $A, B \supset \emptyset$. If not, then from (2), it immediately follows that $I(A;B|C) = 0$, or that the CI $A \perp B|C$ trivially holds. Likewise, for every conditional $A \rightarrow B$, we may assume that $B \supset \emptyset$. Otherwise, by (1), we get that $h(B|A) = 0$, which means that the functional dependency $A \rightarrow B$ trivially holds.

Lemma 2.1. *Let A, B, C, and X be pairwise disjoint sets of jointly distributed random variables, then*

$$I(AX;BX|CX) = I(AX;B|CX) = I(A;B|CX).$$

Proof. By definition (see (2)), it holds that

$$\begin{aligned} I(AX;BX|CX) &= h(ACXX) + h(BCXX) - h(ABCXXX) - h(CX) \\ &= h(ACX) + h(BCX) - h(ABCX) - h(CX) \\ &= I(A;B|CX). \end{aligned}$$

□

Lemma 2.1 implies that we may assume w.l.o.g. that in every triple $(A;B|C)$ considered, it holds that $A\cap C=\emptyset$ and $B\cap C=\emptyset$.

The following identity will be useful. It follows immediately from (1) and (2).

$$h(A\rightarrow B)=h(B|A)=I(B;B|A). \tag{8}$$

Lemma 2.2. *Let Ω be a set of jointly distributed RVs, and let $A,B,X\subseteq\Omega$ be pairwise disjoint, jointly distributed RVs. Then*

$$h(AX\rightarrow BX)=h(AX\rightarrow B)=I(B;\Omega\backslash(ABX)|AX)+I(B;B|\Omega\backslash B).$$

Proof. We first show that $h(AX\rightarrow BX)=h(AX\rightarrow B)$. By definition (see (1)), we have that

$$h(AX\rightarrow BX)=h(AXBX)-h(AX)=h(ABX)-h(AX)=h(B|AX)=h(AX\rightarrow B).$$

We now prove that $h(AX\rightarrow B)=I(B;\Omega\backslash(ABX)|AX)+I(B;B|\Omega\backslash B)$.

$$\begin{aligned}&I(B;\Omega\backslash(ABX)|AX)+I(B;B|\Omega\backslash B)\\&=h(ABX)+h(\Omega\backslash B)-h(AX)-h(\Omega)+2h(\Omega)-h(\Omega\backslash B)-h(\Omega)\\&=h(ABX)-h(AX)\\&\overset{\text{def}}{=}h(B|AX)\overset{\text{def}}{=}h(AX\rightarrow B).\end{aligned}$$

□

Lemma 2.2 implies that for every conditional $A\rightarrow B$, we may assume that $A\cap B=\emptyset$. Furthermore, lemma 2.2 establishes that every conditional $A\rightarrow B$ can be expressed as the sum of two saturated CIs. That is, $h(A\rightarrow B)=I(B;\Omega\backslash(AB)|A)+I(B;B|\Omega\backslash B)$. This will be important for establishing our result on relaxation of implications from the set of saturated CIs and conditionals.

3. Exact Implication and its Role in PGMs

In this section, we formally define the notions of exact and approximate implication, and their role in undirected and directed PGMs. This provides the appropriate context for which to present our results on approximate implication in later sections. We fix a set of variables $\Omega=\{X_1,\dots,X_n\}$, and consider triples of the form $\sigma=(X;Y|Z)$, where $X,Y,Z\subseteq\Omega$, which we call a *conditional independence*, CI. An *implication* is a formula $\Sigma\Rightarrow\tau$, where Σ is a set of CIs called *antecedents* and τ is a CI called *consequent*. For an n-dimensional polymatroid $h\in\Gamma_n$, and a CI $\sigma=(X;Y|Z)$, we define $h(\sigma)\overset{\text{def}}{=}I_h(X;Y|Z)$ (see (2)), for a set of CIs Σ, we define $h(\Sigma)\overset{\text{def}}{=}\sum_{\sigma\in\Sigma}h(\sigma)$. We denote by $\mathbf{var}(\sigma)$ the set of RVs mentioned in σ (e.g., if $\sigma=(X_1X_2;X_3|X_4)$ then $\mathbf{var}(\sigma)=\{X_1,X_2,X_3,X_4\}$). Fix a set K s.t. $K\subseteq\Gamma_n$.

Definition 3.1. *The* exact implication *(EI) $\Sigma\Rightarrow\tau$ holds in K, denoted $K\models_{EI}\Sigma\Rightarrow\tau$ if, for all $h\in K$, $h(\Sigma)=0$ implies $h(\tau)=0$. The* λ-approximate implication *(λ-AI) holds in K, in notation $K\models\lambda\cdot h(\Sigma)\geq h(\tau)$, if $\forall h\in K$, $\lambda\cdot h(\Sigma)\geq h(\tau)$. The* approximate implication *holds, in notation $K\models_{AI}(\Sigma\Rightarrow\tau)$, if there exist a finite $\lambda\geq 0$ such that the λ-AI holds.*

Notice that both exact (EI) and approximate (AI) implications are preserved under subsets of K: if $K_1 \subseteq K_2$ and $K_2 \models_x \Sigma \Rightarrow \tau$, then $K_1 \models_x \Sigma \Rightarrow \tau$, for $x \in \{\text{EI,AI}\}$.

Approximate implication always implies its exact counterpart. Indeed, if $h(\tau) \leq \lambda \cdot h(\Sigma)$ and $h(\Sigma) = 0$, then $h(\tau) \leq 0$, which further implies that $h(\tau) = 0$, because $h(\tau) \geq 0$ for every triple τ, and every polymatroid h. In this paper we study the reverse.

Definition 3.2. *Let $\mathcal{L}$ be a syntactically-defined class of implication statements $(\Sigma \Rightarrow \tau)$, and let $K \subseteq \Gamma_n$. We say that $\mathcal{L}$* admits a λ-relaxation *in K, if every exact implication statement $(\Sigma \Rightarrow \tau)$ in $\mathcal{L}$ has a λ-approximation:*

$$K \models_{EI} \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad K \models \lambda \cdot h(\Sigma) \geq h(\tau).$$

Example 3.3. *Let $\Sigma = \{(A;B), (A;C|B)\}$, and $\tau = (A;C)$. Since $I_h(A;C) \leq I_h(A;BC)$, and since $I_h(A;BC) = I_h(A;B) + I_h(A;C|B)$ by the chain rule* (3)*, then the exact implication $\Gamma_n \models \Sigma \Rightarrow \tau$ admits an approximate implication with $\lambda = 1$ (i.e., a 1-AI).*

In this paper, we focus on λ-relaxation in different subsets of Γ_n, and three syntactically-defined classes: 1) Where Σ is a set of saturated CIs and conditionals (Section 3.1), 2) Where Σ is the recursive basis of a Bayesian network (Section 3.2), and 3) Where Σ is a set of marginal CIs. In what follows, we will use $K \models \Sigma \Rightarrow \tau$ as shorthand for $K \models_{EI} \Sigma \Rightarrow \tau$ for any subset of polymatroids $K \subseteq \Gamma_n$. If, in addition, K is clear from the context, we will just use $\Sigma \Rightarrow \tau$. Also, we will use $\lambda \cdot h(\Sigma) \geq h(\tau)$ as shorthand for $\Gamma_n \models \lambda \cdot h(\Sigma) \geq h(\tau)$.

3.1 Markov Networks and Implication from Saturated CIs and Conditionals

Recall that a CI $(X;Y|Z)$ is saturated if $XYZ = \Omega$, and that conditionals are a special case of saturated CIs (Lemma 2.2).

Theorem 3.4. (Geiger & Pearl, 1993; Beeri et al., 1977) *Let Σ be a set of saturated CIs over the set $\Omega \stackrel{\text{def}}{=} \{X_1, \dots, X_n\}$ of random variables, and let Σ^+ denote the closure of Σ with respect to the semigraphoid axioms. Let τ be a saturated CI over Ω. Then*

$$\Gamma_n \models \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad \tau \in \Sigma^+.$$

Theorem 3.4 establishes that the semigraphoid axioms are sound and complete for inferring saturated CIs from a set of saturated CIs (and conditionals). Gyssens et al. (Gyssens, Niepert, & Gucht, 2014) improve this result by dropping any restrictions on the consequent τ. In Section 7, we prove that if Σ is a set of saturated CIs and conditionals, then the exact implication $\Gamma_n \models \Sigma \Rightarrow \tau$ has an $n/2$-relaxation for any implied CI τ. In other words, $\Gamma_n \models n/2 \cdot h(\Sigma) \geq h(\tau)$.

When restricted to polymatroids that are also graphoids, then CI relations can be represented by an undirected graph. Let $G(V,E)$ be an undirected graph, and let $u,v \in V$. We say that u and v are *adjacent* if $(u,v) \in E$. A *path* $t = (v_1, \dots, v_n)$ is a sequence of vertices $(v_1, \dots, v_n)$ such that $(v_i, v_{i+1}) \in E$ for every $i \in \{1, \dots, n-1\}$. We say that u and v are *connected* if there is a path $(u = v_1, \dots, v_n = v)$ starting at u and ending at v; otherwise, we say that u and v are *disconnected*. Let $X, Y \subseteq V$ be disjoint sets of vertices. We say that X and Y are disconnected if x and y are disconnected for every $x \in X$ and $y \in Y$. Let $V' \subseteq V$. The graph *induced by V'* denoted $G[V']$ is the graph $G'(V', E')$ where

$E' \overset{\text{def}}{=} \{(u,v) \in E \mid u,v \in V'\}$. We say that $Z \subseteq V$ is an *X,Y-separator* if, in the graph $G[V \backslash Z]$, that results from G by removing the vertex-set Z and the edges adjacent to Z, X and Y are disconnected. For an undirected graph $G(V,E)$, we denote by $X \perp_G Y|Z$ the fact that Z is an X,Y-separator.

Let p be a joint probability distribution over the random variables $\Omega = \{X_1, \dots, X_n\}$. We define $\Sigma_{\text{pair}} \overset{\text{def}}{=} \{(u;v|\Omega \backslash uv) : p \models u \perp v | \Omega \backslash uv\}$. That is, Σ_{pair} is the set of vertex-pairs that are conditionally independent given the rest of the variables. Observe that every CI in Σ_{pair} is, by definition, a saturated CI. We define the *independence graph* $G(\Omega, E)$, where $E \overset{\text{def}}{=} \{(u,v) : (u;v|\Omega \backslash uv) \notin \Sigma_{\text{pair}}\}$. That is, the edges of the independence graph are between vertices that are not independent given the rest of the variables.

Theorem 3.5. (Geiger & Pearl, 1993) *Let p be a joint probability distribution over the random variables $\Omega = \{X_1, \dots, X_n\}$, with entropy function h_p. Let $G(\Omega, E)$ be the independence graph generated from $\Sigma_{pair} \overset{\text{def}}{=} \{(u;v|\Omega \backslash uv) : p \models u \perp v | \Omega \backslash uv\}$. Let Σ^+_{pair} denote the closure of Σ_{pair} with respect to the graphoid axioms. If h_p is a graphoid (e.g., p is strictly positive), then for any three disjoint sets $X, Y, Z \subseteq \Omega$, where $XYZ = \Omega$, it holds that*

$$(X;Y|Z) \in \Sigma^+_{pair} \qquad \textit{if and only if} \qquad X \perp_G Y|Z.$$

Theorem 3.5 establishes that the graphoid axioms are sound and complete for inferring saturated CIs from Σ_{pair} which, in turn, correspond to graph-separation in the independence graph. In Section 5, we show that the intersection axiom does not relax. An immediate consequence is that no relaxation exists for exact implications whose derivation includes the intersection axiom. Specifically, if an implication of the form $\Sigma_{\text{pair}} \Rightarrow (X;Y|Z)$ requires the application of the intersection axiom, then it does not translate to an inequality $I(X;Y|Z) \leq \lambda h(\Sigma_{\text{pair}})$ that holds for all positive distributions, and where λ is finite. Practically, this means that the current method of inferring CIs in Markov Networks does not extend to the case where the CIs in Σ_{pair} do not hold exactly; we illustrate in Example 3.6.

Example 3.6. *Let $p(A,B,C,D)$ be a strictly positive joint probability distribution over the RVs A, B, C, and D. That is, p is a graphoid. Let $\Sigma_{pair} \overset{\text{def}}{=} \{A \perp C|BD, A \perp D|BC\}$. The independence graph associated with Σ_{pair} is presented in Figure 1. Since p is a graphoid, then by Theorem 3.5, it holds that $p \models A \perp CD|B$ because B is an A, CD-separator in the independence graph. In fact, the CI $A \perp CD|B$ is derived from the intersection axiom (see* (7)*): $A \perp C|BD, A \perp D|BC \Rightarrow A \perp CD|B$.*

Let $\varepsilon > 0$ be a small constant. In Section 5, we show that even if $I(A;C|BD) \leq \varepsilon$ and $I(A;D|BC) \leq \varepsilon$, there is no guaranteed upper bound on the value of $I(A;CD|B)$ which holds for all strictly positive distributions.

3.2 Bayesian Networks

Let $G(V,E)$ be a Directed Acyclic Graph (DAG), and let $u, v \in V$. We say that u is a *parent* of v, and v a *child* of u if $(u \to v) \in E$. A *directed path* $t = (v_1, v_2, \dots, v_n)$ is a sequence of vertices $(v_1, v_2, \dots, v_n)$ such that there is an edge $(v_i \to v_{i+1}) \in E$ for every $i \in \{1, \dots, n-1\}$. We say that v is a *descendant* of u, and u an *ancestor* of v if there is a directed path from u to v. A *trail* $t = (v_1, v_2, \dots, v_n)$ is a sequence of vertices

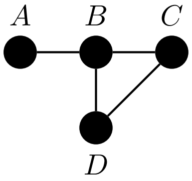

Figure 1: Example 3.6 of implication from Σ_{pair}.

$(v_1, v_2, \ldots, v_n)$ such that there is an edge between v_i and v_{i+1} for every $i \in \{1, \ldots, n-1\}$. That is, $(v_i \rightarrow v_{i+1}) \in E$ or $(v_i \leftarrow v_{i+1}) \in E$ for every $i \in \{1, \ldots, n-1\}$. A vertex v_i is said to be *head-to-head* with respect to t if $(v_{i-1} \rightarrow v_i) \in E$ and $(v_i \leftarrow v_{i+1}) \in E$. A trail $t = (v_1, v_2, \ldots, v_n)$ is *active* given $Z \subseteq V$ if (1) every v_i that is a head-to-head vertex with respect to t either belongs to Z or has a descendant in Z, and (2) every v_i that is not a head-to-head vertex with respect to t does not belong to Z. If a trail t is not active given Z, then it is *blocked* given Z. Let $X, Y, Z \subseteq V$ be pairwise disjoint. We say that Z *d-separates* X from Y if every trail between $x \in X$ and $y \in Y$ is blocked given Z. We denote by $X \perp_{dsep} Y | Z$ that Z d-separates X from Y in G.

A Bayesian network encodes the CIs of a probability distribution using a DAG. Each node X_i in a Bayesian network corresponds to the variable $X_i \in \Omega$, a set of nodes α correspond to the set of variables X_α, and $x_i \in \mathcal{D}_i$ is a value from the domain of X_i. Each node X_i in the network represents the distribution $p(X_i \mid X_{\pi(i)})$ where $X_{\pi(i)}$ is a set of variables that correspond to the parent nodes $\pi(i)$ of i. The distribution represented by a Bayesian network is

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi(i)}). \tag{9}$$

(when i has no parents then $X_{\pi(i)} = \emptyset$).

Equation 9 implicitly encodes a set of n conditional independence statements, called the *recursive basis* for the network:

$$\Sigma_{\text{RB}} \stackrel{\text{def}}{=} \{(X_i; X_1 \ldots X_{i-1} \backslash \pi(X_i) \mid \pi(X_i)) : i \in [n]\}. \tag{10}$$

The implication problem associated with Bayesian Networks is to determine whether $\Gamma_n \models \Sigma_{\text{RB}} \Rightarrow \tau$ for a CI τ. Let $G(\Omega, E)$ be a DAG generated by the recursive basis Σ_{RB}. That is, the vertices of G are the RVs Ω, and its edges are $E = \{X_i \rightarrow X_j | X_i \in \pi(X_j)\}$. Given a CI $\tau = (A; B|C)$, the d-separation algorithm efficiently determines whether C d-separates A from B in G. It has been shown that $\Gamma_n \models \Sigma_{\text{RB}} \Rightarrow \tau$ if and only if C d-separates A from B (Geiger et al., 1990).

Theorem 3.7. (Geiger et al., 1990; Geiger & Pearl, 1988; Verma & Pearl, 1988) *Let $f \in \Gamma_n$ be a polymatroid, and $G(\Omega, E)$ be the DAG generated by the recursive basis Σ_{RB} (see* (10)*). Let Σ_{RB}^+ denote the closure of Σ_{RB} with respect to the semigraphoid axioms. The following holds for every three disjoint sets $A, B, C \subseteq \Omega$:*

$$\Gamma_n \models \Sigma_{RB} \Rightarrow (A; B|C) \qquad \textit{iff} \qquad (A; B|C) \in \Sigma_{RB}^+ \qquad \textit{iff} \qquad A \perp_{dsep} B | C.$$

Theorem 3.7 establishes that both the semigraphoid axioms, and the d-separation criterion, are sound and complete for inferring CI statements from the recursive basis. Since the semigraphoid axioms follow from the Shannon inequalities ((1) and (2)), Theorem 3.7 estsablishes that the Shannon inequalities are both sound and complete for inferring CI statements from the recursive basis. In Section 8, we show that the exact implication $\Gamma_n \models \Sigma_{\text{RB}} \Rightarrow (A;B|C)$ admits a 1-relaxation.

4. Formal Statement of Results and Practical Implications

In this section, we formally state the results proved in the paper. We begin by establishing a negative result concerning the intersection axiom (see (7)).

Theorem 4.1 (The intersection axiom does not relax)**.** *For any finite $\lambda > 0$, there exists a strictly positive probability distribution $p_\lambda(A,B,C)$, such that*

$$\lambda \left(I_{h_{p_\lambda}}(A;B|C) + I_{h_{p_\lambda}}(A;C|B) \right) < I_{h_{p_\lambda}}(A;BC),$$

where h_{p_λ} is the entropy function of p_λ.

Theorem 4.1 establishes that the intersection axiom (see (7)) does not relax, even when restricted to strictly positive distributions! This result has the following practical implication. Recall from Theorem 3.5 that every X,Y-separator Z in the independence graph $G(V,E)$, generated using the saturated set of CIs $\Sigma_{\text{pair}} \overset{\text{def}}{=} \{(u;v|V\backslash uv) : I_h(u;v|V\backslash uv) = 0\}$, corresponds to the CI $I_h(X;Y|Z) = 0$. In other words, by Theorem 3.5, the exact implication $\Sigma_{\text{pair}} \Rightarrow (X;Y|Z)$ holds for all graphoids. We ask whether the implication $\Sigma_{\text{pair}} \Rightarrow (X;Y|Z)$, derived using the graphoid axioms, and the intersection axiom in particular, translates to an inequality of the form $\lambda \cdot h(\Sigma_{\text{pair}}) \geq I(X;Y|Z)$, where λ is finite and holds for all strictly positive probability distributions. Theorem 4.1 establishes that the answer is negative. In other words, approximate CIs cannot be derived from Σ_{pair}.

Colollary 4.2. *There exist Σ_{pair}, τ with three RVs, such that $\Sigma_{pair} \Rightarrow \tau$ holds for all graphoids, but it does not relax.*

Proof. Let $\Sigma_{\text{pair}} \overset{\text{def}}{=} \{A \perp B|C, A \perp C|B\}$. By the intersection axiom, $\Sigma_{\text{pair}} \Rightarrow A \perp BC$ holds for all graphoids. By Theorem 4.1, the implication $\Sigma_{\text{pair}} \Rightarrow A \perp BC$ does not relax. □

In stark contrast to the negative result of Theorem 4.1, we show that approximate CIs can be derived using the semigraphoid axioms, and the d-separation algorithm in Bayesian networks.

We recall that a *conditional* is a statement of the form $A \to B$, which holds in a probability distribution p if and only if $h(B|A) = 0$.

Theorem 4.3. *Let Σ be a set of saturated CIs and conditionals over the set of variables $\Omega \overset{\text{def}}{=} \{X_1, \dots, X_n\}$, and let $\tau \overset{\text{def}}{=} (A;B|C)$ be any CI. Then*

$$\Gamma_n \models \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad \Gamma_n \models I_h(A;B|C) \leq \min\{|A|,|B|\} h(\Sigma).$$

Theorem 4.3 generalizes Theorem 3.4 by establishing that every exact implication from a set of saturated CIs and conditionals relaxes to the inequality $\Gamma_n \models I_h(A;B|C) \leq \min\{|A|,|B|\}h(\Sigma)$. In other words, the implication is derived from the inequality. Note that the only-if direction of Theorem 4.3 is immediate, and follows from the non-negativity of Shannon's information measures. In previous work (Kenig & Suciu, 2022), it was shown that $\Gamma_n \models \Sigma \Rightarrow \tau$ if and only if $\Gamma_n \models I_h(A;B|C) \leq |A|{\cdot}|B|{\cdot}h(\Sigma) \leq {}^{n^2}\!/4 \cdot h(\Sigma)$. Noting that $\min\{|A|,|B|\} \leq {}^{n}\!/2$, the result of Theorem 4.3 tightens the bound by an order of magnitude.

Theorem 4.4. *Let Σ be a recursive set of CIs (see* (10)*), and let $\tau = (A;B|C)$. Then*

$$\Gamma_n \models \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad \Gamma_n \models h(\Sigma) \geq h(\tau). \tag{11}$$

The result of Theorem 4.4 has the following practical implication. Theorem 3.7 establishes that if X and Y are d-separated given Z in the DAG $G(V,E)$, generated by the recursive basis Σ_{RB} (see (10)), then $I_h(X;Y|Z) = 0$. Now, let $\varepsilon > 0$, and suppose that for every $(X_i; X_1 \dots X_{i-1}|\pi(X_i)) \in \Sigma_{\text{RB}}$, it holds that $I_h(X_i; X_1 \dots X_{i-1}|\pi(X_i)) \leq \varepsilon$. We ask whether we can bound the value of $I_h(X;Y|Z)$. Theorem 4.4 establishes that $I_h(X;Y|Z) \leq h(\Sigma_{\text{RB}})$. Theorem 4.4 was first proved in Kenig (Kenig, 2021). In this paper, we simplify the proof, and relate it to d-separation.

Finally, we consider implications from the set of marginal CIs.

Theorem 4.5. *Let Σ be a set of marginal CIs, and $\tau = (A;B|C)$ be any CI. Then*

$$\Gamma_n \models \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad \Gamma_n \models I_h(A;B|C) \leq \min\{|A|,|B|\}h(\Sigma). \tag{12}$$

Theorem 4.5 generalizes the result of (Geiger et al., 1991b), which proved that the semi-graphoid axioms are sound and complete for deriving marginal CIs. In previous work (Kenig, 2021), it was shown that $\Gamma_n \models \Sigma \Rightarrow \tau$ if and only if $\Gamma_n \models I_h(A;B|C) \leq |A|{\cdot}|B|{\cdot}h(\Sigma) \leq {}^{n^2}\!/4 \cdot h(\Sigma)$. Noting that $\min\{|A|,|B|\} \leq {}^{n}\!/2$, the result of Theorem 4.5 tightens the bound by an order of magnitude.

5. Intersection Axiom Does Not Relax

It is well-known that the intersection axiom (7) does not hold for all probability distributions. From this, we can immediately conclude that the intersection axiom does not relax for all polymatroids. This follows from the fact that the entropic function associated with any probability distribution is a polymatroid, and that approximate implication generalizes exact implication. In this section, we prove Theorem 4.1, establishing that the intersection axiom does not relax even for strictly positive probability distributions. In other words, we prove the (surprising) result that even for the class of distributions in which the intersection axiom holds, it does not relax.

Let $p(A,B,C)$ be a strictly positive probability distribution, where $I_h(A;B|C) = 0$ and $I_h(A;C|B) = 0$. According to the intersection axiom (7), it holds that $I_h(A;BC) = 0$.

To prove the Theorem, we describe the following "helper" distribution. Let $A_1, \dots, A_7$ denote mutually independent, random variables. The random variable A_7 is defined $A_7 \overset{\text{def}}{=} (A_{7,1}, A_{7,2}, A_{7,3})$, where $A_{7,j}$ is a binary RV for all $j \in \{1,2,3\}$. For $i \in \{4,5,6\}$,

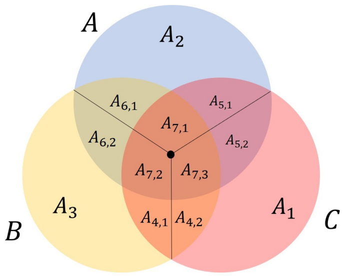

Figure 2: The information diagram for the joint probability $p(A, B, C)$ where A, B and C are defined in (13)–(15).

$A_i \overset{\text{def}}{=} (A_{i,1}, A_{i,2})$, where $A_{i,1}$ and $A_{i,2}$ are binary RVs. The joint distributions of $A_7 \overset{\text{def}}{=} (A_{7,1}, A_{7,2}, A_{7,3})$, and $A_i \overset{\text{def}}{=} (A_{i,1}, A_{i,2})$ for $i \in \{4, 5, 6\}$, are defined as follows:

$A_{7,1}$	$A_{7,2}$	$A_{7,3}$	P
0	0	0	$\frac{1}{2} - 3y$
0	0	1	y
0	1	0	y
0	1	1	y
1	0	0	y
1	0	1	y
1	1	0	y
1	1	1	$\frac{1}{2} - 3y$

Table 2: Joint distribution of $A_7 \overset{\text{def}}{=} (A_{7,1}, A_{7,2}, A_{7,3})$, where $y \in (0, \frac{1}{6})$.

$A_{i,1}$	$A_{i,2}$	P
0	0	$1 - 3x$
0	1	x
1	0	x
1	1	x

Table 3: Joint distribution of $A_i \overset{\text{def}}{=} (A_{i,1}, A_{i,2})$ for $i \in \{4, 5, 6\}$, where $x \in (0, \frac{1}{3})$.

Finally, for A_1, A_2, A_3 we have that $P(A_i = 1) = P(A_i = 0) = \frac{1}{2}$.
We define the RVs A, B, and C as follows:

$$A \overset{\text{def}}{=} (A_2, A_{6,1}, A_{7,1}, A_{5,1}), \tag{13}$$

$$B \overset{\text{def}}{=} (A_3, A_{6,2}, A_{7,2}, A_{4,1}), \text{ and} \tag{14}$$

$$C \overset{\text{def}}{=} (A_1, A_{5,2}, A_{7,3}, A_{4,2}). \tag{15}$$

The information diagram corresponding to $P(A, B, C)$ is presented in Figure 2.

Lemma 5.1. *The joint distribution $P : \mathcal{D}(A) \times \mathcal{D}(B) \times \mathcal{D}(C) \to (0, 1)$ is strictly positive.*

Proof. Take any assignment $d \in \mathcal{D}(A) \times \mathcal{D}(B) \times \mathcal{D}(C)$. Since A, B, and C have no common RVs, then d uniquely maps to an assignment to the binary RVs A_1,A_2,A_3,$A_{4,1}$,$A_{4,2}$,$A_{5,1}$, etc. In particular, d maps to a unique assignment to $A_1, A_2, A_3, A_4, A_5, A_6$, and A_7. For all $i \in \{1, \ldots, 7\}$, denote by d_i the assignment to A_i induced by d. By definition, these RVs are mutually independent. Hence

$$\begin{aligned} P(d) &= P(A_1 = d_1, A_2 = d_2, A_3 = d_3, A_4 = d_4, A_5 = d_5, A_6 = d_6, A_7 = d_7) \\ &= P(A_1 = d_1) \cdots P(A_7 = d_7). \end{aligned}$$

By definition, for every A_i it holds that $p(A_i)$ is strictly positive (see Tables 2 and 3). This proves the claim. □

We define the following functions where $x \in (0, \frac{1}{3})$ and $y \in (0, \frac{1}{6})$:

$$\delta_1(x) \stackrel{\mathsf{def}}{=} -\big((1-3x)\log(1-3x) + 3x\log x\big) \qquad \text{where } x \in (0, \frac{1}{3}), \tag{16}$$

$$\delta_2(x) \stackrel{\mathsf{def}}{=} -\big((1-2x)\log(1-2x) + 2x\log 2x\big) \qquad \text{where } x \in (0, \frac{1}{3}), \tag{17}$$

$$f_1(y) \stackrel{\mathsf{def}}{=} -\left(2(\frac{1}{2} - 3y)\log(\frac{1}{2} - 3y) + 6y\log y\right) \qquad \text{where } y \in (0, \frac{1}{6}), \text{ and} \tag{18}$$

$$f_2(y) \stackrel{\mathsf{def}}{=} -\left(2(\frac{1}{2} - 2y)\log(\frac{1}{2} - 2y) + 4y\log 2y\right) \qquad \text{where } y \in (0, \frac{1}{6}). \tag{19}$$

By definition of $\delta_1(x)$, $\delta_2(x)$, $f_1(y)$, and $f_2(y)$ (see (16)–(19)), it easily follows that

$$\lim_{x \to 0} \delta_1(x) = \lim_{x \to 0} \delta_2(x) = 0 \text{ and} \tag{20}$$

$$\lim_{y \to 0} f_1(y) = \lim_{y \to 0} f_2(x) = 1. \tag{21}$$

Lemma 5.2. *The following holds for any $x \in (0, \frac{1}{3})$, and $y \in (0, \frac{1}{6})$.*

1. *$H(A_i) = 1$ for $i \in \{1, 2, 3\}$.*
2. *$H(A_i) = H(A_{i,1}, A_{i,2}) = \delta_1(x)$ for $i \in \{4, 5, 6\}$.*
3. *$H(A_{i,1}) = H(A_{i,2}) = \delta_2(x)$ for $i \in \{4, 5, 6\}$.*
4. *$H(A_{7,j}) = 1$ for $j \in \{1, 2, 3\}$.*
5. *$H(A_{7,j}, A_{7,k}) = f_2(y)$ for $j \neq k$ and $j, k \in \{1, 2, 3\}$.*
6. *$H(A_7) = H(A_{7,1}, A_{7,2}, A_{7,3}) = f_1(y)$.*

where $\delta_1, \delta_2, f_1,$ and f_2 are defined in (16)–(19).

Proof Overview. The proof follows from the definition of the RVs $A_1, A_2, \ldots, A_7$ in Tables 2, and 3. The complete technical details are deferred to Section A in the Appendix. □

Lemma 5.3. *The following holds:*

1. *$H(A) = H(B) = H(C) = 2 + 2\delta_2(x)$.*
2. *$H(AC) = H(AB) = H(BC) = 2 + \delta_1(x) + 2\delta_2(x) + f_2(y)$.*
3. *$H(ABC) = 3 + 3\delta_1(x) + f_1(y)$.*

Proof Overview. The proof follows from the definition of the RVs $A_1, A_2, \ldots, A_7$ (see Tables 2, and 3), the fact that they are mutually independent, the definition of RVs A, B, and C (see (13)–(15)), and the application of Lemma 5.2. The complete technical details are deferred Section A in the Appendix. □

An immediate consequence from Lemma 5.3 is that

$$\begin{aligned} I(A;B|C) &= H(AC)+H(BC)-H(C)-H(ABC) \\ &= 2(\delta_1(x)+f_2(y)+2+2\delta_2(x))-(2+2\delta_2(x))-(3+3\delta_1(x)+f_1(y)) \\ &= 2\delta_2(x)-\delta_1(x)+2f_2(y)-f_1(y)-1. \end{aligned} \tag{22}$$

By symmetry, $I(A;C|B)=I(B;C|A)=2\delta_2(x)-\delta_1(x)+2f_2(y)-f_1(y)-1$ as well. And,

$$\begin{aligned} I(A;B) &= H(A)+H(B)-H(AB) \\ &= 2(2+2\delta_2(x))-(2+\delta_1(x)+2\delta_2(x)+f_2(y)) \\ &= 4+4\delta_2(x)-2-\delta_1(x)-2\delta_2(x)-f_2(y) \\ &= 2\delta_2(x)-\delta_1(x)-f_2(y)+2. \end{aligned} \tag{23}$$

By symmetry, $I(A;C)=I(B;C)=2\delta_2(x)-\delta_1(x)-f_2(y)+2$ as well.

Theorem 4.1. *For any finite $\lambda>0$, there exists a strictly positive probability distribution $p_\lambda(A,B,C)$, such that*

$$\lambda\left(I_{h_{p_\lambda}}(A;B|C)+I_{h_{p_\lambda}}(A;C|B)\right)<I_{h_{p_\lambda}}(A;BC),$$

where h_{p_λ} is the entropy function of p_λ.

Proof. Suppose otherwise, and consider the joint probability distribution $p:\mathcal{D}(A)\times\mathcal{D}(B)\times\mathcal{D}(C)\rightarrow(0,1)$, where A, B and C are defined in (13)–(15). By Lemma 5.1, p is a strictly positive distribution for all $x\in(0,\frac{1}{3})$, and $y\in(0,\frac{1}{6})$. Then there exists a fixed, finite λ such that for all $x\in(0,\frac{1}{3})$, and all $y\in(0,\frac{1}{6})$ it holds that

$$\lambda(I(A;B|C)+I(A;C|B))\geq I(A;B).$$

From (22) and (23), it means that for all $x\in(0,\frac{1}{3})$, and all $y\in(0,\frac{1}{6})$ it holds that

$$2\lambda(2\delta_2(x)-\delta_1(x)+2f_2(y)-f_1(y)-1)\geq 2\delta_2(x)-\delta_1(x)-f_2(y)+2. \tag{24}$$

By the assumption, (24) holds for all $x\in(0,\frac{1}{3})$, and all $y\in(0,\frac{1}{6})$. In particular, it holds in the limits $x\rightarrow 0$ and $y\rightarrow 0$. That is,

$$\lim_{\substack{x\rightarrow 0,\\ y\rightarrow 0}}\left(2\lambda(2\delta_2(x)-\delta_1(x)+2f_2(y)-f_1(y)-1)\right)\underbrace{\geq}_{(24)}\lim_{\substack{x\rightarrow 0,\\ y\rightarrow 0}}\left(2\delta_2(x)-\delta_1(x)-f_2(y)+2\right). \tag{25}$$

From (20) and (21), we have that $\lim_{x\rightarrow 0}\delta_1(x)=\lim_{x\rightarrow 0}\delta_2(x)=0$, and $\lim_{y\rightarrow 0}f_1(y)=\lim_{y\rightarrow 0}f_2(x)=1$. Substituting into (25), we get that

$$2\lambda\left(2\cdot 0-0+2-1-1\right)\geq\left(2\cdot 0-0-1+2\right)\text{, and thus}$$
$$2\lambda\cdot 0\geq 1.$$

Hence, no such finite λ exists, and the intersection axiom does not relax. □

Information Measures	μ^*
$H(X)$	$\mu^*(\mathrm{m}(X))$
$H(XY)$	$\mu^*\left(\mathrm{m}(X) \cup \mathrm{m}(Y)\right)$
$H(X\|Y)$	$\mu^*\left(\mathrm{m}(X) \cap \mathrm{m}^c(Y)\right)$
$I_H(X;Y)$	$\mu^*\left(\mathrm{m}(X) \cap \mathrm{m}(Y)\right)$
$I_H(X;Y\|Z)$	$\mu^*\left(\mathrm{m}(X) \cap \mathrm{m}(Y) \cap \mathrm{m}^c(Z)\right)$

Table 4: Information measures and associated I-Measure.

6. Properties of Exact Implication

This section proves various technical lemmas that establish some general properties of exact implication in the set Γ_n of n-dimensional polymatroids, and a certain subset of polymatroids called *positive polymatroids*, to be defined later. The lemmas in this section will be used for proving the approximate implication guarantees presented in Section 4. A central tool in our analysis of exact and approximate implication is the *I-Measure theory* (Yeung, 1991, 2008). We present the required background on the I-Measure theory in Section 6.1.

In what follows, $\Omega \stackrel{\text{def}}{=} \{X_1, \dots, X_n\}$ is a set of n jointly-dstributed RVs, Σ denotes a set of triples $(A;B|C)$, and τ denotes a single triple. We denote by $\mathbf{var}(\sigma)$ the set of RVs mentioned in σ (e.g., if $\sigma = (X_1X_2;X_3|X_4)$ then $\mathbf{var}(\sigma) = \{X_1, X_2, X_3, X_4\}$).

6.1 The I-Measure

The I-Measure (Yeung, 1991, 2008) is a theory which establishes a one-to-one correspondence between Shannon's information measures and set theory. Let $h \in \Gamma_n$ denote a polymatroid defined over the variables $\{X_1, \dots, X_n\}$. Every variable X_i is associated with a set $\mathrm{m}(X_i)$, and its complement $\mathrm{m}^c(X_i)$. The universal set is $\Lambda \stackrel{\text{def}}{=} \bigcup_{i=1}^n \mathrm{m}(X_i)$. Let $\alpha \subseteq [n]$. We denote by $X_\alpha \stackrel{\text{def}}{=} \{X_j \mid j \in \alpha\}$. For the variable-set X_α, we define

$$\mathrm{m}(X_\alpha) \stackrel{\text{def}}{=} \bigcup_{i \in \alpha} \mathrm{m}(X_i) \qquad \text{and} \qquad \mathrm{m}^c(X_\alpha) \stackrel{\text{def}}{=} \bigcap_{i \in \alpha} \mathrm{m}^c(X_i). \tag{26}$$

Definition 6.1. (Yeung, 1991, 2008) *The field $\mathcal{F}_n$ generated by sets $\mathrm{m}(X_1), \dots, \mathrm{m}(X_n)$ is the collection of sets which can be obtained by any sequence of usual set operations (union, intersection, complement, and difference) on $\mathrm{m}(X_1), \dots, \mathrm{m}(X_n)$.*

The *atoms* of $\mathcal{F}_n$ are sets of the form $\bigcap_{i=1}^n Y_i$, where Y_i is either $\mathrm{m}(X_i)$ or $\mathrm{m}^c(X_i)$. We denote by $\mathcal{A}$ the atoms of $\mathcal{F}_n$. We consider only atoms in which at least one set appears in positive form (i.e., the atom $\bigcap_{i=1}^n \mathrm{m}^c(X_i) \stackrel{\text{def}}{=} \emptyset$ is empty). There are $2^n - 1$ non-empty atoms and 2^{2^n-1} sets in $\mathcal{F}_n$ expressed as the union of its atoms. A function $\mu : \mathcal{F}_n \to \mathbb{R}$ is *set additive* if, for every pair of disjoint sets A and B, it holds that $\mu(A \cup B) = \mu(A) + \mu(B)$. A real function μ defined on $\mathcal{F}_n$ is called a *signed measure* if it is set additive, and $\mu(\emptyset) = 0$.

The *I*-Measure μ^* on $\mathcal{F}_n$ is defined by $\mu^*(m(X_\alpha)) \stackrel{\text{def}}{=} H(X_\alpha)$ for all nonempty subsets $\alpha \subseteq \{1, \dots, n\}$, where H is the entropy (4). Table 4 summarizes the extension of this definition to the rest of the Shannon measures. Yeung's I-Measure Theorem establishes the one-to-one correspondence between Shannon's information measures and μ^*.

Theorem 6.2. (Yeung, 1991, 2008) [I-Measure Theorem] *μ^* is the unique signed measure on $\mathcal{F}_n$ which is consistent with all Shannon's information measures (i.e., entropies, conditional entropies, mutual information, and conditional mutual information).*

Let $\sigma = (X;Y|Z)$. We denote by $\mathrm{m}(\sigma) \overset{\text{def}}{=} \mathrm{m}(X) \cap \mathrm{m}(Y) \cap \mathrm{m}^c(Z)$ the set associated with σ (see Table 4). For a set of triples Σ, we define

$$\mathrm{m}(\Sigma) \overset{\text{def}}{=} \bigcup_{\sigma \in \Sigma} \mathrm{m}(\sigma). \tag{27}$$

Example 6.3. *Let A, B, and C be three disjoint sets of RVs defined as follows: $A{=}A_1A_2A_3$, $B{=}B_1B_2$ and $C{=}C_1C_2$. By Theorem 6.2:*

$$\begin{aligned} H(A) &= \mu^*(\mathrm{m}(A)) = \mu^*(\mathrm{m}(A_1) \cup \mathrm{m}(A_2) \cup \mathrm{m}(A_3)), \\ H(B) &= \mu^*(\mathrm{m}(B)) = \mu^*(\mathrm{m}(B_1) \cup \mathrm{m}(B_2)), \text{ and} \\ \mu^*(\mathrm{m}^c(C)) &= \mu^*(\mathrm{m}^c(C_1) \cap \mathrm{m}^c(C_2)). \end{aligned}$$

By Table 4, $I(A;B|C){=}\mu^*(\mathrm{m}(A) \cap \mathrm{m}(B) \cap \mathrm{m}^c(C))$.

Theorem 6.2 establishes that every polymatroid $h \in \mathbb{R}^{2^n}$ is associated with a unique signed measure $\mu^* : \mathcal{F}_n \to \mathbb{R}$, termed I-Measure. The I-Measure is not necessarily positive for entropic functions, as illustrated in the following example.

Example 6.4. *Let A, B, C be binary RVs, we define the* parity distribution *as follows:*

$$p(a,b,c) = \begin{cases} \frac{1}{4} & \textit{if } a+b+c \mod 2 = 0 \\ 0 & \textit{otherwise} \end{cases}$$

The entropy function h associated with p is $h(A) = h(B) = h(C) = 1$, and $h(AB) = h(AC) = h(BC) = h(ABC) = 2$. Therefore

$$\begin{aligned} I_h(A;B) &= h(A) + h(B) - h(AB) = 1 + 1 - 2 = 0 \text{ and} \\ I_h(A;B|C) &= h(AC) + h(BC) - h(C) - h(ABC) = 2 + 2 - 1 - 2 = 1. \end{aligned}$$

Consequently, for the I-Measure μ^* associated with h it holds that

$$\begin{aligned} \mu^*(\mathrm{m}(A) \cap \mathrm{m}(B) \cap \mathrm{m}(C)) &= \mu^*(\mathrm{m}(A) \cap \mathrm{m}(B)) - \mu^*(\mathrm{m}(A) \cap \mathrm{m}(B) \cap \mathrm{m}^c(C)) \\ &\underbrace{=}_{\textit{Table } 4} I_h(A;B) - I_h(A;B|C) \\ &= 0 - 1 = -1. \end{aligned}$$

By Theorem 6.2, every polymatroid (and hence every entropic function) is associated with a unique signed measure μ^*. The following Theorem characterizes I-Measures that correspond to a specific subset of entropic functions.

Theorem 6.5. (Yeung, 2008) *If there is no constraint on $X_1, \dots, X_n$, then μ^* can take any set of nonnegative values on the nonempty atoms of $\mathcal{F}_n$. In other words, if there is no constraint on $X_1, \dots, X_n$, and $\mu^* : \mathcal{F}_n \to \mathbb{R}_{\geq 0}$, then the vector $h \in \mathbb{R}^{2^n}_{\geq 0}$ where $h(U) \overset{\text{def}}{=} \mu^*(m(U))$ for all subsets $U \subseteq \{X_1, \dots, X_n\}$ is entropic.*

We say that an I-Measure $\mu^* : \mathcal{F}_n \to \mathbb{R}$ is *positive* if $\mu^*(a) \geq 0$ for every $a \in \mathcal{F}_n$. Theorem 6.5 implies that every positive I-Measure μ^* corresponds to an entropic function $h \in \mathbb{R}^{2^n}_{\geq 0}$. We denote by Δ_n the set of n-dimensional entropic functions that have a positive I-Measure. Since every entropic function is a polymatroid, then $\Delta_n \subset \Gamma_n$. We refer to Δ_n as the set of n-dimensional *positive polymatroids.*

6.2 Exact Implication in the Set of Positive Polymatroids

Lemma 6.6. *The following holds:*

$$\Delta_n \models \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad \mathrm{m}(\Sigma) \supseteq \mathrm{m}(\tau).$$

Proof. Suppose that $\mathrm{m}(\tau) \not\subseteq \mathrm{m}(\Sigma)$, and let $b \in \mathcal{F}_n$ be an atom such that $b \in \mathrm{m}(\tau) \backslash \mathrm{m}(\Sigma)$. By Theorem 6.5, there exists a positive polymatroid in Δ_n with an I-Measure μ^* that takes the following non-negative values on its atoms: $\mu^*(b) = 1$, and $\mu^*(a) = 0$ for any atom $a \in \mathcal{F}_n$ where $a \neq b$. Since $b \notin \mathrm{m}(\Sigma)$, then $\mu^*(\Sigma) = 0$ while $\mu^*(\tau) = 1$. Hence, $\Delta_n \not\models \Sigma \Rightarrow \tau$.

Now, suppose that $\mathrm{m}(\Sigma) \supseteq \mathrm{m}(\tau)$. Then for any positive I-Measure $\mu^* : \mathcal{F}_n \to \mathbb{R}_{\geq 0}$, we have that $\mu^*(\mathrm{m}(\Sigma)) \geq \mu^*(\mathrm{m}(\tau))$. By Theorem 6.2, μ^* is the unique signed measure on $\mathcal{F}_n$ that is consistent with all of Shannon's information measures. Therefore, $h(\Sigma) \geq h(\tau)$. The result follows from the non-negativity of the Shannon information measures. □

An immediate consequence of Lemma 6.6 is that $\mathrm{m}(\Sigma) \supseteq \mathrm{m}(\tau)$ is a necessary condition for implication between polymatroids.

Colollary 6.7. *If* $\Gamma_n \models \Sigma \Rightarrow \tau$ *then* $\mathrm{m}(\Sigma) \supseteq \mathrm{m}(\tau)$.

Proof. If $\Gamma_n \models \Sigma \Rightarrow \tau$ then it must hold for any subset of polymatroids, and in particular, $\Delta_n \models \Sigma \Rightarrow \tau$. The result follows from Lemma 6.6. □

Lemma 6.8. *Let* $\Delta_n \models \Sigma \Rightarrow \tau$*, and let* $\sigma \in \Sigma$ *such that* $\mathrm{m}(\sigma) \cap \mathrm{m}(\tau) = \emptyset$*. Then* $\Delta_n \models \Sigma \backslash \{\sigma\} \Rightarrow \tau$*.*

Proof. Let $\Sigma' = \Sigma \backslash \{\sigma\}$, and suppose that $\Delta_n \not\models \Sigma' \Rightarrow \tau$. By Lemma 6.6, we have that $\mathrm{m}(\Sigma') \not\supseteq \mathrm{m}(\tau)$. In other words, there is an atom $a \in \mathcal{F}_n$ such that $a \in \mathrm{m}(\tau) \backslash \mathrm{m}(\Sigma')$. In particular, $a \notin \mathrm{m}(\sigma) \cup \mathrm{m}(\Sigma') = \mathrm{m}(\Sigma)$. Hence, $\mathrm{m}(\tau) \not\subseteq \mathrm{m}(\Sigma)$, and by Lemma 6.6 we get that $\Delta_n \not\models \Sigma \Rightarrow \tau$. □

Colollary 6.9. *Let* $\Delta_n \models \Sigma \Rightarrow \tau$ *where* $\tau = (A; B|C)$*, and let* $\sigma = (X; Y|Z) \in \Sigma$*. If* $A \subseteq Z$*,* $B \subseteq Z$*,* $X \subseteq C$*, or* $Y \subseteq C$*, then* $\Delta_n \models \Sigma \backslash \{\sigma\} \Rightarrow \tau$*.*

Proof. By definition, it holds that $\mathrm{m}(\tau) = \mathrm{m}(A) \cap \mathrm{m}(B) \cap \mathrm{m}^c(C) = (\cup_{a \in A} \mathrm{m}(a)) \cap (\cup_{b \in B} \mathrm{m}(b)) \cap (\cap_{c \in C} \mathrm{m}^c(c))$, and likewise $\mathrm{m}(\sigma) = (\cup_{x \in X} \mathrm{m}(x)) \cap (\cup_{y \in Y} \mathrm{m}(y)) \cap (\cap_{z \in Z} \mathrm{m}^c(z))$. If $A \subseteq Z$, then $\mathrm{m}(\sigma) \subseteq \mathrm{m}^c(Z) \subseteq \mathrm{m}^c(A)$, while $\mathrm{m}(\tau) \subseteq \mathrm{m}(A)$. Therefore, $\mathrm{m}(\tau) \cap \mathrm{m}(\sigma) = \emptyset$. Similarly, if $B \subseteq Z$, then $\mathrm{m}(\sigma) \subseteq \mathrm{m}^c(Z) \subseteq \mathrm{m}^c(B)$, while $\mathrm{m}(\tau) \subseteq \mathrm{m}(B)$, and hence $\mathrm{m}(\tau) \cap \mathrm{m}(\sigma) = \emptyset$. Similarly, it is shown that if $X \subseteq C$ or $Y \subseteq C$, then $\mathrm{m}(\tau) \cap \mathrm{m}(\sigma) = \emptyset$. The corollary then directly follows from Lemma 6.8. □

Lemma 6.10. *Let* $\Delta_n \models \Sigma \Rightarrow \tau = (A; B|C)$*. There exists a CI* $\sigma = (X; Y|Z) \in \Sigma$ *such that* $X \not\subseteq C$*,* $Y \not\subseteq C$*, and* $Z \subseteq C$*.*

Proof. Let $\Sigma' \overset{\text{def}}{=} \{(X;Y|Z) \in \Sigma \mid X \not\subseteq C \text{ and } Y \not\subseteq C\}$. By Corollary 6.9, if $\Delta_n \models \Sigma \Rightarrow \tau$, then $\Delta_n \models \Sigma' \Rightarrow \tau$. Suppose, by way of contradiction, that for every $\sigma = (X;Y|Z) \in \Sigma'$, it holds that $Z \not\subseteq C$. Consider the atom $t \overset{\text{def}}{=} \bigcap_{a \in \Omega \backslash C} \mathrm{m}(a) \cap \mathrm{m}^c(C)$. Clearly, $t \in \mathrm{m}(\tau)$. Now, take any $\sigma = (X;Y|Z) \in \Sigma'$. Since $Z \not\subseteq C$, then there exists a variable $a \in Z \backslash C$, and $\mathrm{m}(\sigma) \subseteq \mathrm{m}^c(Z) \subseteq \mathrm{m}^c(a)$. On the other hand, since $a \notin C$, then $t \in \mathrm{m}(a)$, and hence $t \notin \mathrm{m}(\sigma)$. Since this is the case for all $\sigma \in \Sigma'$, then $t \notin \mathrm{m}(\Sigma')$, and thus $\mathrm{m}(\tau) \not\subseteq \mathrm{m}(\Sigma')$. By Lemma 6.6, it holds that $\Delta_n \not\models \Sigma' \Rightarrow \tau$, and this brings us to a contradiction. □

6.3 Exact Implication in the Set of Polymatroids

The main technical result of this section is Lemma 6.12, that establishes an essential characterization of exact implication in the set of of polymatroids (and entropic functions). We start with a short technical lemma.

Lemma 6.11. *Let $\sigma = (X;Y|Z)$ and $\tau = (A;B|C)$ be CIs such that $A \subseteq X$, $B \subseteq Y$, $Z \subseteq C$, and $C \subseteq XYZ$. Then, $\Gamma_n \models h(\tau) \leq h(\sigma)$.*

Proof. By Lemma 2.1, we may assume that $C \cap AB = \emptyset$, and that $Z \cap XY = \emptyset$. Since $Z \subseteq C \subseteq XYZ$, then $C = C_X C_Y Z$, where $C_X \overset{\text{def}}{=} X \cap C$, and $C_Y \overset{\text{def}}{=} Y \cap C$. Also, denote by $X' \overset{\text{def}}{=} X \backslash (C_X \cup A)$, $Y' \overset{\text{def}}{=} Y \backslash (C_Y \cup B)$. So, we have that $I(X;Y|Z) = I(X'C_XA;Y'C_YB|Z)$. By the chain rule (see (3)), and submodularity (i.e., $I(\cdot;\cdot|\cdot) \geq 0$), we have that

$$\begin{aligned}
I(X'C_XA;Y'C_YB|Z) &= I(X'C_XA;C_Y|Z) + I(X'C_XA;Y'B|C_YZ) && \text{see (3)}\\
&\geq I(X'C_XA;Y'B|C_YZ)\\
&= I(C_XA;Y'B|C_YZ) + I(X';Y'B|C_YC_XZA) && \text{see (3)}\\
&\geq I(C_XA;Y'B|C_YZ)\\
&= I(C_X;Y'B|C_YZ) + I(A;Y'B|C_YC_XZ) && \text{see (3)}\\
&\geq I(A;Y'B|C_YC_XZ)\\
&= I(A;B|C_XC_YZ) + I(A;Y'|C_XC_YBZ) && \text{see (3)}\\
&\geq I(A;B|C_XC_YZ)\\
&= I(A;B|C).
\end{aligned}$$

The final step follows from the fact that $C = ZC_XC_Y$. Hence, we get that $I(A;B|C) \leq I(X;Y|Z)$ as required. □

Lemma 6.12. *Let $\tau = (A;B|C)$. If $\Gamma_n \models \Sigma \Rightarrow \tau$ then there exists a triple $\sigma = (X;Y|Z) \in \Sigma$ such that*

1. *$XYZ \supseteq ABC$, and*
2. *$ABC \cap X \neq \emptyset$ and $ABC \cap Y \neq \emptyset$.*

Proof. Let $\tau = (A;B|C)$, where $A = a_1 \dots a_m$, $B \backslash A = b_1 \dots b_\ell$, $C = c_1 \dots c_k$, and $U \overset{\text{def}}{=} \Omega \backslash ABC$. By Lemma 2.1, we assume that $AB \cap C = \emptyset$. Following (Geiger et al., 1991a), we construct the parity distribution $P(\Omega)$ as follows. We let all the RVs, except a_1, be independent binary RVs with probability $1/2$ for each of their two values, and let a_1 be determined from $ABC \backslash \{a_1\}$ as follows:

$$a_1 = \sum_{i=2}^{m} a_i + \sum_{i=1}^{\ell} b_i + \sum_{i=1}^{k} c_i \pmod 2. \tag{28}$$

Let $D \subseteq \Omega$ and $\boldsymbol{d} \in \mathcal{D}(D)$. We denote by $D_{ABC} \overset{\mathsf{def}}{=} D \cap ABC$, and by $\boldsymbol{d}_{ABC}$ the assignment $\boldsymbol{d}$ restricted to the RVs D_{ABC}. We show that if $D_{ABC} \subsetneq ABC$ then the RVs in D are mutually independent. By the definition of P we have that

$$P(D = \boldsymbol{d}) = \left(\frac{1}{2}\right)^{|D \cap U|} P(D_{ABC} = \boldsymbol{d}_{ABC}).$$

There are two cases with respect to D. If $a_1 \notin D$ then, by definition, $P(D_{ABC} = \boldsymbol{d}_{ABC}) = \left(\frac{1}{2}\right)^{|D_{ABC}|}$, and overall we get that $P(D = \boldsymbol{d}) = \left(\frac{1}{2}\right)^{|D|}$, proving that the RVs in D are mutually independent. If $a_1 \in D$, then since $D_{ABC} \subsetneq ABC$ it holds that $P(a_1 | D_{ABC} \backslash \{a_1\}) {=} P(a_1)$. To see this, observe that

$$P(a_1 = 1 | D_{ABC} \backslash \{a_1\}) = \begin{cases} \frac{1}{2} & \text{if } \sum_{y \in D_{ABC} \backslash \{a_1\}} y \pmod 2{=}0 \\ \frac{1}{2} & \text{if } \sum_{y \in D_{ABC} \backslash \{a_1\}} y \pmod 2{=}1 \end{cases}$$

because if, w.l.o.g., $\sum_{y \in D_{ABC} \backslash \{a_1\}} y \pmod 2 = 0$, then $a_1 = 1$ implies that $\sum_{y \in ABC \backslash D} y \pmod 2 = 1$, and this is the case for precisely half of the assignments $ABC \backslash D {\rightarrow} \{0,1\}^{|ABC \backslash D|}$. Hence, for any $D \subseteq \Omega$ such that $D \cap \mathbf{var}(\tau) \subsetneq ABC$, it holds that $P(D = \boldsymbol{d}) = \prod_{y \in D} P(y = \boldsymbol{d}_y) = \left(\frac{1}{2}\right)^{|D|}$, and therefore the RVs in D are mutually independent.

By definition of entropy (see (4)) we have that $H(X_i) = 1$ for every binary RV in Ω. Since the RVs in D are mutually independent, then $H(D) = \sum_{y \in D} H(y) = |D|$ (see Section 2.2). Furthermore, for any $(X;Y|Z) \in \Sigma$ such that $XYZ \not\supseteq ABC$ we have that

$$\begin{aligned} I(X;Y|Z) &= H(XZ) + H(YZ) - H(Z) - H(XYZ) \\ &= |XZ| + |YZ| - |Z| - |XYZ| \\ &= |X| + |Y| + |Z| - |XYZ| \\ &= 0. \end{aligned}$$

On the other hand, letting $A' {\overset{\mathsf{def}}{=}} A \backslash \{a_1\}$, then by chain rule for entropies (see (5)), and noting that, by (28), $ABC \backslash a_1 \rightarrow a_1$, then

$$\begin{aligned} H(\mathbf{var}(\tau)) = H(ABC) &= H(a_1 A' BC) \\ &= H(a_1 | A' BC) + H(A' BC) \\ &= 0 + |ABC| - 1 = |ABC| - 1. \end{aligned}$$

Therefore,

$$\begin{aligned} I(A;B|C) &= H(AC) + H(BC) - H(C) - H(ABC) \\ &= |AC| + |BC| - |C| - (|ABC| - 1) \qquad (29) \\ &= 1. \end{aligned}$$

In other words, the parity distribution P of (28) has an entropic function $h_P \in \Gamma_n$, such that $h_P(\sigma) = 0$ for all $\sigma \in \Sigma$ where $\mathbf{var}(\sigma) \not\supseteq ABC$, while $h_P(\tau) = 1$. Hence, if $\Gamma_n \models \Sigma \Rightarrow \tau$, then there must be a triple $\sigma = (X;Y|Z) \in \Sigma$ such that $XYZ \supseteq ABC$.

Now, suppose that $ABC \subseteq XYZ$ and that $ABC \cap Y = \emptyset$. In other words, $ABC \subseteq XZ$. We denote $X_{ABC} \overset{\text{def}}{=} X \cap ABC$ and $Z_{ABC} \overset{\text{def}}{=} Z \cap ABC$. Therefore, we can write $I(X;Y|Z)$ as $I(X_{ABC}X';Y|Z_{ABC}Z')$ where $X' \overset{\text{def}}{=} X \backslash X_{ABC}$ and $Z' \overset{\text{def}}{=} Z \backslash Z_{ABC}$. Since $X_{ABC}Z_{ABC} = ABC$, and due to the properties of the parity distribution, we get

$$\begin{aligned}
&I(X_{ABC}X';Y|Z_{ABC}Z') = \\
&H(X_{ABC}Z_{ABC}X'Z') + H(YZ_{ABC}Z') - H(Z_{ABC}Z') - H(X_{ABC}Z_{ABC}X'Z'Y) = \\
&H(ABCX'Z') + H(YZ_{ABC}Z') - H(Z_{ABC}Z') - H(ABCX'Z'Y) = \\
&H(ABC) + H(X'Z') + H(Y) + H(Z_{ABC}Z') - H(Z_{ABC}Z') - H(ABC) - H(X'Z'Y) = 0.
\end{aligned}$$

Symmetrically, if $ABC \subseteq YZ$, then $I(X;Y|Z) = 0$.

Overall, we showed that for all triples $(X;Y|Z) \in \Sigma$ that do not meet the conditions of the lemma, it holds that $I_{h_P}(X;Y|Z) = 0$, while $I_{h_P}(A;B|C) = 1$ (see (29)) where h_P is the entropic function associated with the parity function P in (28). Therefore, there must be a triple $\sigma \in \Sigma$ that meets the conditions of the lemma. Otherwise, we arrive at a contradiction to the assumption that $\Gamma_n \models \Sigma \Rightarrow \tau$. □

7. Approximate Implication For Saturated CIs

In this section we prove Theorem 4.3. In fact, we prove the following stronger statement.

Theorem 7.1. *Let $\tau = (A;B|C)$, and let Σ be a set of saturated CIs and conditionals. Then*

$$\Delta_n \models \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad \Gamma_n \models h(\tau) \leq \min\{|A|,|B|\} \cdot h(\Sigma). \tag{30}$$

Since $\Delta_n \subset \Gamma_n$, then if $\Gamma_n \models \Sigma \Rightarrow \tau$, then clearly $\Delta_n \models \Sigma \Rightarrow \tau$. Theorem 7.1 establishes that if Σ is a set of saturated CIs and conditionals, then $\Delta_n \models \Sigma \Rightarrow \tau$ implies the (stronger!) statement that $\Gamma_n \models \Sigma \Rightarrow \tau$. Previously, it was shown that if Σ is a set of saturated CIs, and $\Delta_n \models \Sigma \Rightarrow \tau$, then $\Gamma_n \models h(\tau) \leq |A|{\cdot}|B|{\cdot}h(\Sigma)$ (Kenig & Suciu, 2022). In particular, $h(\tau) \leq {}^{n^2}\!/\!_4{\cdot}h(\Sigma)$. Since $|A \cup B| \leq n$, then $\min\{|A|,|B|\} \leq {}^{n}\!/\!_2$, leading to the significantly tighter bound of $h(\tau) \leq {}^{n}\!/\!_2 \cdot h(\Sigma)$.

Before proceeding, we show that w.l.o.g. we can assume that Σ consists of only saturated CIs. If Σ contains a non-saturated term, then it must be a conditional $X \rightarrow Y$. By Lemma 2.2, it holds that

$$h(X \rightarrow Y) = h(Y|X) = I_h(Y;\Omega\backslash(XY)|X) + I_h(Y;Y|\Omega\backslash Y). \tag{31}$$

We create a new set of CIs by replacing every conditional $X \rightarrow Y \in \Sigma$ with the two saturated CIs $(Y;\Omega\backslash(XY)|X)$ and $(Y;Y|\Omega\backslash Y)$. Denoting by Σ' the new set of CIs, it follows from (31) that $h(\Sigma) = h(\Sigma')$. Therefore, we assume w.l.o.g. that Σ contains only saturated CIs. The proof of Theorem 7.1 relies on the following Lemma, proved in Section 7.1.

Lemma 7.2. *Let $\tau = (a;B|C)$ where $a \in \Omega$ is a singleton, and $B, C \subseteq \Omega$, and let Σ be a set of saturated CIs. Then*

$$\Delta_n \models \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad \Gamma_n \models h(\tau) \leq h(\Sigma). \tag{32}$$

Proof of Theorem 7.1

If $\Gamma_n \models h(\tau) \leq h(\Sigma)$, then whenever $h(\Sigma) = 0$, it holds that $h(\tau) \leq h(\Sigma) = 0$. By the non-negativity of the Shannon information measures, we have that $\Gamma_n \models 0 \leq h(\tau)$. Therefore, if $h(\Sigma) = 0$, then $h(\tau) = 0$; or $\Gamma_n \models \Sigma \Rightarrow \tau$. Since $\Delta_n \subseteq \Gamma_n$, then $\Delta_n \models \Sigma \Rightarrow \tau$.

Consider the consequent $\tau = (A; B|C)$. By Lemma 2.1, we may assume that $AB \cap C = \emptyset$. We do not assume that $A \cap B = \emptyset$. Suppose, without loss of generality, that $|A| \leq |B|$, and that $A = a_1 \dots a_K$. By the chain rule of mutual information (see (3)), we have that

$$h(\tau) = I_h(a_1 \dots a_K; B|C) = I_h(a_1; B|C) + I_h(a_2; B|a_1C) + \cdots + I_h(a_K; B|a_1 \dots a_{K-1}C).$$

By Lemma 7.2, we have that $I_h(a_i; B|a_1 \dots a_{i-1}C) \leq h(\Sigma)$ for every $i \in \{1, \dots, K\}$. Hence, we get that $h(\tau) \leq |A| \cdot h(\Sigma)$. We remark that Lemma 7.2 holds also if $a_i \in B$. Hence, the theorem holds also if $A \cap B \neq \emptyset$. In particular, the theorem holds if $\tau = (A; A|C)$, in which case $\tau = (C \to A)$.

7.1 Proof of Lemma 7.2

If $\Gamma_n \models h(\tau) \leq h(\Sigma)$, then whenever $h(\Sigma) = 0$, it holds that $h(\tau) \leq h(\Sigma) = 0$. By the non-negativity of the Shannon information measures, we have $\Gamma_n \models 0 \leq h(\tau)$. Therefore, if $h(\Sigma) = 0$, then $h(\tau) = 0$; or $\Gamma_n \models \Sigma \Rightarrow \tau$. Since $\Delta_n \subseteq \Gamma_n$, then $\Delta_n \models \Sigma \Rightarrow \tau$.

Now, suppose that $\Delta_n \models \Sigma \Rightarrow \tau$, and $|\Omega| = n$. We prove the claim by reverse induction on $|C|$.

Base. There are two base cases, where $|C| = n - 1$, and $|C| = n - 2$. If $|C| = n - 1$, then $\tau = (a; a|C)$ where $a \in \Omega$. That is, $\tau = (C \to a)$ is a conditional, and $aC = \Omega$. By Lemma 6.10, there exists a CI $\sigma = (X; Y|Z) \in \Sigma$ such that $Z \subseteq C$, $X \not\subseteq C$, and $Y \not\subseteq C$. Since $X \not\subseteq C$, then $X \cap (\Omega \backslash C) = X \cap \{a\} \neq \emptyset$, and hence $a \in X$. Likewise, we get that $a \in Y$. We denote by $X_C \overset{\mathsf{def}}{=} X \cap C$, and $Y_C \overset{\mathsf{def}}{=} Y \cap C$. By Lemma 6.11, we have that

$$h(\sigma) = I_h(aX_C; aY_C|Z) \underbrace{\geq}_{\text{Lemma 6.11}} I_h(a; a|ZX_CY_C) \underbrace{=}_{Z \subseteq C \subseteq XYZ} I_h(a; a|C) = h(\tau).$$

Since $\sigma \in \Sigma$, we get that $\Gamma_n \models h(\tau) \leq h(\Sigma)$.

Now, assume that $|C| = n - 2$. This means that $\tau = (a; b|C)$ where $a, b \in \Omega$ are singletons, and that $|abC| = n$. Therefore, $abC = \Omega$. By Lemma 6.10, it holds that there exists a CI $\sigma = (X; Y|Z) \in \Sigma$ such that $Z \subseteq C$, $X \not\subseteq C$, and $Y \not\subseteq C$. Since $X \not\subseteq C$, then $X \cap (\Omega \backslash C) \neq \emptyset$, and hence $X \cap ab \neq \emptyset$. Similarly, since $Y \not\subseteq C$, then $Y \cap ab \neq \emptyset$. Since $Z \subseteq C$, then $ab \cap Z = \emptyset$, and since Σ is saturated, then $ab \subseteq XYZ$, and $ab \subseteq XY$. So, we may assume w.l.o.g. that $a \in X$ and $b \in Y$. We denote by $X_C \overset{\mathsf{def}}{=} X \cap C$, and $Y_C \overset{\mathsf{def}}{=} Y \cap C$. By Lemma 6.11, we have that

$$h(\sigma) = I_h(aX_C; bY_C|Z) \underbrace{\geq}_{\text{Lemma 6.11}} I_h(a; b|ZX_CY_C) \underbrace{=}_{Z \subseteq C \subseteq XYZ} I_h(a; b|C) = h(\tau).$$

Since $\sigma \in \Sigma$, we get that $\Gamma_n \models h(\tau) \leq h(\Sigma)$.

Step. We now assume that the claim holds for all $C \subseteq \Omega$, where $|C| \geq k+1$, and we prove the claim for the case where $|C| = k$. By Lemma 6.10, it holds that there exists a CI $\sigma = (X;Y|Z) \in \Sigma$ such that $Z \subseteq C$, $X \not\subseteq C$, and $Y \not\subseteq C$. Since $XYZ \supseteq aBC$, and $Z \subseteq C$, then $XY \supseteq aB$. So, we may assume w.l.o.g. that $a \in X$. As before, we denote by $X_C \stackrel{\text{def}}{=} X \cap C$, $X_B \stackrel{\text{def}}{=} X \cap B$, $X' \stackrel{\text{def}}{=} X \setminus aX_BX_C$, $Y_C \stackrel{\text{def}}{=} Y \cap C$, $Y_B \stackrel{\text{def}}{=} Y \cap B$, and $Y' \stackrel{\text{def}}{=} Y \setminus Y_CY_B$. We write $h(\sigma) = I_h(aX_BX_CX';Y_BY_CY'|Z)$ as follows.

$$I_h(aX_BX_CX';Y_BY_CY'|Z) = I_h\underbrace{(aX_BX_CX';Y_C|Z)}_{\stackrel{\text{def}}{=}\alpha_1} + I_h(aX_BX_CX';Y_BY'|Y_CZ) \tag{33}$$

$$\underbrace{=}_{(3)} h(\alpha_1) + I_h\underbrace{(aX_BX_CX';Y_B|Y_CZ)}_{\stackrel{\text{def}}{=}\sigma_1} + I_h\underbrace{(aX_BX_CX';Y'|Y_BY_CZ)}_{\stackrel{\text{def}}{=}\sigma_2}. \tag{34}$$

Now, we consider two options for σ: (1) $Y_B \neq \emptyset$, and (2) $Y_B = \emptyset$.

We first consider the case where $Y_B \neq \emptyset$. In that case, we express $h(\tau)$ as

$$h(\tau) = I_h(a;B|C) = I_h(a;X_BY_B|C) \underbrace{=}_{(3)} I_h\underbrace{(a;Y_B|C)}_{\tau_1} + I_h\underbrace{(a;X_B|Y_BC)}_{\tau_2}. \tag{35}$$

Since $\mathbf{var}(\sigma_1) \supseteq \mathbf{var}(\tau_1)$, and $Y_CZ \subseteq C$, then by Lemma 6.11, it holds that $h(\tau_1) \leq h(\sigma_1)$. We define $\Sigma_2 \stackrel{\text{def}}{=} (\Sigma \backslash \{\sigma\}) \cup \{\sigma_2\}$. Since $\mathbf{var}(\sigma_2) = \mathbf{var}(\sigma) = \Omega$, then σ_2 is saturated, and hence Σ_2 is a set of saturated CIs. We claim that $\Delta_n \models \Sigma_2 \Rightarrow \tau_2$. This completes the proof for the case where $Y_B \neq \emptyset$, because $|Y_BC| > |C| = k$. By the induction hypothesis $\Gamma_n \models h(\tau_2) \leq h(\Sigma_2)$. Therefore,

$$h(\tau) = h(\tau_1) + h(\tau_2) \leq h(\sigma_1) + h(\Sigma_2) \leq h(\alpha_1) + h(\sigma_1) + h(\Sigma_2) = h(\Sigma).$$

So, we show that $\Delta_n \models \Sigma_2 \Rightarrow \tau_2$. Since $\Delta_n \models \Sigma \Rightarrow \tau$, then by (35), it holds that $\Delta_n \models \Sigma \Rightarrow \tau_2$. Observe that $\Sigma = \Sigma_2 \cup \{\sigma_1, \alpha_1\}$. Since Y_B belongs to the conditioned part in τ_2 (i.e., $\mathrm{m}(\tau_2) \subseteq \mathrm{m}^c(Y_B)$), and since $Y_C \subseteq C$, then by Corollary 6.9, it holds that $\Delta_n \models \Sigma_2 \Rightarrow \tau_2$, as required.

We now consider the case where $Y_B = \emptyset$. Since $Z \subseteq C$, and $XYZ \supseteq aBC$, this means that $B \subseteq X$. In this case, by repeated application of the chain rule for mutual information (see (3)), we can express $h(\sigma)$ as

$$\begin{aligned} h(\sigma) = I_h(aBX_CX';Y_CY'|Z) &= I_h\underbrace{(aBX_CX';Y_C|Z)}_{\stackrel{\text{def}}{=}\alpha_1} + I_h(aBX_CX';Y'|Y_CZ) \\ &= h(\alpha_1) + I_h\underbrace{(X_C;Y'|Y_CZ)}_{\stackrel{\text{def}}{=}\alpha_2} + I_h(aBX';Y'|X_CY_CZ) \\ &= h(\alpha_1) + h(\alpha_2) + I_h\underbrace{(a;Y'|C)}_{\stackrel{\text{def}}{=}\sigma_1} + I_h\underbrace{(B;Y'|aC)}_{\stackrel{\text{def}}{=}\sigma_2} + I_h\underbrace{(X';Y'|aBC)}_{\stackrel{\text{def}}{=}\alpha_3}. \end{aligned} \tag{36}$$

where $\sigma_1 \stackrel{\text{def}}{=} (a;Y'|C)$, $\sigma_2 \stackrel{\text{def}}{=} (B;Y'|aC)$, and $\alpha_3 \stackrel{\text{def}}{=} (X';Y'|aBC)$. Since $Y \not\subseteq C$, then $Y' \neq \emptyset$. In particular, we have that $\Delta_n \models \Sigma \Rightarrow \sigma_2$, and $\Delta_n \models \Sigma \Rightarrow \tau$. By the chain rule (see (3)), we have that

$$h(\tau)+h(\sigma_2)=I_h(a;B|C)+I_h(B;Y'|aC)=I_h\underbrace{(aY';B|C)}_{\tau'},$$

where we denote $\tau' \overset{\text{def}}{=} (aY';B|C)$. Hence, it holds that $\Delta_n \models \Sigma \Rightarrow \tau'$. We express $h(\tau')$ as

$$h(\tau')=I_h(aY';B|C)=I_h\underbrace{(B;Y'|C)}_{\tau_1}+I_h\underbrace{(a;B|Y'C)}_{\tau_2},$$

where we denote $\tau_1 \overset{\text{def}}{=} (B;Y'|C)$, and $\tau_2 \overset{\text{def}}{=} (a;B|Y'C)$. Since $\Delta_n \models \Sigma \Rightarrow \tau'$, then $\Delta_n \models \Sigma \Rightarrow \tau_2$. Since $Y_CY' \subseteq Y'C$, then by Corollary 6.9, it holds that $\Delta_n \models \Sigma\backslash\{\sigma\} \Rightarrow \tau_2$. Letting $\Sigma_2 \overset{\text{def}}{=} \Sigma\backslash\{\sigma\}$, we get that $\Delta_n \models \Sigma_2 \Rightarrow \tau_2$.

Since $\Sigma_2 \subseteq \Sigma$, then all CIs in Σ_2 are saturated. Since $\tau_2=(a;B|Y'C)$ where $Y' \neq \emptyset$, then $|Y'C| > |C| = k$. Hence, by the induction hypothesis, we have that $\Gamma_n \models h(\tau_2) \leq h(\Sigma_2)$. Also, by Lemma 6.11, we have that $h(\tau_1)=I_h(B;Y'|C) \leq h(\sigma)$. Overall, we get that

$$h(\tau) \leq h(\tau') = h(\tau_1)+h(\tau_2) \leq h(\sigma)+h(\Sigma_2)=h(\Sigma).$$

This completes the proof.

8. Approximate Implication for Recursive CIs

In this section, w e prove Theorem 4.4. Let Σ be the recursive basis (see (10)) over the variable set Ω=$\{X_1,\dots,X_n\}$, and let G be the DAG generated by the recursive set Σ. By definition, $X_1,\dots,X_n$ correspond to a topological order of the vertices in G. Let $\sigma=(A;B|C)$ where $A,B,C \subseteq \Omega$ are pairwise disjoint. We denote by $I_G(A;B|C)$ or $I_G(\sigma)$ the fact that A is d-separated from B given C in G (see Section 3.2). Theorem 3.7 establishes that d-separation is sound and complete for inferring CIs from the recursive basis. This means that $I_G(A;B|C)$ if and only if $\Gamma_n \models \Sigma \Rightarrow (A;B|C)$. In our proof, we will make use of the following.

Lemma 8.1. (Pearl, 1989) *Let $X,Y,Z \subseteq \Omega$ be pairwise disjoint, and let $\gamma \in \Omega\backslash(XYZ)$.*

If $I_G(X;Y|Z)$ and $I_G(X;Y|Z\gamma)$, then $I_G(X;\gamma|Z)$ or $I_G(Y;\gamma|Z)$.

We prove Theorem 4.4 by induction on the highest RV-index mentioned in any triple of Σ. The claim vacuously holds for $n=1$ (since no conditional independence statements are implied), so we assume correctness when the highest RV-index mentioned in Σ is $\leq n-1$, and prove for n.

The recursive basis contains n CIs, $\Sigma=\{\sigma_1,\dots,\sigma_n\}$, where $\sigma_i=(X_i;Y_i|Z_i)$, where $Y_iZ_i=\{X_1,\dots,X_{i-1}\}$. In particular, only $\sigma_n=(X_n;Y_n|Z_n)$ mentions the RV X_n, and it is saturated (i.e., $X_nY_nZ_n=\Omega$). By Lemma 2.1, we may assume that in the consequent $\tau=(A;B|C)$, it holds that $AB \cap C = \emptyset$. We claim that if $\Gamma_n \models \Sigma \Rightarrow \tau$, then $A \cap B = \emptyset$. Suppose otherwise, and let $d \in A \cap B$. By Lemma 6.11, we have that $h(\tau) \geq I(d;d|C)$. Therefore, if $\Gamma_n \models \Sigma \Rightarrow \tau$, then $\Gamma_n \models \Sigma \Rightarrow (d;d|C)$. By Lemma 6.12, we have that there exists a triple $(X_i;Y_i|Z_i) \in \Sigma$, such that $X_i \cap \{d\} \neq \emptyset$ and $Y_i \cap \{d\} \neq \emptyset$. But this means that $d = X_i$ and $d \in Y_i$. By definition of the recursive basis, we have that $X_i \notin Y_i$, which brings us to a contradiction. Therefore, we may assume that $A \cap B = \emptyset$. We denote by $\Sigma' \overset{\mathsf{def}}{=} \Sigma \backslash \{\sigma_n\}$, and note that $X_n \notin \mathbf{var}(\Sigma')$. The induction hypothesis states that

$$\Gamma_n \models \Sigma' \Rightarrow \tau' \qquad \text{if and only if} \qquad \Gamma_n \models h(\Sigma') \geq h(\tau'). \tag{37}$$

Now, we consider $\tau = (A; B|C)$. We divide into three cases, and treat each one separately:

1. $X_n \notin ABC$.
2. $X_n \in A$ (or, symmetrically, $X_n \in B$).
3. $X_n \in C$.

Case 1: $X_n \notin ABC$. By Theorem 3.7, it holds that the d-separation criterion is complete with respect to the recursive basis. Therefore, if $\Gamma_n \models \Sigma \Rightarrow (A; B|C)$, then $I_G(A; B|C)$. That is, A and B are d-separated given C in G. Let G' be the graph that results from G by removing X_n an all edges adjacent to X_n. We claim that $I_{G'}(A; B|C)$. If not, then there is an active trail $P = (a, v_1, \dots, v_k, b)$ in G' between a vertex $a \in A$ and $b \in B$, given C. Since all vertices and edges in P are included in G, and since the addition of vertices and edges cannot block a trail (see Section 3.2), then P is an active trail given C between a and b in G. By the completeness of d-separation, this implies that $\Gamma_n \not\models \Sigma \Rightarrow \tau$, bringing us to a contradiction. Therefore, it holds that $I_{G'}(A; B|C)$. Since the recursive basis associated with G' is $\Sigma' \overset{\mathsf{def}}{=} \Sigma \backslash \{\sigma_n\}$, and since d-separation is sound, we get that $\Gamma_n \models \Sigma' \Rightarrow \tau$. Since $X_n \notin \mathbf{var}(\Sigma')$, then by the induction hypothesis we get that $\Gamma_n \models h(\Sigma') \geq h(\tau)$.

Case 2: $\tau = (AX_n; B|C)$. Recall that $\sigma_n = (X_n; Y_n|Z_n)$. We claim that $B \subseteq Y_n$. Suppose otherwise, and let $b \in B \backslash Y_n$. Since σ_n is saturated, and $b \notin \{X_n\} \cup Y_n$, then $b \in Z_n$. Consider the atom $t \overset{\mathsf{def}}{=} \mathrm{m}(X_n) \cap \mathrm{m}(b) \cap \mathrm{m}^c(\Omega \backslash bX_n)$. Clearly, $t \in \mathrm{m}(\tau)$. On the other hand, $t \notin \mathrm{m}(\sigma_n)$ because $Y_n \subseteq \Omega \backslash bX_n$. For every $\sigma = (X_i; Y_i|Z_i) \in \Sigma'$ either $X_i \in \Omega \backslash X_n b$, or $Y_i \subseteq \Omega \backslash X_n b$, and hence $t \notin \mathrm{m}(\sigma)$. Consequently, $\mathrm{m}(\tau) \not\subseteq \mathrm{m}(\Sigma') \cup \mathrm{m}(\sigma_n) = \mathrm{m}(\Sigma)$. From Corollary 6.7, we get that $\Gamma_n \not\models \Sigma \Rightarrow \tau$, which brings us to a contradiction. Therefore, $B \subseteq Y_n$.

Since σ_n is saturated, then $ABC \subseteq X_n Y_n Z_n$. We denote by $A_Y \overset{\mathsf{def}}{=} A \cap Y_n$, $A_Z \overset{\mathsf{def}}{=} A \cap Z_n$, and $Y' \overset{\mathsf{def}}{=} Y_n \backslash ABC$. Similarly, we define C_Y, C_Z, and Z'. Since $B \subseteq Y_n$, we can express $h(\sigma_n)$ as

$$\begin{aligned} h(\sigma_n) = I_h(X_n; BA_Y C_Y Y'|A_Z C_Z Z') &\geq I_h(X_n; BY'|A_Y C_Y A_Z C_Z Z') \\ &= I_h(X_n; BY'|ACZ') \\ &\geq I_h(X_n; B|ACZ'). \end{aligned}$$

By the chain rule (see (3)), we express $h(\tau)$ as

$$h(\tau) = I_h(AX_n; B|C) = I_h \underbrace{(A; B|C)}_{\tau_1} + I_h \underbrace{(X_n; B|AC)}_{\tau_2}. \tag{38}$$

Since $\Gamma_n \models \Sigma \Rightarrow \tau$ then $\Gamma_n \models \Sigma \Rightarrow \tau_1$, and $\Gamma_n \models \Sigma \Rightarrow \tau_2$.

We claim that $\Gamma_n \models \Sigma \Rightarrow (B; Z'|AC)$. Suppose, by way of contradiction that $\Gamma_n \not\models \Sigma \Rightarrow (B; Z'|AC)$. By the soundness of the d-separation algorithm (Theorem 3.7), there is an active trail P from $b \in B$ to $z \in Z'$ given AC in G. By construction, G contains an edge $(z \to X_n)$ for every $z \in Z_n$. Since $Z' \subseteq Z_n$, then G contains the edge $(z \to X_n)$, where $z \in Z'$. Therefore, the trail P can be augmented with the edge $(z \to X_n)$ to form an active trail from b to X_n (given AC). Since the d-separation algorithm is complete, this means that $\Gamma_n \not\models \Sigma \Rightarrow (X_n; B|AC)$. But then, $\Gamma_n \not\models \Sigma \Rightarrow \tau$, which brings us to a contradiction. Therefore, $\Gamma_n \models \Sigma \Rightarrow (B; Z'|AC)$. Therefore, we have that $\Gamma_n \models \Sigma \Rightarrow (B; Z'|AC), (A; B|C)$. By the chain rule, we get that $\Gamma_n \models \Sigma \Rightarrow (AZ'; B|C)$. Since $X_n \notin ABCZ'$, then as in Case 1, we have that $\Gamma_n \models \Sigma' \Rightarrow (AZ'; B|C)$, and by the induction hypothesis that $\Gamma_n \models h(\Sigma') \geq I_h(AZ'; B|C)$.

Since $\Gamma_n \models \Sigma' \Rightarrow (AZ'; B|C)$, and $\Gamma_n \models h(\sigma_n) \geq I_h(X_n; B|ACZ')$, we get that

$$\begin{aligned} h(\tau) = I_h(AX_n; B|C) \leq I(AX_nZ'; B|C) &= I_h(AZ'; B|C) + I_h(X_n; B|ACZ') \\ &\leq h(\Sigma') + h(\sigma_n) \\ &= h(\Sigma). \end{aligned}$$

Case 3: $\tau = (A; B|CX_n)$. We claim that $\Gamma_n \models \Sigma \Rightarrow (A; B|C)$. Suppose, by way of contradiction, that $\Gamma_n \not\models \Sigma \Rightarrow (A; B|C)$. By the soundness of the d-separation algorithm, it holds that A and B are not d-separated, given C, in the DAG G. By construction, X_n is a sink vertex in G (i.e., it has only incoming edges). Consequently, this means that A and B are not d-separated given CX_n. But then, by the completeness of the d-separation algorithm it holds that $\Gamma_n \not\models \Sigma \Rightarrow (A; B|CX_n)$, which brings us to a contradiction. Therefore, it holds that $\Gamma_n \models \Sigma \Rightarrow (A; B|C), (A; B|CX_n)$. By Lemma 8.1, this means that $\Gamma_n \models \Sigma \Rightarrow (A; X_n|C)$, or $\Gamma_n \models \Sigma \Rightarrow (B; X_n|C)$. We divide into cases accordingly. If $\Gamma_n \models \Sigma \Rightarrow (A; X_n|C)$, then by the chain rule, we have that

$$\Gamma_n \models \Sigma \Rightarrow (A; X_n|C), (A; B|CX_n) \Longrightarrow \Gamma_n \models \Sigma \Rightarrow (A; BX_n|C). \tag{39}$$

If $\Gamma_n \models \Sigma \Rightarrow (B; X_n|C)$, then by the chain rule, we have that

$$\Gamma_n \models \Sigma \Rightarrow (B; X_n|C), (A; B|CX_n) \Longrightarrow \Gamma_n \models \Sigma \Rightarrow (AX_n; B|C). \tag{40}$$

Both cases (39) and (40) bring us back to case 2. Hence, we have that

$$\begin{aligned} &h(\Sigma) \underbrace{\geq}_{(39)} I_h(A; BX_n|C) \underbrace{\geq}_{\text{Lemma 6.11}} I_h(A; B|CX_n) && \text{if (39) holds, and} \\ &h(\Sigma) \underbrace{\geq}_{(40)} I_h(AX_n; B|C) \underbrace{\geq}_{\text{Lemma 6.11}} I_h(A; B|CX_n) && \text{if (40) holds.} \end{aligned}$$

So, in both cases we get that $h(\Sigma) \geq h(\tau)$ as required. This completes the proof.

8.1 Tightness of Bound

Consider a probability distribution P over $\Omega = \{X_1, \dots, X_n\}$, such that the following recursive set of CIs holds in P:

$$\Sigma = \{(X_1; X_i|X_2 \dots X_{i-1}) : i \in \{2, \dots, n\}\}. \tag{41}$$

Let $\tau = (X_1; X_2X_3 \dots X_n)$. It is not hard to see that by the chain rule

$$I(X_1; X_2X_3 \dots X_n) = \sum_{i=2}^{n} I(X_1; X_i | X_2 \dots X_{i-1}) = h(\Sigma). \tag{42}$$

Hence, $\Gamma_n \models \Sigma \Rightarrow \tau$, and the bound of (42) is tight.

9. Approximate Implication for Marginal CIs

In this section, we prove Theorem 4.5. Let Σ be a set of marginal mutual information terms, and let $\tau = (A; B|C)$ such that $\Gamma_n \models \Sigma \Rightarrow \tau$. By the chain rule (see (3)), we can write $\tau = (a_1 \dots a_K; B|C)$ as

$$h(\tau) = I_h(a_1 \dots a_K; B|C) = I_h(a_1; B|C) + \dots + I_h(a_K; B|Ca_1 \dots a_{K-1}). \tag{43}$$

We show, in Theorem 9.4, that if $\Gamma_n \models \Sigma \Rightarrow (a_i; B|Ca_1 \dots a_{i-1})$, then $\Gamma_n \models h(\Sigma) \geq I_h(a_i; B|Ca_1 \dots a_{i-1})$, and thus $\Gamma_n \models \min\{|A|, |B|\} \cdot h(\Sigma) \geq h(\tau)$, as required.

Let Σ be a set of marginal CIs defined over variables Ω, and let $U \subseteq \Omega$. We denote by $\Sigma(U)$ the set of CIs projected onto the random variables U. Formally,

$$\Sigma(U) \overset{\mathsf{def}}{=} \{(X'; Y') : (X, Y) \in \Sigma, X' = X \cap U, Y' = Y \cap U, X' \supset \emptyset, Y' \supset \emptyset\}. \tag{44}$$

Example 9.1. *Suppose that* $\Sigma = \{(abc; e), (def; ac)\}$, *then* $\Sigma(eac) = \{(ac; e)\}$, *while* $\Sigma(def) = \emptyset$.

Lemma 9.2. *Let* Σ *be a set of marginal mutual information terms, and let* $\tau = (a; b|C)$ *be an elemental mutual information term where* $a, b \in \Omega$, *and* $C \subseteq \Omega$. *The following holds:*

$$\Gamma_n \models \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad \Gamma_n \models \Sigma(\mathbf{var}(\tau)) \Rightarrow \tau. \tag{45}$$

Proof. Since Σ is a set of marginal CIs, then by Lemma 6.11, it holds that $h(\Sigma) \geq h(\Sigma(\mathbf{var}(\tau)))$. Therefore, if $\Gamma_n \models \Sigma(\mathbf{var}(\tau)) \Rightarrow \tau$, then clearly $\Gamma_n \models \Sigma \Rightarrow \tau$.

We prove the other direction by induction on $|C|$. When $|C| = 0$, then $\tau = (a; b)$. By Lemma 6.12, it holds that there exists a CI $\sigma = (X; Y)$ such that (1) $XY \supseteq ab$, and (2) $ab \cap X \neq \emptyset$ and $ab \cap Y \neq \emptyset$. In other words, $\sigma = (aX'; bY')$, where $X' \overset{\mathsf{def}}{=} X \backslash \{a\}$ and $Y' \overset{\mathsf{def}}{=} Y \backslash \{b\}$. Since $\sigma(\mathbf{var}(\tau)) = (a; b)$, we get that $\Gamma_n \models \sigma(\mathbf{var}(\tau)) \Rightarrow \tau$, and hence $\Gamma_n \models \Sigma(\mathbf{var}(\tau)) \Rightarrow \tau$. This proves the lemma for the case where $|C| = 0$.

So, we assume correctness for elemental terms $(a; b|C)$ where $|C| \leq k$, and prove for $|C| = k + 1$. Since $\Gamma_n \models \Sigma \Rightarrow \tau$, then by Lemma 6.12 there exists a mutual information term $\sigma = (X; Y) \in \Sigma$ such that $XY \supseteq abC$, $abC \cap X \neq \emptyset$, and $abC \cap Y \neq \emptyset$. Hence, we denote $C = C_X C_Y$, where $C_X \overset{\mathsf{def}}{=} X \cap C$ and $C_Y \overset{\mathsf{def}}{=} Y \cap C$. We also denote $X' \overset{\mathsf{def}}{=} X \setminus abC$, and $Y' \overset{\mathsf{def}}{=} Y \setminus abC$. There are two cases. If $\sigma = (aC_X X'; bC_Y Y')$, then $\sigma(\mathbf{var}(\tau)) = (aC_X; bC_Y)$. By Lemma 6.11, it holds that $I_h(aC_X; bC_Y) \geq I_h(a; b|C) = h(\tau)$. Therefore, $\Gamma_n \models \Sigma(\mathbf{var}(\tau)) \Rightarrow \tau$.

Otherwise, w.l.o.g., $\sigma = (abC_X X'; C_Y Y')$. Since $abC \cap Y \neq \emptyset$, then $C_Y \neq \emptyset$. We define

$$\alpha_1 \overset{\mathsf{def}}{=} (a; C_Y | C_X) \qquad \text{and} \qquad \alpha_2 \overset{\mathsf{def}}{=} (a; C_Y | bC_X). \tag{46}$$

By Lemma 6.11, we have that $h(\sigma) \geq h(\alpha_1)$ and $h(\sigma) \geq h(\alpha_2)$. In particular, it holds that

$$h(\sigma(\mathbf{var}(\tau))) = I_h(abC_X; C_Y) \geq I_h(a; C_Y|bC_X) = h(\alpha_2),$$

and thus $\Gamma_n \models \sigma(\mathbf{var}(\tau)) \Rightarrow \alpha_2$.

So, we have that $\Gamma_n \models \Sigma \Rightarrow \{\alpha_1, \alpha_2, \tau\}$, where $\tau = (a; b|C_X C_Y)$. By the chain rule (see (3)), it holds that $(a; C_Y|C_X), (a; b|C_X C_Y) \Leftrightarrow (a; bC_Y|C_X)$. Therefore

$$\Gamma_n \models \Sigma \Rightarrow (a; C_Y|C_X), (a; b|C_X C_Y) \Rightarrow (a; bC_Y|C_X) \Rightarrow \underbrace{(a; b|C_X)}_{\overset{\mathsf{def}}{=}\tau_1}.$$

In other words, we have that $\Gamma_n \models \Sigma \Rightarrow (a; b|C_X)$. We have established that $C_Y \neq \emptyset$, and thus $C_X \subsetneq C$. Therefore, $|C_X| < |C_X C_Y| = |C| = k+1$, and thus $|C_X| \leq k$. By the induction hypothesis, it holds that $\Gamma_n \models \Sigma(abC_X) \Rightarrow \tau_1$. In particular, this means that $\Gamma_n \models \Sigma \backslash \{\sigma\} \Rightarrow \tau_1$ because $C_Y Y' \cap abC_X = \emptyset$. Denoting $\Sigma_1 \overset{\mathsf{def}}{=} \Sigma \backslash \{\sigma\}$, we have that

$$\Gamma_n \models \Sigma_1(\mathbf{var}(\tau_1)) \Rightarrow \tau_1 \qquad \text{and} \qquad \Gamma_n \models \sigma(\mathbf{var}(\tau)) \Rightarrow \alpha_2.$$

By the chain rule (see (3)), this means that

$$\Gamma_n \models \Sigma_1(\mathbf{var}(\tau_1)) \cup \sigma(\mathbf{var}(\tau)) \Rightarrow (a; bC_Y|C_X) \Rightarrow (a; b|C_X C_Y) = (a; b|C) = \tau. \tag{47}$$

Since $\mathbf{var}(\tau_1) = abC_X \subset \mathbf{var}(\tau)$, then by Lemma 6.11 it holds that $h(\Sigma_1(\mathbf{var}(\tau_1))) \leq h(\Sigma_1(\mathbf{var}(\tau)))$. Since $\Gamma_n \models \Sigma_1(\mathbf{var}(\tau_1)) \Rightarrow \tau_1$, then $\Gamma_n \models \Sigma_1(\mathbf{var}(\tau)) \Rightarrow \tau_1$. Since $\Gamma_n \models \sigma(\mathbf{var}(\tau)) \Rightarrow \alpha_2$, then from (47), we get that

$$\Gamma_n \models \Sigma_1(\mathbf{var}(\tau)) \cup \sigma(\mathbf{var}(\tau)) \Rightarrow \tau, \text{ and therefore}$$
$$\Gamma_n \models \Sigma(\mathbf{var}(\tau)) \Rightarrow \tau,$$

as required. This completes the proof. □

Colollary 9.3. *Let Σ be a set of marginal mutual information terms, and let $\tau = (A; B|C)$. It holds that*

$$\Gamma_n \models \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad \Gamma_n \models \Sigma(\mathbf{var}(\tau)) \Rightarrow \tau. \tag{48}$$

Proof. By the chain rule of mutual information, $\Gamma_n \models \Sigma \Rightarrow \tau$ if and only if $\Gamma_n \models \Sigma \Rightarrow (a; b|CA'B')$ for every $a \in A$, $b \in B$, $A' \subseteq A \backslash \{a\}$, and $B' \subseteq B \backslash \{b\}$. By Lemma 9.2, this holds only if $\Gamma_n \models \Sigma(\mathbf{var}(\tau)) \Rightarrow \tau$. The other direction follows from the fact that $\Gamma_n \models h(\Sigma) \geq h(\Sigma(\mathbf{var}(\tau)))$. □

Theorem 9.4. *Let Σ be a set of marginal CIs, and let $\tau = (a; B|C)$, where $a \in \Omega$, and $BC \subseteq \Omega$. It holds that*

$$\Gamma_n \models \Sigma \Rightarrow \tau \qquad \textit{if and only if} \qquad \Gamma_n \models h(\Sigma) \geq h(\tau). \tag{49}$$

Proof. We make the assumption that $\Gamma_n \not\models \Sigma \Rightarrow (a; Bc|C\backslash\{c\})$, for every $c \in C$. This is without loss of generality because otherwise, we prove the claim for $\tau' \overset{\text{def}}{=} (a; Bc|C\backslash\{c\})$. By Lemma 6.11, it holds that $h(\tau') \geq h(\tau)$. Therefore, if $\Gamma_n \models h(\Sigma) \geq h(\tau')$ then $\Gamma_n \models h(\Sigma) \geq h(\tau)$.

By Corollary 9.3, it holds that $\Gamma_n \models \Sigma \Rightarrow \tau$ if and only if $\Gamma_n \models \Sigma(\mathbf{var}(\tau)) \Rightarrow \tau$. So, we prove the claim for $\Sigma(\mathbf{var}(\tau))$. This gives us the desired result because, by Lemma 6.11, it holds that $h(\Sigma) \geq h(\Sigma(\mathbf{var}(\tau)))$. By definition, for every $\sigma \in \Sigma(\mathbf{var}(\tau))$, it holds that $\mathbf{var}(\sigma) \subseteq \mathbf{var}(\tau) = aBC$.

We prove the claim by induction on $|BC|$. If $|BC| = 1$, then $\tau = (a; b)$. By Lemma 6.12, there exists a CI $\sigma = (X; Y) \in \Sigma(ab)$, such that $XY \supseteq ab$. Since, $\mathbf{var}(\Sigma(ab)) \subseteq ab$, we get that $ab \subseteq XY \subseteq ab$, and hence $XY = ab$. Therefore, it must hold that $(X; Y) = (a; b)$, and hence $h(\Sigma) \geq h(\sigma) = h(\tau)$.

So, we assume that the claim holds for $|BC| \leq k$, and prove the claim for the case where $|BC| = k + 1$. Since $\Gamma_n \models \Sigma(aBC) \Rightarrow \tau$, then by Lemma 6.12, there exists a CI $\sigma = (X; Y) \in \Sigma(aBC)$, such that $XY \supseteq aBC$. Since, $\mathbf{var}(\Sigma(aBC)) \subseteq aBC$, we get that $aBC \subseteq XY \subseteq aBC$, and hence $XY = aBC = \mathbf{var}(\tau)$. We denote by $B_X = B \cap X$, $C_X = C \cap X$, $B_Y = B \cap Y$, and $C_Y = C \cap Y$. Since $B = B_X B_Y$ and $C = C_X C_Y$, we can express $\tau = (a; B|C) = (a; B_X B_Y|C_X C_Y)$. There are three options: (1) $\sigma = (X; Y) = (aBC_X; C_Y)$ (symmetrically $\sigma = (C_X; aBC_Y)$), (2) $\sigma = (X; Y) = (BC_X; aC_Y)$ (or, symmetrically that $\sigma = (aC_X; BC_Y)$), and (3) $\sigma = (X; Y) = (aB_X C_X; B_Y C_Y)$ where $B_Y \neq \emptyset$ and $B_X \neq \emptyset$. We prove the claim for each one of these options.

If $\sigma = (X; Y) = (aBC_X; C_Y)$, then $\Gamma_n \models \Sigma \Rightarrow (a; C_Y|C_X), (a; B|C_X C_Y)$ because $\Gamma_n \models \Sigma \Rightarrow \tau = (a; B|C_X C_Y)$, and by Lemma 6.11 it holds that $\sigma \Rightarrow (a; C_Y|C_X)$. By the chain rule, this means that $\Gamma_n \models \Sigma \Rightarrow (a; BC_Y|C_X)$. But this is a contradiction to our assumption that $\Gamma_n \not\models \Sigma \Rightarrow (a; Bc|C\backslash\{c\})$, for every $c \in C$.

If $\sigma = (X; Y) = (BC_X; aC_Y)$, then the claim clearly follows from Lemma 6.11. So, we consider the case where $\sigma = (X; Y) = (aB_X C_X; B_Y C_Y)$ where $B_X \neq \emptyset$ and $B_Y \neq \emptyset$. Using the chain rule (see (3)) we can write $h(\sigma) = I(aB_X C_X; B_Y C_Y)$ as

$$\begin{aligned} I(aB_XC_X; B_YC_Y) &= I\underbrace{(C_Y; aB_XC_X)}_{\overset{\text{def}}{=}\sigma_1} + I\underbrace{(B_Y; aB_XC_X|C_Y)}_{\overset{\text{def}}{=}\sigma_2} \\ &\underbrace{\leq}_{\text{Lemma 6.11}} I(C_Y; aB_XC_X) + I\underbrace{(B_Y; aB_XC)}_{\overset{\text{def}}{=}\sigma_3}. \end{aligned} \tag{50}$$

On the other hand, we can write

$$h(\tau) = I(a; B|C) = I(a; B_X B_Y|C) = I\underbrace{(a; B_X|C)}_{\overset{\text{def}}{=}\tau_1} + I(a; B_Y|CB_X).$$

Since $\Gamma_n \models \Sigma \Rightarrow \tau$, then $\Gamma_n \models \Sigma \Rightarrow \tau_1$. By Corollary 9.3, we have that $\Gamma_n \models \Sigma(aB_XC) \Rightarrow \tau_1$.

Let $\Sigma' \overset{\text{def}}{=} \Sigma\backslash\{\sigma\} \cup \{\sigma_1, \sigma_3\}$. By (50), we have that Σ' is marginal, and since $h(\sigma) \leq h(\sigma_1) + h(\sigma_3)$, then $\Gamma_n \models \Sigma' \Rightarrow \Sigma$, and $h(\Sigma) \leq h(\Sigma')$. Therefore, $\Gamma_n \models \Sigma' \Rightarrow \tau$, and in particular, $\Gamma_n \models \Sigma' \Rightarrow \tau_1$. By Corollary 9.3, we have that $\Gamma_n \models \Sigma'(aB_XC) \Rightarrow \tau_1$. Since Σ'

is marginal, and $B_Y \cap aB_XC = \emptyset$, then $\Gamma_n \models \Sigma'(aB_XC)\backslash\{\sigma_3\} \Rightarrow \tau_1$. We observe that

$$h(\Sigma) \geq h(\underbrace{\Sigma\backslash\{\sigma\} \cup \{\sigma_1\}}_{\overset{\mathsf{def}}{=}\Sigma_1}) = h(\Sigma'\backslash\{\sigma_3\}) \geq h(\Sigma'(aB_XC)\backslash\{\sigma_3\}) = h(\Sigma'(aB_XC)).$$

Hence, $\Gamma_n \models \Sigma'(aB_XC) \Rightarrow \tau_1$, and $\Gamma_n \models \Sigma_1 \Rightarrow \Sigma'(aB_XC)$. Consequently, we get that $\Gamma_n \models \Sigma_1 \Rightarrow \tau_1$, where Σ_1 is marginal. From (50), we have that $h(\Sigma_1) = h(\Sigma) - h(\sigma_2)$. Since $B_Y \neq \emptyset$, then $|B_XC| < |BC| = k+1$, and hence $|B_XC| \leq k$. Therefore, by the induction hypothesis, we get that $\Gamma_n \models h(\tau_1) \leq h(\Sigma_1)$. Now, we get that

$$h(\tau) \leq h(\tau_1) + h(\sigma_2) \leq h(\Sigma_1) + h(\sigma_2) = h(\Sigma) - h(\sigma_2) + h(\sigma_2) = h(\Sigma).$$

This completes the proof. □

10. Conclusion and Future Work

We consider the problem of approximate implication for conditional independence. In the general case, approximate implication does not hold (Kenig & Suciu, 2022). Therefore, we establish results and approximation guarantees under various restrictions to the derivation rules, and antecedents (our results are summarized in Table 1). We establish new and tighter approximation bounds when the set of antecedents are saturated, or marginal. We also prove a negative result showing that approximate CIs cannot be inferred from the independence graph associated with a Markov network. The results in this paper characterize settings in which approximate CIs can be used in place of exact CIs. This, for example, holds for the important case of Bayesian Networks.

As part of future work, we intend to investigate restrictions to probability distributions that allow the intersection axiom to relax. In addition, we intend to explore how our results can be used to efficiently generate separating candidate sets that represent approximate CIs, which can be used to synthesize *decomposable models* (de Campos & Huete, 1997), that provide a good approximation of the underlying data or distribution (Kenig & Weinberger, 2023).

Acknowledgments

This work was supported by the US-Israel Binational Science Foundation (BSF) Grant No. 2020751.

APPENDIX

A. Missing Proofs from Section 5

Lemma 5.2. *The following holds for any $x \in (0, \frac{1}{3})$, and $y \in (0, \frac{1}{6})$.*

1. *$H(A_i) = 1$ for $i \in \{1, 2, 3\}$.*
2. *$H(A_i) = H(A_{i,1}, A_{i,2}) = \delta_1(x)$ for $i \in \{4, 5, 6\}$.*

3. $H(A_{i,1}) = H(A_{i,2}) = \delta_2(x)$ *for* $i \in \{4,5,6\}$.

4. $H(A_{7,j}) = 1$ *for* $j \in \{1,2,3\}$.

5. $H(A_{7,j}, A_{7,k}) = f_2(y)$ *for* $j \neq k$ *and* $j,k \in \{1,2,3\}$.

6. $H(A_7) = H(A_{7,1}, A_{7,2}, A_{7,3}) = f_1(y)$.

where $\delta_1, \delta_2, f_1,$ *and* f_2 *are defined in* (16)–(19).

Proof. The proof is by definition, and we provide the technical details. For $i \in \{1,2,3\}$:

$$H(A_i) = 2 \cdot \frac{1}{2} \log 2 = 1$$

By the definition of A_4 in Table 3, we have that

$$H(A_4) = -\big((1-3x)\log(1-3x) + 3x \log x\big) = \delta_1(x).$$

Proof is the same for $H(A_5)$ and $H(A_6)$.

We now compute $H(A_{4,1})$. From Table 3, we have that $P(A_{4,1} = 0) = 1-3x+x = 1-2x$, and $P(A_{4,1} = 1) = 2x$. Therefore,

$$H(A_{4,1}) = -\big((1-2x)\log(1-2x) + 2x \log 2x\big) = \delta_2(x).$$

For symmetry reasons, the same holds for $H(A_{4,2})$ and $H(A_{i,j})$ for $i \in \{5,6\}$ and $j \in \{1,2\}$.

From Table 2, we have that for all $i \in \{1,2,3\}$:

$$P(A_{7,i} = 0) = P(A_{7,i} = 1) = \frac{1}{2}.$$

Therefore, $H(A_{7,i}) = 1$ for all $i \in \{1,2,3\}$.

Now, take any $i, j \in \{1,2,3\}$ where $i < j$. Then, from Table 2, we have that

$$P(A_{7,i} = a, A_{7,j} = b) = \begin{cases} \frac{1}{2} - 2y & \text{if } a = b \\ 2y & \text{otherwise} \end{cases}$$

Therefore, we have that

$$H(A_{7,i}, A_{7,j}) = -\left(2(\frac{1}{2} - 2y)\log(\frac{1}{2} - 2y) + 2 \cdot 2y \log 2y\right) = f_2(y).$$

Finally, by Table 2, we have that

$$H(A_7) = H(A_{7,1}, A_{7,2}, A_{7,3}) = -\left(2(\frac{1}{2} - 3y)\log(\frac{1}{2} - 3y) + 6y \log y\right) = f_1(y).$$

□

Lemma 5.3. *The following holds:*

1. $H(A) = H(B) = H(C) = 2 + 2\delta_2(x)$.

2. $H(AC) = H(AB) = H(BC) = 2 + \delta_1(x) + 2\delta_2(x) + f_2(y)$.

3. $H(ABC) = 3 + 3\delta_1(x) + f_1(y)$.

Proof.

$$\begin{aligned}
H(A) &\underbrace{=}_{(13)} H(A_2, A_{6,1}, A_{7,1}, A_{5,1}) \\
&\underbrace{=}_{\substack{A_1,\ldots,A_7 \text{ are} \\ \text{mutually-independent}}} H(A_2) + H(A_{6,1}) + H(A_{7,1}) + H(A_{5,1}) \\
&\underbrace{=}_{\text{Lemma 5.2}} 1 + \delta_2(x) + 1 + \delta_2(x) \\
&= 2 + 2\delta_2(x)
\end{aligned}$$

By symmetry, we have that $H(B) = H(C) = 2 + 2\delta_2(x)$ as well.

We now compute $H(A|C)$.

$$\begin{aligned}
H(A|C) &\underbrace{=}_{(13),(15)} H(A_2, A_{6,1}, A_{7,1}, A_{5,1}|A_1, A_{5,2}, A_{7,3}, A_{4,2}) \\
&\underbrace{=}_{\substack{\text{chain rule} \\ \text{for entropy}}} H(A_{5,1}, A_{7,1}|A_1, A_{5,2}, A_{7,3}, A_{4,2}) + H(A_2, A_{6,1}|A_1, A_{5,2}, A_{7,3}, A_{4,2}, A_{5,1}, A_{7,1}) \\
&\underbrace{=}_{\text{independence}} H(A_{5,1}, A_{5,2}, A_{7,1}, A_{7,3}, A_1, A_{4,2}) - H(A_1, A_{5,2}, A_{7,3}, A_{4,2}) + H(A_2, A_{6,1}) \\
&= H(A_{5,1}, A_{5,2}) + H(A_{7,1}, A_{7,3}) + H(A_1) + H(A_{4,2}) \\
&\quad - H(A_1) - H(A_{5,2}) - H(A_{7,3}) - H(A_{4,2}) + H(A_2) + H(A_{6,1}) \\
&= H(A_{5,1}, A_{5,2}) + H(A_{7,1}, A_{7,3}) - H(A_{5,2}) - H(A_{7,3}) + H(A_2) + H(A_{6,1}) \\
&\underbrace{=}_{\text{Lemma 5.2}} \delta_1(x) + f_2(y) - \delta_2(x) - 1 + 1 + \delta_2(x) \\
&= \delta_1(x) + f_2(y)
\end{aligned}$$

Now, since $H(AC) = H(A|C) + H(C)$, then by the above, and the fact that $H(C) = 2 + 2\delta_2(x)$, we get that

$$H(AC) = H(A|C) + H(C) = \delta_1(x) + f_2(y) + 2 + 2\delta_2(x)$$

as required.

Finally, we compute $H(A|BC)$.

$$\begin{aligned}
H(A|BC) \underbrace{=}_{(13),(14),(15)} & H(A_2, A_{6,1}, A_{7,1}, A_{5,1}|A_1, A_{5,2}, A_{7,3}, A_{4,2}, A_3, A_{6,2}, A_{7,2}, A_{4,1}) \\
&= H(A_1, A_2, A_3, (A_{4,1}, A_{4,2}), (A_{6,1}, A_{6,2}), (A_{7,1}, A_{7,2}, A_{7,3}), (A_{5,1}, A_{5,2})) \\
&\quad - H(A_1, A_3, (A_{4,1}, A_{4,2}), A_{5,2}, A_{6,2}, A_{7,2}, A_{7,3}) \\
&= H(A_1) + H(A_2) + H(A_3) + H(A_4) + H(A_5) + H(A_6) + H(A_7) \\
&\quad - H(A_1) - H(A_3) - H(A_4) - H(A_{5,2}) - H(A_{6,2}) - H(A_{7,2}, A_{7,3}) \\
&= H(A_2) + H(A_5) + H(A_6) + H(A_7) - H(A_{5,2}) - H(A_{6,2}) - H(A_{7,2}, A_{7,3}) \\
&\underbrace{=}_{\text{Lemma 5.2}} 1 + \delta_1(x) + \delta_1(x) + f_1(y) - \delta_2(x) - \delta_2(x) - f_2(y) \\
&= 1 + 2\delta_1(x) + f_1(y) - f_2(y) - 2\delta_2(x)
\end{aligned}$$

Since $H(ABC) = H(A|BC) + H(BC)$, we get that

$$\begin{aligned}
H(ABC) &= 1 + 2\delta_1(x) + f_1(y) - f_2(y) - 2\delta_2(x) + (\delta_1(x) + f_2(y) + 2 + 2\delta_2(x)) \\
&= 3 + 3\delta_1(x) + f_1(y).
\end{aligned}$$

□

References

Armstrong, W. W., & Delobel, C. (1980). Decomposition and functional dependencies in relations. *ACM Trans. Database Syst.*, *5*(4), 404–430.

Beeri, C. (1980). On the menbership problem for functional and multivalued dependencies in relational databases. *ACM Trans. Database Syst.*, *5*(3), 241–259.

Beeri, C., Fagin, R., & Howard, J. H. (1977). A complete axiomatization for functional and multivalued dependencies in database relations. In *Proceedings of the 1977 ACM SIGMOD International Conference on Management of Data, Toronto, Canada, August 3-5, 1977.*, pp. 47–61.

Chen, X., Anantha, G., & Lin, X. (2008). Improving bayesian network structure learning with mutual information-based node ordering in the k2 algorithm. *IEEE Transactions on Knowledge and Data Engineering*, *20*(5), 628–640.

Cheng, J., Greiner, R., Kelly, J., Bell, D., & Liu, W. (2002). Learning bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, *137*(1), 43 – 90.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(1), 1–31.

de Campos, L. M. (2006). A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, *7*(77), 2149–2187.

de Campos, L. M., & Huete, J. F. (1997). Algorithms for learning decomposable models and chordal graphs. In Geiger, D., & Shenoy, P. P. (Eds.), *UAI '97: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, Brown University, Providence, Rhode Island, USA, August 1-3, 1997*, pp. 46–53. Morgan Kaufmann.

Geiger, D., Paz, A., & Pearl, J. (1991a). Axioms and algorithms for inferences involving probabilistic independence. *Inf. Comput.*, *91*(1), 128–141.

Geiger, D., Paz, A., & Pearl, J. (1991b). Axioms and algorithms for inferences involving probabilistic independence. *Information and Computation*, *91*(1), 128 – 141.

Geiger, D., & Pearl, J. (1988). On the logic of causal models. In *UAI '88: Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence, Minneapolis, MN, USA, July 10-12, 1988*, pp. 3–14.

Geiger, D., & Pearl, J. (1993). Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics*, *21*(4), 2001–2021.

Geiger, D., Verma, T., & Pearl, J. (1989). d-separation: From theorems to algorithms. In Henrion, M., Shachter, R. D., Kanal, L. N., & Lemmer, J. F. (Eds.), *UAI '89: Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence, Windsor, Ontario, Canada, August 18-20, 1989*, pp. 139–148. North-Holland.

Geiger, D., Verma, T., & Pearl, J. (1990). Identifying independence in Bayesian networks. *Networks*, *20*(5), 507–534.

Gyssens, M., Niepert, M., & Gucht, D. V. (2014). On the completeness of the semigraphoid axioms for deriving arbitrary from saturated conditional independence statements. *Inf. Process. Lett.*, *114*(11), 628–633.

Herrmann, C. (1995). On the undecidability of implications between embedded multivalued database dependencies. *Inf. Comput.*, *122*(2), 221–235.

Kenig, B. (2021). Approximate implication with d-separation. In de Campos, C. P., Maathuis, M. H., & Quaeghebeur, E. (Eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI 2021, Virtual Event, 27-30 July 2021*, Vol. 161 of *Proceedings of Machine Learning Research*, pp. 301–311. AUAI Press.

Kenig, B., Mundra, P., Prasaad, G., Salimi, B., & Suciu, D. (2020). Mining approximate acyclic schemes from relations. In Maier, D., Pottinger, R., Doan, A., Tan, W., Alawini, A., & Ngo, H. Q. (Eds.), *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, pp. 297–312. ACM.

Kenig, B., & Suciu, D. (2020). Integrity constraints revisited: From exact to approximate implication. In Lutz, C., & Jung, J. C. (Eds.), *23rd International Conference on Database Theory, ICDT 2020, March 30-April 2, 2020, Copenhagen, Denmark*, Vol. 155 of *LIPIcs*, pp. 18:1–18:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Kenig, B., & Suciu, D. (2022). Integrity constraints revisited: From exact to approximate implication. *Log. Methods Comput. Sci.*, *18*(1).

Kenig, B., & Weinberger, N. (2023). Quantifying the loss of acyclic join dependencies. In Geerts, F., Ngo, H. Q., & Sintos, S. (Eds.), *Proceedings of the 42nd ACM SIGMOD-*

SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2023, Seattle, WA, USA, June 18-23, 2023, pp. 329–338. ACM.

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models - Principles and Techniques*. MIT Press.

Kontinen, J., Link, S., & Väänänen, J. (2013). Independence in database relations. In Libkin, L., Kohlenbach, U., & de Queiroz, R. (Eds.), *Logic, Language, Information, and Computation*, pp. 179–193, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lee, T. T. (1987). An information-theoretic analysis of relational databases - part I: data dependencies and information metric. *IEEE Trans. Software Eng.*, *13*(10), 1049–1061.

Li, C. T. (2023). Undecidability of network coding, conditional information inequalities, and conditional independence implication. *IEEE Trans. Inf. Theory*, *69*(6), 3493–3510.

Maier, D. (1983). *Theory of Relational Databases*. Computer Science Pr.

Pearl, J. (1989). *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann.

Pearl, J., Geiger, D., & Verma, T. (1989). Conditional independence and its representations. *Kybernetika*, *25*(7), 33–44.

Pearl, J., & Paz, A. (1986). Graphoids: Graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z?. In *ECAI*, pp. 357–363.

Sayrafi, B., Van Gucht, D., & Gyssens, M. (2008). The implication problem for measure-based constraints. *Information Systems*, *33*(2), 221 – 239. Performance Evaluation of Data Management Systems.

Studený, M. (1990). Conditional independence relations have no finite complete characterization. In *11th Prague Conf. Information Theory, Statistical Decision Foundation and Random Processes*, pp. 377–396. Norwell, MA.

Studený, M. (2018). Conditional independence and markov properties for basic graphs. In Maathuis, M., Drton, M., Lauritzen, S., & Wainwright, M. (Eds.), *Handbook of Graphical Models*, pp. 3–38. CRC Press.

Verma, T., & Pearl, J. (1988). Causal networks: semantics and expressiveness. In Shachter, R. D., Levitt, T. S., Kanal, L. N., & Lemmer, J. F. (Eds.), *UAI '88: Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence, Minneapolis, MN, USA, July 10-12, 1988*, pp. 69–78. North-Holland.

Verma, T., & Pearl, J. (1990). Causal networks: Semantics and expressiveness. In SHACHTER, R. D., LEVITT, T. S., KANAL, L. N., & LEMMER, J. F. (Eds.), *Uncertainty in Artificial Intelligence*, Vol. 9 of *Machine Intelligence and Pattern Recognition*, pp. 69–76. North-Holland.

Yeung, R. W. (1991). A new outlook of shannon's information measures. *IEEE Trans. Information Theory*, *37*(3), 466–474.

Yeung, R. W. (2008). *Information Theory and Network Coding* (1 edition). Springer Publishing Company, Incorporated.

Zhao, J., Zhou, Y., Zhang, X., & Chen, L. (2016). Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences*, *113*(18), 5130–5135.