Approximate Implication with d-Separation

Batya Kenig

Technion, Israel Institute of Technology
Haifa, Israel
batyak@technion.ac.il

Abstract

The graphical structure of Probabilistic Graphical Models (PGMs) encodes the conditional independence (CI) relations that hold in the modeled distribution. Graph algorithms, such as *d-separation*, use this structure to infer additional conditional independencies, and to query whether a specific CI holds in the distribution. 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. 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 prove an approximation guarantee for independence relations derived from *marginal* 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 [17, 21].

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 [10, 12, 15, 16, 22, 23]. Known algorithms for deriving CIs from the topological structure of the graphical model are, in fact, an instance of implication. Notably, the DAG structure of Bayesian Networks is generated based on a set of CIs termed the *recursive basis* [11], and the d-separation algorithm is used to derive additional CIs, implied by this set. The d-separation algorithm is a sound and complete method for deriving CIs in probability distributions represented by DAGs [10, 11], and hence completely characterizes the CIs that hold in the distribution. The foundation of deriving CIs in both directed and undirected models is the *semigraphoid axioms* [6, 9, 13].

Current systems for inferring CIs, and the semigraphoid 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 [3, 4, 7, 16, 30].

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 we can convert an exact implication to an approximate one. When relaxation holds, then any system-of-inference for deriving exact implications, (e.g. the semigraphoid axioms,

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

d-separation), can be used to infer approximate implications as well.

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. Hence, 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 triples $(X;Y|Z)$, and h is the conditional mutual information measure $I(\cdot;\cdot|\cdot)$. 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 $I(\cdot;\cdot|\cdot) \geq 0$ is a nonnegative measure. 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 distribution, and it is called *marginal* if $C = \emptyset$.

We show that every conditional independence relation $(X;Y|Z)$ read off a DAG by the d-separation algorithm [10], admits a 1-approximation. In other words, if Σ is the *recursive basis* of CIs used to build the Bayesian network [10], then it is guaranteed that $I(X;Y|Z) \leq \sum_{i \in \Sigma} h(\sigma_i)$. Furthermore, we present a family of implications for which our 1-approximation is tight (i.e., $I(X;Y|Z) = \sum_{i \in \Sigma} h(\sigma_i)$). We also prove that every CI $(X;Y|Z)$ implied by a set of marginal CIs admits an $|X| \cdot |Y|$-approximation (i.e., where $|X|$ denotes the number of RVs in the set X). The exact variant of implication from these classes of CIs were extensively studied [8, 9, 10, 11, 12] (see below the related work). Here, we study their approximation.

Of independent interest is the technique used for proving the approximation guarantees. The *I-measure* [28] 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. Ours is the first to apply this technique to the study of CI implication.

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* [22] are sound and complete for deriving CI statements that are implied by saturated CIs [9], marginal CIs [9], and *recursive CIs* that are used in Bayesian networks [8, 11]. 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* [27]. Studený proved that in the general case, when no assumptions are made on the antecendents, no finite axiomatization exists [25]. That is, there does not exist a finite set of axioms (deductive rules) from which all general conditional independence implications can be deduced.

The database community has also studied the EI problem for integrity constraints [1, 2, 18, 20], and showed that the implication problem is decidable and axiomatizable when the antecedents are Functional Dependencies or *Multivalued Dependencies* (which correspond to saturated CIs, see [15, 19]), and undecidable for *Embedded Multivalued Dependencies* [14].

The relaxation problem was first studied by Kenig and Suciu in the context of database dependencies [15], 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 [15]). Consequently, it is not straightforward that exact implication necessarily imply its approximation counterpart, and arriving at meaningful approximation guarantees requires making certain assumptions on the antecedents, consequent, or both.

Organization. We start in Section 2 with preliminaries. We formally define the relaxation problem in Section 3, and formally state our results in Section 4. In Section 5 we establish, through a series of lemmas, properties of exact implication that will be used for proving our results. In Section 6 we prove that every implication from a set of recursive CIs admits a 1-relaxation, and in Section 7 we prove that every implication $\Sigma \Rightarrow (X;Y|Z)$ from a set of marginal CIs admits an $|X||Y|$-relaxation. We conclude in Section 8.

2 PRELIMINARIES

We denote by $[n] = \{1, 2, \ldots, n\}$. 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. 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 $p \models X_\alpha \rightarrow X_\beta$.

An assertion $X \perp Y | Z$ is called a *Conditional Independence* statement, or a CI; this includes $Z \to Y$ as a special case. 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 [29]. For $n > 0$, we identify the functions $2^{[n]} \to \mathbb{R}$ with the vectors in $\mathbb{R}^{2^n}$.

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]

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

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

Then, $\forall h \in \Gamma_n$, $I_h(B;C|A) \geq 0$ by submodularity, and $h(B|A) \geq 0$ by monotonicity. We say that A *functionally determines* B, in notation $A \to B$ 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 $|B||C| \leq \frac{n^2}{4}$ 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

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

For a set of jointly distributed random variables $\Omega = \{X_1, \ldots, X_n\}$ we define the function $h : 2^{[n]} \to \mathbb{R}$ as $h(\alpha) \stackrel{\text{def}}{=} H(X_\alpha)$; 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 iff $I_h(B;C|A) = 0$, and similarly $p \models A \to B$ iff $h(B|A) = 0$, thus, entropy provides us with an alternative characterization of CIs.

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

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 1: Information measures and associated I-measure

2.2.1 The I-measure

The I-measure [28, 29] 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 $\{X_1, \ldots, X_n\}$. Every variable X_i is associated with a set $\mathrm{m}(X_i)$, and it's 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 $X_\alpha \stackrel{\text{def}}{=} \{X_j \mid j \in \alpha\}$, and $\mathrm{m}(X_\alpha) \stackrel{\text{def}}{=} \bigcup_{i \in \alpha} \mathrm{m}(X_i)$.

Definition 2.1. ([28, 29]) *The field $\mathcal{F}_n$ generated by sets $\mathrm{m}(X_1), \ldots, \mathrm{m}(X_n)$ is the collection of sets which can be obtained by any sequence of usual set operations (union, intersection, complement, and difference) on $\mathrm{m}(X_1), \ldots, \mathrm{m}(X_n)$.*

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 $\mu^*(m(X_\alpha)) = H(X_\alpha)$ for all nonempty subsets $\alpha \subseteq \{1, \ldots, n\}$, where H is the entropy (4). Table 1 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 2.2. ([28, 29]) [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, and mutual information).*

Let $\sigma = (X;Y|Z)$. We denote by $\mathrm{m}(\sigma) \stackrel{\text{def}}{=} \mathrm{m}(X) \cap \mathrm{m}(Y) \cap \mathrm{m}^c(Z)$ the set associated with σ (see Table 1). For a set of triples Σ, we define:

$$\mathrm{m}(\Sigma) \stackrel{\text{def}}{=} \bigcup_{\sigma \in \Sigma} \mathrm{m}(\sigma) \tag{5}$$

Example 2.3. *Let A, B, and C be three disjoint sets of RVs defined as follows: $A{=}A_1A_2A_3$, $B{=}B_1B_2$ and $C{=}C_1C_2$. Then, by Theorem 2.2: $H(A){=}\mu^*(\mathrm{m}(A)){=}\mu^*(\mathrm{m}(A_1){\cup}\mathrm{m}(A_2){\cup}\mathrm{m}(A_3))$, $H(B){=}\mu^*(\mathrm{m}(B)){=}\mu^*(\mathrm{m}(B_1){\cup}\mathrm{m}(B_2))$, and $\mu^*(\mathrm{m}^c(C)){=}\mu^*(\mathrm{m}^c(C_1){\cap}\mathrm{m}^c(C_2))$. By Table 1: $I(A;B|C){=}\mu^*(\mathrm{m}(A) \cap \mathrm{m}(B) \cap \mathrm{m}^c(C))$.*

We denote by Δ_n the set of signed measures $\mu^* : \mathcal{F}_n \rightarrow \mathbb{R}_{\geq 0}$ that assign non-negative values to the atoms $\mathcal{F}_n$. We call these *positive I-measures*.

Theorem 2.4. ([29]) *If there is no constraint on $X_1, \ldots, X_n$, then μ^* can take any set of nonnegative values on the nonempty atoms of $\mathcal{F}_n$.*

Theorem 2.4 implies that every positive I-measure μ^* corresponds to a function that is consistent with the Shannon inequalities, and is thus a polymatroid. Hence, $\Delta_n \subset \Gamma_n$ is the set of polymatroids with a positive I-measure that we call *positive polymatroids*.

2.3 BAYESIAN NETWORKS

A Bayesian network encodes the CIs of a probability distribution using a Directed Acyclic Graph (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

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi(i)}) \tag{6}$$

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

Equation 6 implicitly encodes a set of n conditional independence statements, called the *recursive basis* for the network:

$$\Sigma \stackrel{\text{def}}{=} \{(X_i; X_1 \ldots X_{i-1} \backslash \pi(X_i) \mid \pi(X_i)) : i \in [n]\} \tag{7}$$

The implication problem associated with Bayesian Networks is to determine whether $\Sigma \Rightarrow \tau$ for a CI τ. Geiger and Pearl have shown that $\Sigma \Rightarrow \tau$ iff τ can be derived from Σ using the *semigraphoid axioms* [11]. Their result establishes that the semigraphoid axioms are sound and complete for inferring CI statements from the recursive basis.

3 THE RELAXATION PROBLEM

We now formally define the relaxation problem. We fix a set of variables $\Omega = \{X_1, \ldots, X_n\}$, and consider triples of the form $\sigma = (Y; Z|X)$, 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 a CI $\sigma = (Y; Z|X)$, we define $h(\sigma) \stackrel{\text{def}}{=} I_h(Y; Z|X)$, for a set of CIs Σ, we define $h(\Sigma) \stackrel{\text{def}}{=} \sum_{\sigma \in \Sigma} h(\sigma)$. 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, forall $h \in K$, $h(\Sigma) = 0$ 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 $x{\in}\{EI, AI\}$.

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:*

$$K \models_{EI} \Sigma \Rightarrow \tau \text{ iff } K \models_{AI} \lambda \cdot h(\Sigma) \geq h(\tau).$$

In this paper, we focus on λ-relaxation in the set Γ_n of polymatroids, and two syntactically-defined classes: 1) Where Σ is the recursive basis of a Bayesian network (see (7)), and 2) Where Σ is a set of marginal CIs.

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

4 FORMAL STATEMENT OF RESULTS

We generalize the results of Geiger et al. [10, 13], by proving that implicates $\tau{=}(X; Y|Z)$ of the recursive set [10], and of marginal CIs [13], admit a 1, and $|X||Y|$-approximation respectively, and thus continue to hold also approximately.

4.1 IMPLICATION FROM RECURSIVE CIS

Geiger et al. [10] prove that the semigraphoid axioms are sound and complete for the implication from the recursive set (see (7)). They further showed that the set of implicates can be read off the appropriate DAG via the d-separation procedure. We show that every such exact implication can be relaxed, admitting a 1-relaxation, guaranteeing a bounded approximation for the implicates (CI relations) read off the DAG by d-separation.

We recall the definition of the recursive basis Σ from (7):

$$\Sigma \stackrel{\text{def}}{=} \{(X_i; R_i|B_i) : i \in [1, n], R_iB_i = U^{(i)}\} \tag{8}$$

where $B_i \stackrel{\text{def}}{=} \pi(X_i)$ and $U^{(i)} \stackrel{\text{def}}{=} \{X_1, \ldots, X_{i-1}\}$. We observe that $|\Sigma|=n$, there is a single triple $\sigma_n=(X_n; R_n|B_n) \in \Sigma$ that mentions X_n, and that σ_n is saturated.

We recall that $\Delta_n \subset \Gamma_n$ is the set of polymatroids whose I-measure assigns non-negative values to the atoms $\mathcal{F}_n$ (see Section 2.2.1).

Theorem 4.1. *Let Σ be a recursive set of CIs (see* (8)*), and let $\tau = (A; B|C)$. Then the following holds:*

$$\Delta_n \models_{EI} \Sigma \Rightarrow \tau \qquad \textit{iff} \qquad \Gamma_n \models h(\Sigma) \geq h(\tau) \tag{9}$$

We note that the only-if direction of Theorem 4.1 is immediate, and follows from the non-negativity of Shannon's information measures. We prove the other direction in Section 6. Theorem 4.1 states that it is enough that the exact implication holds on all of the positive polymatroids Δ_n, because this implies the (even stronger!) statement $\Gamma_n \models h(\Sigma) \geq h(\tau)$.

4.2 IMPLICATION FROM MARGINAL CIS

We show that *any* implicate $\tau=(A; B|C)$ from a set of marginal CIs has an $|A|\cdot|B|$-approximation. This generalizes the result of Geiger, Paz, and Pearl [13], which proved that the semigraphoid axioms are sound and complete for deriving marginal CIs.

Theorem 4.2. *Let Σ be a set of marginal CIs, and $\tau = (A; B|C)$ be any CI.*

$$\Gamma_n \models_{EI} \Sigma \Rightarrow \tau \quad \textit{iff} \quad \Gamma_n \models (|A||B|)h(\Sigma) \geq h(\tau) \tag{10}$$

Also here, the only-if direction of Theorem 4.2 is immediate, and we prove the other direction in Section 7.

5 PROPERTIES OF EXACT IMPLICATION

In this section, we use the I-measure to characterize some general properties of exact implication in the set of positive polymatroids Δ_n (Section 5.1), and the entire set of polymatroids Γ_n (Section 5.2). The lemmas in this section will be used for proving the approximate implication guarantees presented in Section 4.

In what follows, $\Omega = \{X_1, \ldots, X_n\}$ is a set of n RVs, Σ denotes a set of triples $(A; B|C)$ representing mutual information terms, and τ denotes a single triple. We denote by $\mathbf{var}(\sigma)$ the set of RVs mentioned in σ (e.g., if $\sigma = (X_1X_2; X_3|X_5)$ then $\mathbf{var}(\sigma) = X_1 \ldots X_5$).

5.1 EXACT IMPLICATION IN THE SET OF POSITIVE POLYMATROIDS

Lemma 5.1. *The following holds:*

$$\Delta_n \models_{EI} \Sigma \Rightarrow \tau \qquad \textit{iff} \qquad \mathrm{m}(\Sigma) \supseteq \mathrm{m}(\tau)$$

Proof. Suppose that $\mathrm{m}(\tau) \not\subseteq \mathrm{m}(\Sigma)$, and let $b \in \mathrm{m}(\tau) \setminus \mathrm{m}(\Sigma)$. By Theorem 2.4 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 $\mu^*: \mathcal{F}_n \rightarrow \mathbb{R}_{\geq 0}$, we have that $\mu^*(\mathrm{m}(\Sigma)) \geq \mu^*(\mathrm{m}(\tau))$. By Theorem 2.2, 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 5.1 is that $\mathrm{m}(\Sigma) \supseteq \mathrm{m}(\tau)$ is a necessary condition for implication between polymatroids.

Corollary 5.2. *If $\Gamma_n \models_{EI} \Sigma \Rightarrow \tau$ then $\mathrm{m}(\Sigma) \supseteq \mathrm{m}(\tau)$.*

Proof. If $\Gamma_n \models_{EI} \Sigma \Rightarrow \tau$ then it must hold for any subset of polymatroids, and in particular, $\Delta_n \models_{EI} \Sigma \Rightarrow \tau$. The result follows from Lemma 5.1. □

Lemma 5.3. *Let $\Delta_n \models_{EI} \Sigma \Rightarrow \tau$, and let $\sigma \in \Sigma$ such that $\mathrm{m}(\sigma) \cap \mathrm{m}(\tau) = \emptyset$. Then $\Delta_n \models_{EI} \Sigma \setminus \{\sigma\} \Rightarrow \tau$.*

Proof. Let $\Sigma' = \Sigma \setminus \{\sigma\}$, and suppose that $\Delta_n \not\models_{EI} \Sigma' \Rightarrow \tau$. By Lemma 5.1, 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 $a \in \mathrm{m}(\tau) \setminus \mathrm{m}(\Sigma')$. 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 5.1 we get that $\Delta_n \not\models_{EI} \Sigma \Rightarrow \tau$. □

5.2 EXACT IMPLICATION IN THE SET OF POLYMATROIDS

The main technical result of this section is Lemma 5.6. We start with two short technical lemmas.

Lemma 5.4. *Let $\sigma = (A; B|C)$ and $\tau = (X; Y|Z)$ be CIs such that $X \subseteq A$, $Y \subseteq B$, $C \subseteq Z$ and $Z \subseteq ABC$. Then, $\Gamma_n \models h(\tau) \leq h(\sigma)$.*

Proof. Since $Z \subseteq ABC$, we denote by $Z_A = A \cap Z$, $Z_B = B \cap Z$, and $Z_C = C \cap Z$. Also, denote by $A' = A \setminus (Z_A \cup X)$, $B' = B \setminus (Z_B \cup Y)$. So, we have that: $I(A; B|C) = I(Z_A A' X; Z_B B' Y|C)$. By the chain rule, we have that:

$$\begin{aligned}
&I(Z_A A' X; Z_B B' Y|C) = \\
&I(Z_A; Z_B|C) + I(A'X; Z_B|CZ_A) \\
&+ I(Z_A; B'Y|Z_B C) + \boldsymbol{I(X; Y|CZ_A Z_B)} \\
&+ I(X; B'|CZ_A Z_B Y) + I(A'; B'Y|CZ_A Z_B X)
\end{aligned}$$

Noting that $Z = CZ_A Z_B$, we get that $I(X; Y|Z) \leq I(A; B|C)$ as required. □

Lemma 5.5. *Let $\Sigma = \{\sigma_1, \dots, \}$ be a set of triples such that $\mathbf{var}(\sigma_i) \subseteq \{X_1, \dots, X_{n-1}\}$ for all $\sigma_i \in \Sigma$. Likewise, let τ be a triple such that $\mathbf{var}(\tau) \subseteq \{X_1, \dots, X_{n-1}\}$. Then:*

$$\Gamma_n \models_{EI} \Sigma \Rightarrow \tau \qquad \textit{iff} \qquad \Gamma_{n-1} \models_{EI} \Sigma \Rightarrow \tau \tag{11}$$

Proof. Suppose that $\Gamma_n \not\models_{EI} \Sigma \Rightarrow \tau$. Then there exists a polymatroid (Section 2.2) $f : 2^{[n]} \to \mathbb{R}$ such that $f(\sigma) = 0$ for all $\sigma \in \Sigma$, and $f(\tau) \neq 0$. We define $g : 2^{[n-1]} \to \mathbb{R}$ as follows:

$$g(A) = f(A) \quad \text{ for all } \quad A \subseteq \{X_1, \dots, X_{n-1}\} \tag{12}$$

Since f is a polymatroid, then so is g. Further, since Σ does not mention X_n then, by (12), we have that $g(\sigma) = f(\sigma)$ for all $\sigma \in \Sigma$. Hence, $\Gamma_{n-1} \not\models_{EI} \Sigma \Rightarrow \tau$.

If $\Gamma_{n-1} \not\models_{EI} \Sigma \Rightarrow \tau$. Then there exists a polymatroid $g : 2^{[n-1]} \to \mathbb{R}$ such that $g(\sigma) = 0$ for all $\sigma \in \Sigma$, and $g(\tau) \neq 0$. Define $f : 2^{[n]} \to \mathbb{R}$ as follows:

$$f(A) = g(A \setminus X_n) \quad \text{ for all } \quad A \subseteq \{X_1, \dots, X_n\} \tag{13}$$

We claim that $f \in \Gamma_n$ (i.e., f is a polymatroid). It then follows that $\Gamma_n \not\models \Sigma \Rightarrow \tau$ because by the assumption that $\mathbf{var}(\Sigma)$ and $\mathbf{var}(\tau)$ are subsets of $\{X_1, \dots, X_{n-1}\}$, then $f(\sigma) = g(\sigma)$ for all $\sigma \in \Sigma$. Hence, $f(\Sigma) = g(\Sigma) = 0$ while $f(\tau) = g(\tau) \neq 0$.

We now prove the claim. First, by (13), we have that $f(\emptyset) = g(\emptyset) = 0$. We show that f is monotonic. So let $A \subseteq B \subseteq \{X_1, \dots, X_n\}$. If $X_n \notin B$ then $X_n \notin A$ and we have that:

$$f(B) - f(A) = g(B) - g(A) \underbrace{\geq}_{\substack{B \supseteq A \\ g \in \Gamma_{n-1}}} 0$$

If $X_n \in B \setminus A$ then we let $B = B'X_n$, and we have:

$$f(B'X_n) - f(A) \underbrace{=}_{(13)} g(B') - g(A) \underbrace{\geq}_{B' \supseteq A} 0$$

Finally, if $X_n \in A \subseteq B$, then by letting $B = B'X_n$, $A = A'X_n$, we have that:

$$f(B'X_n) - f(A'X_n) \underbrace{=}_{(13)} g(B') - g(A') \geq 0$$

We now show that f is submodular. Let $A, B \subseteq \{X_1, \dots, X_n\}$. If $X_n \notin A \cup B$ then $f(Y) = g(Y)$ for every set $Y \in \{A, B, A \cup B, A \cap B\}$. Since g is submodular, then $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$. If $X_n \in A \setminus B$ then we write $A = A'X_n$ and observe that, by (13): $f(A'X_n) = g(A')$, $f(A \cup B) = f(A'X_n \cup B) = g(A' \cup B)$, that $f(B) = g(B)$, and that $f(A \cap B) = f(A' \cap B) = g(A' \cap B)$. Hence: $f(A) + f(B) = g(A') + g(B) \geq g(A' \cup B) + g(A' \cap B)$. The case where $X_n \in B \setminus A$ is symmetrical. Finally, if $X_n \in A \cap B$ then $X_n \in Y$ for all $Y \in \{A, B, A \cap B, A \cup B\}$. Hence, for every Y in this set, we write $Y = Y'X_n$. In particular, by (13) we have that $f(Y) = f(Y'X_n) = g(Y')$, and the claim follows since $g \in \Gamma_n$. □

Lemma 5.6. *Let $\tau = (A; B|C)$. If $\Gamma_n \models_{EI} \Sigma \Rightarrow \tau$ 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 $A = a_1 \dots a_m$, $B = b_1 \dots b_\ell$, $C = c_1 \dots c_k$, and $U = \Omega \setminus ABC$. Following [12], we construct the parity distribution $P(\Omega)$ as follows. We let all the RVs, except a_1, be independent binary RVs with probability $\frac{1}{2}$ for each of their two values, and let a_1 be determined from $ABC \setminus \{a_1\}$ as follows:

$$a_1 = \sum_{i=2}^{m} a_i + \sum_{i=1}^{\ell} b_i + \sum_{i=1}^{k} c_i \pmod 2 \tag{14}$$

Let $D \subseteq \Omega$ and $\boldsymbol{d} \in \mathcal{D}(D)$. We denote by $D_{ABC} = D \cap ABC$, 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 pairwise 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 $P(D = \boldsymbol{d}) = \left(\frac{1}{2}\right)^{|D|}$. Hence, the RVs in D are pairwise independent. If $a_1 \in D$, then since $D_{ABC} \subsetneq ABC$ it holds that $P(a_1 | D_{ABC} \setminus \{a_1\}) = P(a_1)$. To see this, observe that:

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

because if, w.l.o.g, $\sum_{y \in D_{ABC} \setminus \{a_1\}} y \pmod 2{=}0$, then $a_1{=}1$ implies that $\sum_{y \in ABC \setminus D} y \pmod 2{=}1$, and this is the case for precisely half of the assignments $ABC \setminus D {\rightarrow} \{0,1\}^{|ABC \setminus D|}$. 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 are pairwise 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 pairwise independent then $H(D) = \sum_{y \in D} H(y) = |D|$[2]. Furthermore, for any $(X;Y|Z) \in \Sigma$ s.t. $XYZ \not\subseteq ABC$ 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 $A' \stackrel{\text{def}}{=} A \setminus \{a_1\}$, then by chain rule for entropies, and noting that, by (14), $ABC \setminus a_1 \rightarrow a_1$, then:

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

and thus

$$\begin{aligned} I(A;B|C) &= H(AC) + H(BC) - H(C) - H(ABC) \\ &= |AC| + |BC| - |C| - (|ABC| - 1) \\ &= 1 \end{aligned} \quad (15)$$

In other words, the parity distribution P of (14) 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 $X_{ABC} \stackrel{\text{def}}{=} X \cap ABC$ and $Z_{ABC} = Z \cap ABC$. Therefore, we can write $I(X;Y|Z)$ as $I(X_{ABC}X';Y|Z_{ABC}Z')$ where $X' = X \setminus X_{ABC}$ and $Z' = Z \setminus Z_{ABC}$. It is easily shown that if $ABC \subseteq X$ or $ABC \subseteq Z$ then $I(X;Y|Z) = 0$. Otherwise (i.e., $X_{ABC} \neq \emptyset$ and $Z_{ABC} \neq \emptyset$), then due to the properties of the parity function, we have that $H(YZ'Z_{ABC}) = H(Y) + H(Z') + H(Z_{ABC})$. Noting that $X_{ABC}Z_{ABC} = ABC$, we get that $I(X_{ABC}X';Y|Z_{ABC}Z') = 0$.

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 (15)) where h_P is the entropic function associated with the parity function P in (14). Therefore, there must be a triple $\sigma \in \Sigma$ that meets the conditions of the lemma. Otherwise, we arrive at a contradiction to the EI. $\square$

[2]This is due to the chain rule of entropy, and the fact that if X and Y are independent RVs then $H(Y|X) = H(Y)$.

6 APPROXIMATE IMPLICATION FOR RECURSIVE CIS

We prove Theorem 4.1. Let P be a multivariate distribution over $\Omega{=}\{X_1, \dots, X_n\}$, and Σ be a recursive set (see (8)). We prove Theorem 4.1 by induction on the highest RV-index mentioned in any triple of Σ.

The claim trivially 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.

We recall that $\Sigma = \{\sigma_1, \dots, \sigma_n\}$ where $\sigma_i = (X_i; R_i|B_i)$ where $R_iB_i = \{X_1, \dots, X_{i-1}\}$. In particular, only $\sigma_n = (X_n; R_n|B_n)$ mentions the RV X_n, and it is saturated (i.e., $X_nR_nB_n = \Omega$). We denote by $\Sigma' = \Sigma \setminus \{\sigma_n\}$, and note that $X_n \notin \mathbf{var}(\Sigma')$. The induction hypothesis states that:

$$\Delta_n \models_{EI} \Sigma' \Rightarrow \tau \quad \text{ iff } \quad \Gamma_n \models h(\Sigma') \geq h(\tau) \quad (16)$$

Equivalently, by Lemma 5.1, and due to the one-to-one correspondence between Shannon's information measures and μ^* (Theorem 2.2), we can state the induction hypothesis:

$$\mathrm{m}(\Sigma') \supseteq \mathrm{m}(\tau) \quad \text{ iff } \quad \mu^*(\mathrm{m}(\Sigma')) \geq \mu^*(\mathrm{m}(\tau)) \quad (17)$$

Now, we consider $\tau = (X;Y|Z)$. We divide to three cases, and treat each one separately.

1. $X_n \notin XYZ$
2. $X_n \in Z$
3. $X_n \in X$ (or, symmetrically, $X_n \in Y$)

Case 1: $X_n \notin XYZ$**.** We will show that $\Delta_n \models_{EI} \Sigma' \Rightarrow \tau$, and the claim will follow from the induction hypothesis (16) because Σ' does not mention X_n, and $h(\Sigma) \geq h(\Sigma') \geq h(\tau)$ as required.

Suppose, by way of contradiction, that $\Delta_n \models_{EI} \Sigma' \not\Rightarrow \tau$. Since neither Σ' nor τ mention X_n then, by Lemma 5.5, we have that $\Delta_{n-1} \models_{EI} \Sigma' \not\Rightarrow \tau$. Hence, by Lemma 5.1, we have that $\mathrm{m}(\Sigma') \not\supseteq \mathrm{m}(\tau)$, and there exists an atom $a \in \mathcal{F}_{n-1}$ such that $a \in \mathrm{m}(\tau) \setminus \mathrm{m}(\Sigma')$. Consequently, there exist two atoms $a_1, a_2 \in \mathcal{F}_n$ where:

$$a_1 \stackrel{\text{def}}{=} a \cap \mathrm{m}(X_n) \qquad a_2 \stackrel{\text{def}}{=} a \cap \mathrm{m}^c(X_n)$$

such that $\{a_1, a_2\} {\subseteq} \mathrm{m}(\tau)$ and $\{a_1, a_2\} \cap \mathrm{m}(\Sigma') = \emptyset$. By our observation, $\sigma_n{=}(X_n; R|B)$. Therefore, we have that $\mathrm{m}(\sigma_n) \subseteq \mathrm{m}(X_n)$ (i.e., see Table 1). So, we get that $a_2 {\notin} \mathrm{m}(\sigma_n)$. Overall, we have that $a_2 {\notin} \mathrm{m}(\sigma_n) {\cup} \mathrm{m}(\Sigma') {=} \mathrm{m}(\Sigma)$, and by Lemma 5.1, we get that $\Delta_n \models_{EI} \Sigma \not\Rightarrow \tau$, a contradiction.

Case 2: $\tau = (W;Y|ZX_n)$**.** Then $\mathrm{m}(\tau)\subseteq\mathrm{m}^c(X_n)$, and since σ_n has the form $\sigma_n=(X_n;R|B)$, then $\mathrm{m}(\sigma_n)\subseteq\mathrm{m}(X_n)$ (see Table 1). Hence, $\mathrm{m}(\tau)\cap\mathrm{m}(\sigma_n)=\emptyset$, and by Lemma 5.3, if $\Delta_n\models_{EI}\Sigma\Rightarrow\tau$ then it must hold that $\Delta_n\models_{EI}\Sigma'\Rightarrow\tau$, and the claim follows from the induction hypothesis (16) because Σ' does not mention X_n, and $h(\Sigma)\geq(\Sigma')\geq h(\tau)$.

Case 3: $\tau=(WX_n;Y|Z)$**.** By the chain rule (see (3)):

$$(WX_n;Y|Z)=\underbrace{(W;Y|Z)}_{\tau_1}+\underbrace{(X_n;Y|WZ)}_{\tau_2} \tag{18}$$

Hence, if $\Delta_n{\models_{EI}}\Sigma{\Rightarrow}\tau$ then $\Delta_n{\models_{EI}}\Sigma{\Rightarrow}\tau_1$, and $\Delta_n{\models_{EI}}\Sigma{\Rightarrow}\tau_2$. We have already shown, in case 1, that the former implies $\Delta_n{\models_{EI}}\Sigma'{\Rightarrow}\tau_1$.

Let $\sigma_n=(X_n;R|B)$, and let $Y=Y_1\dots Y_m$ where $m\geq 1$. We claim that $Y\subseteq R$. Since σ_n is saturated then $Y\subseteq RB$. Now, suppose by way of contradiction, that $Y_i\in B$ for some $i\in[1,m]$. Consider the atom

$$a\stackrel{\text{def}}{=}\mathrm{m}(X_n)\cap\mathrm{m}(Y_i)\bigcap_{X\in[n]\setminus\{X_n,Y_i\}}\mathrm{m}^c(X).$$

We observe that $a\in\mathrm{m}(\tau_2)$. Since, by our assumption $Y_i{\in}B$, then $a\notin\mathrm{m}(\sigma_n)$. On the other hand, for every $\sigma=(A;B|C)\in\Sigma'$, we also have that $a\notin\mathrm{m}(\sigma)$. To see why, note that $X_n\notin AB$. Therefore, every atom of $\mathrm{m}(\sigma)$ contains at least two sets in positive form: $\mathrm{m}(X_i)$ for some $X_i\in A$ and $\mathrm{m}(X_j)$ for some $X_j\in B$. Since neither of these are X_n, then at least one of them appears in negative form in a. Overall, we get that $\mathrm{m}(X_n;Y|WZ)\not\subseteq\mathrm{m}(\Sigma)$, and by Lemma 5.1 that $\Delta_n\models_{EI}\Sigma\not\Rightarrow\tau_2$. Hence, from (18), we get that $\Delta_n\models_{EI}\Sigma\not\Rightarrow\tau$, a contradiction.

Since $Y{\subseteq}R$, we can write $\sigma_n{=}(X_n;YR_WR_ZR'|B_WB_ZB')$ where $R_W\stackrel{\text{def}}{=}R\cap W$, $R_Z\stackrel{\text{def}}{=}R\cap Z$, and $R'\stackrel{\text{def}}{=}R\setminus R_WR_ZY$. Likewise, $B_W{\stackrel{\text{def}}{=}}B{\cap}W$, $B_Z{\stackrel{\text{def}}{=}}B{\cap}Z$, and $B'{\stackrel{\text{def}}{=}}B\setminus B_WB_Z$. Further, since σ_n is saturated then $W=R_WB_W$ and $Z=R_ZB_Z$. By the chain rule, we have that:

$$\begin{aligned}h(\sigma_n)&=I_h(X_n;YR_WR_ZR'|B_WB_ZB')\\&=I_h(X_n;YR_WR_Z|B_WB_ZB')+I_h(X_n;R'|WZYB')\\&\geq I_h(X_n;R_WR_Z|B_WB_ZB')+I_h(X_n;Y|WZB')\\&\geq I_h(X_n;Y|ZWB')\end{aligned} \tag{19}$$

Now, if $B'=\emptyset$ then we are done because $h(\sigma_n)\geq h(X_n;Y|ZW)=h(\tau_2)$ and by the induction hypothesis if $\Delta_n\models\Sigma'\Rightarrow\tau_1$ then $h(\Sigma')\geq h(\tau_1)$. So assume that $B'\neq\emptyset$, and consider the following set of atoms:

$$A\stackrel{\text{def}}{=}\mathrm{m}(X_n)\cap\left(\bigcup_{y\in Y}\mathrm{m}(y)\right)\cap\left(\bigcup_{b\in B'}\mathrm{m}(b)\right)\cap\bigcap_{X\in ZW}\mathrm{m}^c(X)$$

We note that $\mathrm{m}(\tau_2)\supseteq A$. By our assumption that $\sigma_n=(X_n;R|B_WB_ZB')$, then $A\cap\mathrm{m}(\sigma_n)=\emptyset$. Since $\Delta_n\models\Sigma\Rightarrow\tau_2$ then by Lemma 5.3, it must hold that $\mathrm{m}(\Sigma')\supseteq A$. Furthermore, since $X_n\notin\mathbf{var}(\Sigma')$ then it must hold that $\mathrm{m}(\Sigma')\supseteq A'$ where:

$$A'\stackrel{\text{def}}{=}\left(\bigcup_{y\in Y}\mathrm{m}(y)\right)\cap\left(\bigcup_{b\in B'}\mathrm{m}(b)\right)\cap\bigcap_{X\in ZW}\mathrm{m}^c(X)$$

Denote by $\tau_3\stackrel{\text{def}}{=}(Y;B'|ZW)$, and hence $\mathrm{m}(\tau_3)=A'$ (see Table 1). In particular, $\mathrm{m}(\Sigma')\supseteq\mathrm{m}(\tau_3)$, and by Lemma 5.1 we have that $\Delta_n\models\Sigma'\Rightarrow\tau_3$. Since neither Σ' nor τ_3 mention X_n, then by the induction hypothesis (17), we have that $\mu^*(\mathrm{m}(\Sigma'))\geq\mu^*(\mathrm{m}(\tau_3))$.

Since $\Delta_n\models\Sigma\Rightarrow\tau_1$, and since $X_n\notin\mathbf{var}(\tau_1)$, then by the argument of case 1 we have that $\Delta_n\models\Sigma'\Rightarrow\tau_1$, and hence by Lemma 5.1 that $\mathrm{m}(\Sigma')\supseteq\mathrm{m}(\tau_1)$. Now, by the previous reasoning, we have also have that $\mathrm{m}(\Sigma')\supseteq\mathrm{m}(\tau_3)$. By noting that $\mathrm{m}(\tau_1)\cap\mathrm{m}(\tau_3)=\emptyset$, and applying Lemma 5.3, we get that $\mathrm{m}(\Sigma')\setminus\mathrm{m}(\tau_3)\supseteq\mathrm{m}(\tau_1)$. Applying the induction hypothesis (17), we get that $\mu^*\left(\mathrm{m}(\Sigma')\setminus\mathrm{m}(\tau_3)\right)\geq\mu^*(\mathrm{m}(\tau_1))$. Now, since μ^* is set-additive, and $\mathrm{m}(\Sigma')\supseteq\mathrm{m}(\tau_3)$, we get that $\mu^*(\mathrm{m}(\Sigma'))-\mu^*(\mathrm{m}(\tau_3))\geq\mu^*(\mathrm{m}(\tau_1))$. And, by the one-to-one correspondence between Shannon's information measures and the I-measure (Theorem 2.2), we get that $h(\Sigma')-h(\tau_3)\geq h(\tau_1)$.

Now, from (19) we have that $h(\sigma_n)\geq I_h(X_n;Y|ZWB')$. By applying the chain rule:

$$\underbrace{I_h(X_n;Y|ZWB')}_{\leq h(\sigma_n)}+\underbrace{I_h(Y;B'|ZW)}_{=h(\tau_3)}=I_h(B'X_n;Y|WZ)\geq h(\tau_2)$$

Overall, we get that:

$$I_h\underbrace{(W;Y|Z)}_{\tau_1}+I_h\underbrace{(X_n;Y|WZ)}_{\tau_2}\leq h(\Sigma')-h(\tau_3)+h(\tau_3)+h(\sigma_n)=h(\Sigma)$$

as required.

TIGHTNESS OF BOUND

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

$$\Sigma=\{(X_1;X_i|X_2\dots X_{i-1}):i\in\{2,\dots,n\}\} \tag{20}$$

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

$$I(X_1;X_2X_3\dots X_n)=\sum_{i=2}^{n}I(X_1;X_i|X_2\dots X_{i-1})=h(\Sigma) \tag{21}$$

Hence, $\Sigma\Rightarrow_{EI}\tau$, and the bound of (21) is tight.

7 APPROXIMATE IMPLICATION FOR MARGINAL CIS

In this section, we prove Theorem 4.2. Let Σ be a set of marginal mutual information terms, and let $\tau = (A;B|D)$ such that $\Gamma_n \models_{EI} \Sigma \Rightarrow \tau$. Then, by the chain rule (3), τ can be written as a sum of at most $|A||B|$ elemental CIs $(a;b|C)$. In Lemma 7.1 we show that for every such elemental triple $(a;b|C)$, there exists a marginal $(X;Y) \in \Sigma$ such that $XY \supseteq abC$, $a \in X$, and $b \in Y$. Consequently, from Lemma 5.4, we get that $h(\Sigma) \geq I(X;Y) \geq I(a;b|C)$. Hence, it follows from lemma 7.1 that $|A||B|h(\Sigma) \geq h(\tau)$, and this will complete the proof for Theorem 4.2.

Lemma 7.1. *Let Σ be a set of marginal mutual information terms, and let $\tau = (a;b|C)$ be an elemental mutual information term. The following holds:*

$$\Gamma_n \models_{EI} \Sigma \Rightarrow \tau \quad \textit{iff} \quad \begin{array}{c} \exists (X;Y) \in \Sigma : \\ XY \supseteq abC \text{ and } a \in X, b \in Y \end{array}$$

Proof. We prove by induction on $|C|$. When $|C| = 0$ then $\tau = (a;b)$. Consider the atom:

$$t \stackrel{\text{def}}{=} \mathrm{m}(a) \cap \mathrm{m}(b) \bigcap_{y \in \Omega \setminus ab} \mathrm{m}^c(y) \tag{22}$$

Clearly, $t \in \mathrm{m}(\tau)$. Suppose, by way of contradiction, that for every $\sigma = (X;Y) \in \Sigma$ it holds that $ab \cap X = \emptyset$ or $ab \cap Y = \emptyset$. If, without loss of generality, we assume the former then clearly $t \notin \mathrm{m}(\sigma)$ because all of the RVs in X appear in negative form in the atom t. If this is the case for all $\sigma \in \Sigma$, then $t \notin \mathrm{m}(\Sigma)$, and $\mathrm{m}(\tau) \not\subseteq \mathrm{m}(\Sigma)$. But then, by Corollary 5.2, it cannot be that $\Gamma_n \models_{EI} \Sigma \Rightarrow \tau$, a contradiction.

So, we assume correctness for elemental terms $(a;b|C)$ where $|C| \leq k-1$, and prove for $|C| = k$. Since $\Gamma_n \models_{EI} \Sigma \Rightarrow \tau$, then by Lemma 5.6 there exists a mutual information term $\sigma = (X;Y) \in \Sigma$ such that $XY \supseteq abC$. Hence, we denote $C = C_X C_Y$, where $C_X = X \cap C$ and $C_Y = Y \cap C$. There are two cases. If $\sigma = (aC_X X_0; bC_Y Y_0)$ then, by Lemma 5.4, we have that $h(\sigma) \geq h(\tau)$, and we are done.

Otherwise, w.l.o.g, $\sigma = (abC_X X_0; C_Y Y_0)$. By item 2 of Lemma 5.6, it holds that $C_Y \neq \emptyset$.

We define:

$$\alpha_1 \stackrel{\text{def}}{=} (a; C_Y | C_X) \qquad \alpha_2 \stackrel{\text{def}}{=} (a; C_Y | bC_X) \tag{23}$$

By Lemma 5.4, we have that $h(\sigma) \geq h(\alpha_1)$ and $h(\sigma) \geq h(\alpha_2)$, and thus $\Gamma_n \models_{EI} \Sigma \Rightarrow \{\alpha_1, \alpha_2\}$. Noting that $\tau = (a;b|C_X C_Y)$, we have that $\Gamma_n \models_{EI} \Sigma \Rightarrow (a;b|C_X C_Y)$. By the chain rule (see (3)) we have that Σ implies:

$$(a;C_Y|C_X), (a;b|C_X C_Y) \Rightarrow (a;bC_Y|C_X) \Rightarrow (a;b|C_X)$$

In other words, we have that $\Gamma_n \models_{EI} \Sigma \Rightarrow (a;b|C_X)$.

By item 2 of Lemma 5.6 it holds that $C_Y \neq \emptyset$. Hence, $C_X \subsetneq C$. Therefore, by the induction hypothesis, there exists an $\alpha_3 \stackrel{\text{def}}{=} (aC_X^1 Z_1; bC_X^2 Z_2) \in \Sigma$ where $C_X = C_X^1 C_X^2$. In particular, by Lemma 5.4, we have that $\alpha_3 \Rightarrow \alpha_4 \stackrel{\text{def}}{=} (a;b|C_X)$, and $h(\alpha_4) \leq h(\alpha_3)$ where $\alpha_3 \in \Sigma$. Furthermore, by our assumption (i.e., that $\sigma = (abC_X X_0; C_Y Y_0)$), then σ and α_3 are distinct. Consequently, we get that:

$$\begin{aligned} &\underbrace{I(a;b|C_X)}_{\leq h(\alpha_3)} + \underbrace{I(a;C_Y|bC_X)}_{\leq h(\sigma)} \\ &= I(a;bC_Y|C_X) \geq I(a;b|C_X C_Y) = h(\tau) \end{aligned} \tag{24}$$

Overall, we get that $h(\tau) \leq h(\alpha_3) + h(\sigma) \leq h(\Sigma)$ because $\alpha_3, \sigma \in \Sigma$ are distinct, by our assumption. This completes the proof. ☐

8 CONCLUSION AND DISCUSSION

We study the approximation variant of the well known implication problem, and showed that d-separation, the popular inference system used to derive CIs in Bayesian networks, continues to be sound and complete for inferring approximate CIs. We prove a tight approximation factor of 1 for the case of recursive CIs, and an approximation factor that depends on the size of the implicate for marginal CIs.

The question that remains is whether there are other classes of CIs that admit a λ-relaxation for a bounded λ. Previous work has shown that without making any assumptions on the antecedents or the inference system, the answer is negative [15], and when the inference system is the polymatroid inequalities (or equivalently, the semigraphoid axioms) then the bound is $(2^n)!$. Despite these negative results, when the set of antecedents fall into certain classes, then they *do* admit bounded relaxation. This is the case for saturated CIs [15], which are the foundation for undirected PGMs. It has been shown that the semigraphoid axioms are sound and complete for deriving constraints from saturated CIs [9]. The semigraphoid axioms are also sound and complete for sets of CIs whose cardinality is at most two [26], and for the *enhanced recursive set* which is a combination of CIs corresponding to a DAG along with functional dependencies [11]. We conjecture that these two sets of CIs also admit a bounded relaxation.

As part of future work we intend to empirically evaluate the extent to which our approach can be applied to the task of extracting the structure of PGMs from observational data. We intend to evaluate our approach along two measures. First, how close the learned model matches the empirical distribution induced by the observed data, and second, how it compares in terms of both accuracy and efficiency to constraint-based algorithms that perform statistical independence tests [5, 24].

References

[1] W. W. Armstrong and C. Delobel. Decomposition and functional dependencies in relations. *ACM Trans. Database Syst.*, 5(4):404–430, 1980. doi: 10.1145/320610.320620.

[2] C. Beeri, R. Fagin, and J. H. Howard. 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.*, pages 47–61, 1977. doi: 10.1145/509404.509414.

[3] X. Chen, G. Anantha, and X. Lin. 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, 2008. doi: 10.1109/TKDE.2007.190732.

[4] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. Learning bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1):43 – 90, 2002. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(02)00191-1.

[5] D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

[6] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979. ISSN 00359246.

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

[8] D. Geiger and J. Pearl. 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*, pages 3–14, 1988.

[9] D. Geiger and J. Pearl. Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics*, 21(4):2001–2021, 1993. ISSN 00905364.

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

[11] D. Geiger, T. Verma, and J. Pearl. Identifying independence in bayesian networks. *Networks*, 20(5):507–534, 1990. doi: 10.1002/net.3230200504.

[12] D. Geiger, A. Paz, and J. Pearl. Axioms and algorithms for inferences involving probabilistic independence. *Inf. Comput.*, 91(1):128–141, 1991. doi: 10.1016/0890-5401(91)90077-F.

[13] D. Geiger, A. Paz, and J. Pearl. Axioms and algorithms for inferences involving probabilistic independence. *Information and Computation*, 91(1):128 – 141, 1991. ISSN 0890-5401. doi: https://doi.org/10.1016/0890-5401(91)90077-F.

[14] C. Herrmann. On the undecidability of implications between embedded multivalued database dependencies. *Inf. Comput.*, 122(2):221–235, Nov. 1995. ISSN 0890-5401. doi: 10.1006/inco.1995.1148.

[15] B. Kenig and D. Suciu. Integrity constraints revisited: From exact to approximate implication. In C. Lutz and J. C. Jung, editors, *23rd International Conference on Database Theory, ICDT 2020, March 30-April 2, 2020, Copenhagen, Denmark*, volume 155 of *LIPIcs*, pages 18:1–18:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi: 10.4230/LIPIcs.ICDT.2020.18.

[16] B. Kenig, P. Mundra, G. Prasaad, B. Salimi, and D. Suciu. Mining approximate acyclic schemes from relations. In D. Maier, R. Pottinger, A. Doan, W. Tan, A. Alawini, and H. Q. Ngo, editors, *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, pages 297–312. ACM, 2020. doi: 10.1145/3318464.3380573.

[17] D. Koller and N. Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009. ISBN 978-0-262-01319-2.

[18] J. Kontinen, S. Link, and J. Väänänen. Independence in database relations. In L. Libkin, U. Kohlenbach, and R. de Queiroz, editors, *Logic, Language, Information, and Computation*, pages 179–193, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-39992-3.

[19] T. T. Lee. An information-theoretic analysis of relational databases - part I: data dependencies and information metric. *IEEE Trans. Software Eng.*, 13(10):1049–1061, 1987. doi: 10.1109/TSE.1987.232847.

[20] D. Maier. *Theory of Relational Databases*. Computer Science Pr, 1983. ISBN 0914894420.

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

[22] J. Pearl and A. Paz. 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*, pages 357–363, 1986.

[23] B. Sayrafi, D. Van Gucht, and M. Gyssens. The implication problem for measure-based constraints. *Information Systems*, 33(2):221 – 239, 2008. ISSN 0306-4379. doi: https://doi.org/10.1016/j.is.2007.07.005. Performance Evaluation of Data Management Systems.

[24] M. Scutari, C. E. Graafland, and J. M. Gutiérrez. Who learns better bayesian network structures: Constraint-based, score-based or hybrid algorithms? In M. Studený and V. Kratochvíl, editors, *International Conference on Probabilistic Graphical Models, PGM 2018, 11-14 September 2018, Prague, Czech Republic*, volume 72 of *Proceedings of Machine Learning Research*, pages 416–427. PMLR, 2018.

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

[26] M. Studený. Semigraphoids and structures of probabilistic conditional independence. *Ann. Math. Artif. Intell.*, 21(1):71–98, 1997. doi: 10.1023/A:1018905100242.

[27] T. VERMA and J. PEARL. Causal networks: Semantics and expressiveness. In R. D. SHACHTER, T. S. LEVITT, L. N. KANAL, and J. F. LEMMER, editors, *Uncertainty in Artificial Intelligence*, volume 9 of *Machine Intelligence and Pattern Recognition*, pages 69–76. North-Holland, 1990. doi: https://doi.org/10.1016/B978-0-444-88650-7.50011-1.

[28] R. W. Yeung. A new outlook of shannon's information measures. *IEEE Trans. Information Theory*, 37(3):466–474, 1991. doi: 10.1109/18.79902.

[29] R. W. Yeung. *Information Theory and Network Coding*. Springer Publishing Company, Incorporated, 1 edition, 2008. ISBN 0387792333, 9780387792330.

[30] J. Zhao, Y. Zhou, X. Zhang, and L. Chen. Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences*, 113(18):5130–5135, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1522586113.