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 $X \perp Y \mid Z$, 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.

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. 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, 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, 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.

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.

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, and the d-separation algorithm) can be used to compute approximate CIs.

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).
We also prove that every CI we use Information Theory. This is the natural semantics for 0. Therefore, if 1, then it is guaranteed that 2.1 CONDITIONAL INDEPENDENCE

We denote by 3 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 are sound and complete for deriving CI statements that are implied by saturated CIs, marginal CIs, and recursive CIs that are used in Bayesian networks. 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. Studeny proved that in the general case, when no assumptions are made on the antecedents, no finite axiomatization exists. 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, 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 15 admits an X1|Y|Z-relaxation. We conclude in Section 8.

2 PRELIMINARIES

We denote by n = {1, 2, ..., n}. If 15 = {X1, ..., Xn}, we denote a set of variables and 15, then we abbreviate the union 15 ∪ 15 with UV.

2.1 CONDITIONAL INDEPENDENCE

Recall that two discrete random variables X, Y are called independent if 10 for all outcomes x, y. Fix 15 = {X1, ..., Xn}, a set of n jointly distributed discrete random variables with finite domains 15, respectively; let p be the probability mass. For 15 ⊆ [n], denote by X1 the joint random variable (X1 : i ∈ 15) with domain 15 = ∏i∈15 15i. We write p ⊨ X1 if X1 is a functionally determined X1; in the special case that X1 is conditionally independent given X2, we write p ⊨ X1 ∩ X2 | Y.

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 ⊥ Y | Z if and only if I(X; Y | Z) = 0, where I is the conditional mutual information. Hence, an exact implication (EI) σ1, ..., σk ⇒ τ is an assertion of the form (h(σ1) = 0 ∧ ⋯ ∧ h(σk) = 0) ⇒ h(τ) = 0, where τ, σ1, σ2, ... are triples (X; Y | Z), and h is the conditional mutual information measure I(·; ·). An approximate implication (AI) is a linear inequality h(σ) ≤ λh(Σ), where h(Σ) = ∑k=1 h(σi), and λ ≥ 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(·; ·) ≥ 0 is a nonnegative measure. Therefore, if 0 ≤ h(τ) ≤ λh(Σ) for some λ ≥ 0, then h(Σ) = 0 ⇒ h(τ) = 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 = ∅.

We show that every conditional independence relation (X; Y | Z) read off a DAG by the d-separation algorithm, admits a 1-approximation. In other words, if Σ is the recursive basis of CIs used to build the Bayesian network, then it is guaranteed that I(X; Y | Z) ≤ ∑i∈Σ h(σi). Furthermore, we present a family of implications for which our 1-approximation is tight (i.e., I(X; Y | Z) = ∑i∈Σ h(σi)). We also prove that every CI (X; Y | Z) implied by a set of marginal CIs admits an |X| · |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 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.
An assertion $X \perp Y \mid Z$ is called a Conditional Independence statement, or a CI; this includes $Z \rightarrow Y$ as a special case. When $XYZ = \Omega$ we call it saturated, and when $Z = \emptyset$ we call it marginal. A set of CIs $\Sigma$ implies a CI $\tau$, in notation $\Sigma \Rightarrow \tau$, if every probability distribution that satisfies $\Sigma$ also satisfies $\tau$.

### 2.2 BACKGROUND ON INFORMATION THEORY

We adopt required notation from the literature on information theory [29]. For $n > 0$, we identify the functions $2^n \rightarrow \mathbb{R}$ 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

$$h(B \mid A) \overset{\text{def}}{=} h(AB) - h(A),$$

$$I_h(B; C \mid A) \overset{\text{def}}{=} h(AB) + h(AC) - h(ABC) - h(A).$$

Then, $\forall h \in \Gamma_n$, $I_h(B; C \mid A) \geq 0$ by submodularity, and $h(B \mid A) \geq 0$ by monotonicity. We say that $A$ functionally determines $B$, in notation $A \rightarrow B$ if $h(B \mid A) = 0$. The chain rule is the identity:

$$I_h(B; CD \mid A) = I_h(B; C \mid A) + I_h(B; D \mid AC).$$

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

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

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 $h(\alpha) \overset{\text{def}}{=} H(X_\alpha)$; $h$ 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 $h(B \mid A)$ and $I_h(B; C \mid A)$ are called the conditional entropy and conditional mutual information respectively. The conditional independence $p \models B \perp C \mid A$

<table>
<thead>
<tr>
<th>Information Measures</th>
<th>$\mu^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H(X)$</td>
<td>$\mu^*(m(X))$</td>
</tr>
<tr>
<td>$H(XY)$</td>
<td>$\mu^*(m(X) \cup m(Y))$</td>
</tr>
<tr>
<td>$H(X</td>
<td>Y)$</td>
</tr>
<tr>
<td>$I_h(X; Y)$</td>
<td>$\mu^*(m(X) \cap m(Y))$</td>
</tr>
<tr>
<td>$I_h(X; Y</td>
<td>Z)$</td>
</tr>
</tbody>
</table>

Table 1: Information measures and associated I-measure

holds iff $I_h(B; C \mid A) = 0$, and similarly $p \models A \rightarrow B$ iff $h(B \mid A) = 0$; thus, entropy provides us with an alternative characterization of CIs.

#### 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 $m(X_i)$, and it’s complement $m^c(X_i)$. The universal set is $\Lambda = \bigcup_{i=1}^n m(X_i)$. Let $\alpha \subseteq [n]$. We denote by $X_{\alpha} \overset{\text{def}}{=} \{X_j \mid j \in \alpha\}$, and $m(X_{\alpha}) = \bigcup_{\alpha \subseteq [n]} m(X_i)$.

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

The atoms of $\mathcal{F}_n$ are sets of the form $\bigcap_{i=1}^n Y_i$, where $Y_i$ is either $m(X_i)$ or $m^c(X_i)$. 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 $\bigcap_{i=1}^n m^c(X_i) = \emptyset$ is empty). There are $2^n - 1$ non-empty atoms and $2^{n-1}$ sets in $\mathcal{F}_n$ expressed as the union of its atoms. A function $\mu : \mathcal{F}_n \rightarrow \mathbb{R}$ 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 $\mu$ defined on $\mathcal{F}_n$ is called a signed measure if it is set additive, and $\mu(\emptyset) = 0$.

The $I$-measure $\mu^*$ 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 $\mu^*$.

**Theorem 2.2.** [28, 29] [I-Measure Theorem] $\mu^*$ 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 $m(\sigma) = m(X) \cap m(Y) \cap m^c(Z)$ the set associated with $\sigma$ (see Table 1). For a set of

\[ \text{Recall that } AB \text{ denotes } A \cup B. \]
triples $\Sigma$, we define:

$$m(\Sigma) \triangleq \bigcup_{\sigma \in \Sigma} m(\sigma)$$

(5)

**Example 2.3.** Let $A$, $B$, and $C$ be three disjoint sets of RVs defined as follows: $A = A_1 A_2 A_3$, $B = B_1 B_2 B_3$, and $C = C_1 C_2 C_3$. Then, by Theorem 2.2,

$$H(A) = \mu^*(m(A)) = \mu^*(m(A_1) \cdot m(A_2) \cdot m(A_3)),$$

$$H(B) = \mu^*(m(B)) = \mu^*(m(B_1) \cdot m(B_2) \cdot m(B_3)).$$

By Table 7, $I(A; B|C) = \mu^*(m(A) \cap m(B) \cap m(C))$.

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

**Theorem 2.4.** (29) If there is no constraint on $X_1, \ldots, X_n$, then $\mu^*$ can take any set of non-negative values on the nonempty atoms of $F_n$.

Theorem 2.4 implies that every positive I-measure $\mu^*$ 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 $\alpha$ correspond to the set of variables $X_{\alpha}$, 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 $p(X_i | X_{\pi(i)})$ 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

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

(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:

$$\Sigma \triangleq \{(X_i; X_1 \ldots X_{i-1} | \pi(X_i) \mid \pi(X_i)) : i \in [n]\}$$

(7)

The implication problem associated with Bayesian Networks is to determine whether $\Sigma \Rightarrow \tau$ for a CI $\tau$. Geiger and Pearl have shown that $\Sigma \Rightarrow \tau$ iff $\tau$ can be derived from $\Sigma$ using the semigraphoid axioms. 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 $\Sigma$ is a set of CIs called antecedents and $\tau$ is a CI called consequent. For a CI $\sigma = (Y; Z | X)$, we define $h(\sigma) \triangleq I_h(Y; Z | X)$, for a set of CIs $\Sigma$, we define $h(\Sigma) \triangleq \sum_{\sigma \in \Sigma} h(\sigma)$. Fix a set $K$ s.t. $K \subseteq \Gamma_n$. We call $\Sigma$ 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 $\lambda$-approximate implication (AI) holds in $K$, in notation $K \models_{AI} \lambda \cdot h(\Sigma) \geq h(\tau)$, if $\forall h \in K$, $h \cdot h(\Sigma) \geq h(\tau)$. The approximate implication holds, in notation $K \models_{AI} \lambda \cdot h(\Sigma) \Rightarrow h(\tau)$, if there exist a finite $\lambda \geq 0$ such that the $\lambda$-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 $\tau$, and every polymatroid $h$. In this paper we study the reverse.

**Definition 3.1.** Let $L$ be a syntactically-defined class of implication statements ($\Sigma \Rightarrow \tau$), and let $K \subseteq \Gamma_n$. We say that $\Sigma$ admits a $\lambda$-relaxation in $K$, if every exact implication statement ($\Sigma \Rightarrow \tau$) in $L$ has a $\lambda$-approximation:

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

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

**Example 3.3.** Let $\Sigma = \{(A; B|\emptyset), (A; C|B)\}$, and $\tau = (A; C|\emptyset)$. Since $I_h(A; C|\emptyset) \leq I_h(A; B|C)$, and since $I_h(A; B|C) = I_h(A; B|\emptyset) + I_h(A; C|B)$ by the chain rule, 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 CTS

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 \( \Sigma \) from (7):
\[
\Sigma \triangleq \{(X_i; R_i|B_i) : i \in [1, n], R_i|B_i = U^{(i)}\}
\]
(8) where \( B_i \triangleq \pi(X_i) \) and \( U^{(i)} \triangleq \{X_1, \ldots, X_{i-1}\} \). 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 \( \sigma_n \) 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 \( F_n \) (see Section 4.1).

**Theorem 4.1.** Let \( \Sigma \) be a recursive set of CIs (see (8)), and let \( \tau = (A; B|C) \). Then the following holds:
\[
\Delta_n \models EI \Sigma \Rightarrow \tau \iff \Gamma_n \models h(\Sigma) \geq h(\tau)
\]
(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 \( \Delta_n \), 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 \( \Sigma \) be a set of marginal CIs, and \( \tau = (A; B|C) \) be any CI.
\[
\Gamma_n \models EI \Sigma \Rightarrow \tau \iff \Gamma_n \models (|A|\cdot|B|)h(\Sigma) \geq h(\tau)
\]
(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 \( \Delta_n \) (Section 5.1), and the entire set of polymatroids \( \Gamma_n \) (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, \( \Sigma \) denotes a set of triples \( (A; B|C) \) representing mutual information terms, and \( \tau \) denotes a single triple. We denote by \( \var(m) \) the set of RVs mentioned in \( \sigma \) (e.g., if \( \sigma = (X_1X_2; X_3X_5) \) then \( \var(m) = X_1 \ldots X_5 \)).

#### 5.1 EXACT IMPLICATION IN THE SET OF POSITIVE POLYMATROIDS

**Lemma 5.1.** The following holds:
\[
\Delta_n \models EI \Sigma \Rightarrow \tau \iff m(\Sigma) \supseteq m(\tau)
\]

**Proof.** Suppose that \( m(\tau) \subseteq m(\Sigma) \), and let \( b \in m(\tau) \setminus m(\Sigma) \). By Theorem 2.4, there exists a positive polymatroid in \( \Delta_n \) with an \( I \)-measure \( \mu \) that takes the following non-negative values on its atoms: \( \mu^*(b) = 1 \), and \( \mu^*(a) = 0 \) for any atom \( a \in F_n \) where \( a \neq b \). Since \( b \notin m(\Sigma) \), then \( \mu^*(\Sigma) = 0 \) while \( \mu^*(\tau) = 1 \). Hence, \( \Delta_n \not \models \tau \).

Now, suppose that \( m(\Sigma) \supseteq m(\tau) \). Then for any positive I-measure \( \mu : F_n \rightarrow \mathbb{R}_{\geq 0} \), we have that \( \mu^*(m(\Sigma)) \geq \mu^*(m(\tau)) \). By Theorem 2.2, \( \mu^* \) is the unique signed measure on \( F_n \) 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 \( m(\Sigma) \supseteq m(\tau) \) is a necessary condition for implication between polymatroids.

**Corollary 5.2.** If \( \Gamma_n \models EI \Sigma \Rightarrow \tau \) then \( m(\Sigma) \supseteq m(\tau) \).

**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 \( m(\sigma) \cap m(\tau) = \emptyset \). 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 \( m(\Sigma') \not \supseteq m(\tau) \). In other words, there is an atom \( a \in F_n \) such that \( a \in m(\tau) \setminus m(\Sigma') \). In particular, \( a \notin m(\sigma) \cup m(\Sigma') = m(\Sigma) \). Hence, \( m(\tau) \not \subseteq m(\Sigma) \), 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, $\Gamma_n \models h(\tau) \iff h(\sigma)$. 

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:

\[ I(Z_A A'X; Z_B B'Y|C) = I(Z_A; Z_B|C) + I(A'; X; Z_B|CZ_A) \]

\[ + I(Z_A; B'|Z_B|C) + I(X; Y'|ZCAZB) \]

\[ + I(X; B'|CZAZB) + I(A'; B'|CZA) \]

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

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

Then:

\[ \Gamma_n \models E1 \Sigma \models \tau \iff \Gamma_{n-1} \models E1 \Sigma \models \tau \quad (11) \]

Proof. Suppose that $\Gamma_n \nvDash E1 \Sigma \models \tau$. 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:

\[ g(A) = f(A) \quad \text{for all} \quad A \subseteq \{X_1, \ldots, X_{n-1}\} \quad (12) \]

Since $f$ is a polymatroid, then so is $g$. Further, since $\Sigma$ does not mention $X_n$, then, by (12), we have that $g(\sigma) = f(\sigma)$ for all $\sigma \in \Sigma$. Hence, $\Gamma_{n-1} \nvDash E1 \Sigma \models \tau$.

If $\Gamma_{n-1} \nvDash E1 \Sigma \models \tau$. Then there exists a polymatroid $g : 2^{[n-1]} \to \mathbb{R}$ such that $g(\tau) = 0$ for all $\tau \in \Sigma$, and $g(\sigma) \neq 0$. Define $f : 2^{[n]} \to \mathbb{R}$ as follows:

\[ f(A) = g(A \setminus X_n) \quad \text{for all} \quad A \subseteq \{X_1, \ldots, X_n\} \quad (13) \]

We claim that $f \in \Gamma_n$ (i.e., $f$ is a polymatroid). It then follows that $\Gamma_n \nvDash E1 \sum \models \tau$ because by the assumption that $\text{var}(\Sigma)$ and $\text{var}(\tau)$ are subsets of $\{X_1, \ldots, X_{n-1}\}$, then $f(\sigma) = g(\sigma)$ for all $\sigma \in \Sigma$. Hence, $f(\Sigma) = g(\Sigma) = 0$ while $f(\tau) = f(\gamma) \neq 0$.

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

\[ f(B) - f(A) = g(B) - g(A) \geq 0 \quad \text{for all} \quad A \subseteq B \subseteq \{X_1, \ldots, X_n\} \quad (14) \]

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

\[ f(B'X_n) - f(A) = g(B') - g(A) \geq 0 \quad \text{for all} \quad A \subseteq B \subseteq \{X_1, \ldots, X_n\} \quad (15) \]

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

\[ f(B'X_n) - f(A'X_n) = g(B') - g(A') \geq 0 \]

We now show that $f$ is submodular. Let $A, B \subseteq \{X_1, \ldots, X_n\}$. If $X_n \notin A \cup B$ then $f(Y) = g(Y)$ for every set $Y \subseteq \{A, B, A \cup B, A \cap B\}$. Since $g$ is submodular, then $f(A) + f(B) = f(A\cup B) + f(A \cap B)$. If $X_n \in A \setminus B$ then we write $A = A'X_n$ and observe that, by (13), $f(A'X_n) = g(A')$. Hence:

\[ f(A) + f(B) = f(A') + g(B) \geq g(A') \]

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 $Y \subseteq \{A, B, A \cap B, A \cup B\}$, and 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 $\Gamma_n \models E1 \Sigma \models \tau$ 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 $A = a_1 \ldots a_m$, $B = b_1 \ldots b_n$, $C = c_1 \ldots c_k$, 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:

\[ a_1 = \begin{cases} m \sum_{i=1}^m a_i + \sum_{i=1}^n b_i + \sum_{i=1}^k c_i \pmod{2} & (14) \end{cases} \]

Let $D \subseteq \Omega$ and $d \in D(D)$. We denote by $D_{ABC} = D \cap ABC$, and by $d_{ABC}$ the assignment $d$ restricted to the RVs $D_{ABC}$. We show that if $D_{ABC} \subseteq ABC$ then the RVs in $D$ are pairwise independent. By the definition of $P$ we have that:

\[ P(D = d) = \left( \frac{1}{2} \right)^{|D \cup U|} P(D_{ABC} = d_{ABC}) \]

There are two cases with respect to $D$. If $a_1 \notin D$ then, by definition, $P(D_{ABC} = d_{ABC}) = \left( \frac{1}{2} \right)^{|D_{ABC}|}$, and overall we get that $P(D = d) = \left( \frac{1}{2} \right)^{|D|}$. Hence, the RVs in $D$ are pairwise independent. If $a_1 \in D$, then since $D_{ABC} \subseteq ABC$ it holds that $P(a_1 \mid D_{ABC} \setminus \{a_1\}) = P(a_1)$. To see this, observe that:

\[ P(a_1 = 1 \mid D_{ABC} \setminus \{a_1\}) = \begin{cases} \frac{1}{2} & \text{if } \sum_{\gamma \in D_{ABC} \setminus \{a_1\}} y \pmod{2} = 0 \\ \frac{1}{2} & \text{if } \sum_{\gamma \in D_{ABC} \setminus \{a_1\}} y \pmod{2} = 1 \end{cases} \]
because if, w.l.o.g., \( \sum_{y \in D \setminus \{a_1\}} y \mod 2 = 0 \), then \( a_1 = 1 \) implies that \( \sum_{y \in AB \cap D \setminus \{a_1\}} y \mod 2 = 1 \), and this is the case for precisely half of the assignments \( ABC \setminus D \rightarrow \{0, 1\}^{D \setminus D} \). Hence, for any \( D \subseteq \Omega \) such that \( D \cap \var{\tau} \subseteq AB \) it holds that \( P(D=\sigma) = \prod_{y \in D} P(y \in \sigma) = \left( \frac{1}{2} \right)^{|D|} \), 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 \( \Omega \). Since the RVs in \( D \) are pairwise independent then \( H(D) = \sum_{y \in D \setminus \{a_1\}} H(y) = |D| \). Further more, for any \((X;Y|Z) \in \Sigma \) s.t. \( XYZ \not\subseteq ABC \) we have that:

\[
I(X;Y|Z) = H(XZ) + H(YZ) - H(Z) - H(XYZ) = |XZ| + |YZ| - |Z| - |XYZ| = |X| + |Y| + |Z| - |XYZ| = 0
\]

On the other hand, letting \( A' = A \setminus \{a_1\} \), then by rule chain for entropies, and noting that, by \([14]\), \( ABC \setminus a_1 \to a_1 \), then:

\[
H(\var{\tau}) = H(ABC) = H(a_1A'BC) = H(a_1) + |A'BC| = 0 + |ABC| - 1 = |ABC| - 1,
\]

and thus

\[
I(A;B|C) = H(AC) + H(BC) - H(C) - H(ABC) = |AC| + |BC| - |C| - |ABC| - 1 \tag{15}
\]

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 \( \var{\sigma} \not\subseteq ABC \), while \( h_P(\tau) = 1 \). Hence, if \( \Gamma_n \models \Sigma \models \tau \), then there must be a triple \( \sigma = (X;Y|Z) \in \Sigma \) such that \( XYZ \subseteq ABC \).

Now, suppose that \( ABC \subseteq XYZ \) and that \( ABC \cap Y = \emptyset \).

In other words, \( ABC \subseteq XYZ \). We denote \( X_{AB \cap X} \), \( Z_{AB \cap Z} \), \( X_{AB \cap X} \), \( Y_{AB \cap Y} \), \( Z_{AB \cap Z} \). Hence, we can write \( I(X;Y|Z) = I(X_{AB \cap X}Y_{AB \cap Z}|Z_{AB \cap Z}) \) where \( X' = X \setminus X_{AB \cap X} \) and \( Z' = Z \setminus Z_{AB \cap Z} \). It is easily shown that if \( ABC \subseteq X \) or \( ABC \subseteq Z \) then \( I(X;Y|Z) = 0 \). Otherwise (i.e., \( X_{AB \cap X} = \emptyset \) and \( Z_{AB \cap Z} = \emptyset \)), then due to the properties of the parity function, we have that \( H(X'Z') = H(Y) + H(Z') \). Noting that \( X_{AB \cap X}Z_{AB \cap X} = ABC \), we get that \( I(X;Y|Z) = 0 \).

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 [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.

\[ \square \]

### 6 APPROXIMATE IMPLICATION FOR RECURSIVE CIS

We prove Theorem 4.1. Let \( P \) be a multivariate distribution over \( \Omega = \{X_1, \ldots, X_n\} \), and \( \Sigma \) be a recursive set (see [8]).

We prove Theorem 4.1 by induction on the highest RV-index mentioned in any triple of \( \Sigma \).

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 \( \Sigma \) is \( \leq n - 1 \), and prove for \( n \).

We recall that \( \Sigma = \{\sigma_1, \ldots, \sigma_n\} \) where \( \sigma_i = (X_i; R_i|B_i) \) where \( R_i \equiv B_i \equiv \{X_1, \ldots, X_{i-1}\} \). In particular, only \( \sigma_n = (X_n; R_n|B_n) \) mentions the RV \( X_n \), and it is saturated (i.e., \( X_n R_n B_n = \Omega \)). We denote by \( \Sigma' = \Sigma \setminus \{\sigma_n\} \), and note that \( X_n \not\in \var{\Sigma'} \). The induction hypothesis states that:

\[
\Delta_n \models \Sigma' \models \Rightarrow \text{if } \Gamma_n \models h(\Sigma') \geq h(\tau) \tag{16}
\]

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

\[
m(\Sigma') \geq m(\tau) \Rightarrow \mu^\star(m(\Sigma')) \geq \mu^\star(m(\tau)) \tag{17}
\]

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

1. \( X_n \not\in XYZ \)
2. \( X_n \in Z \)
3. \( X_n \in X \) (or, symmetrically, \( X_n \in Y \))

**Case 1**: \( X_n \not\in XYZ \). We will show that \( \Delta_n \models \Sigma' \not\models \tau \), and the claim will follow from the induction hypothesis [16] because \( \Sigma' \) does not mention \( X_n \), and \( h(\Sigma') \geq h(\tau) \geq h(\Sigma) \) as required.

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

\[
a_1 \equiv a \land m(X_n) \quad a_2 \equiv a \land m^\star(X_n)
\]

such that \( \{a_1, a_2\} \subseteq m(\tau) \) and \( \{a_1, a_2\} \cap m(\Sigma') = \emptyset \). By our observation, \( \sigma_n = (X_n; R(B)) \). Therefore, we have that \( m(\sigma_n) \subseteq m(X_n) \) (i.e., see Table [1]).

So, we get that \( a_2 \not\models m(\sigma_n) \). Overall, we have that \( a_2 \not\models m(\sigma_n) \cup m(\Sigma') = m(\Sigma) \), and by Lemma 5.1, we get that \( \Delta_n \models \Sigma \not\models \tau \), a contradiction.
We observe that we claim that a form in some of these are some \(X = (X_n; R B)\), then \(m(\sigma_n) \subseteq m(A)\) (see Table 1). Hence, \(m(\sigma) \cap m(\sigma_n) = \emptyset\), and by Lemma 5.3 if \(\Delta_n \models E_1 \Sigma \Rightarrow \tau\) then it must hold that \(\Delta_n \models E_1 \Sigma \Rightarrow \tau\), and the claim follows from the induction hypothesis \([16]\) because \(\Sigma'\) does not mention \(X_n\), and \(h(\Sigma) \geq (\Sigma') \geq h(\tau)\).

**Case 3:** \(\tau = (W X_n; Y | Z)\). By the chain rule (see \([3]\)):

\[
(W X_n; Y | Z) = (W; Y | Z) + (X_n; Y | W Z)
\]

Hence, if \(\Delta_n \models E_1 \Sigma \Rightarrow \tau\) then \(\Delta_n \models E_1 \Sigma \Rightarrow \tau_1\), and \(\Delta_n \models E_1 \Sigma \Rightarrow \tau_2\). We have already shown, in case 1, that the former implies \(\Delta_n \models E_1 \Sigma \Rightarrow \tau_1\).

Let \(\sigma_n = (X_n; R B)\), and let \(Y = Y_1 \ldots Y_m\) where \(m \geq 1\). We claim that \(\Sigma \subseteq R\). Since \(\sigma_n\) is saturated then \(Y \subseteq R B\).

Now, suppose by way of contradiction, that \(Y_i \in B\) for some \(i \in [1, m]\). Consider the atom

\[
a = m(X_n) \cap m(Y_i) \cap_{X \in [n]} m^c(X).
\]

We observe that \(a \in m(\Sigma_2)\). By our assumption \(Y_i \in B\), then \(a \notin m(\Sigma_1)\).

On the other hand, for every \(\sigma = (A; B C) \in \Sigma'\) we also have that \(a \notin m(\sigma)\). To see why, note that \(X_n \notin AB\). Therefore, every atom of \(m(\sigma)\) contains at least two sets in positive form: \(m(X_i)\) for some \(X_i \in A\) and \(m(J_x)\) for some \(J_x \in B\). Since neither of these are \(X_n\), then at least one of them appears in negative form in \(a\). Overall, we get that \(m(X_n; Y | W Z) \notin m(\Sigma)\), and by Lemma 5.1 that \(\Delta_n \models E_1 \Sigma \Rightarrow \tau_2\). Hence, from \([18]\), we get that \(\Delta_n \models E_1 \Sigma \Rightarrow \tau_3\). This contradicts the assumption that \(\Sigma \Rightarrow \tau_3\).

Since \(Y \subseteq R\), we can write \(\sigma_n = (X_n; Y R W R Z R' B W B Z B')\) where \(R W \equiv R \cap W\), \(R Z \equiv R \cap Z\), and \(R' \equiv R' W R W Y\). Likewise, \(B W \equiv B \cap W\), \(B Z \equiv B \cap Z\), and \(B' \equiv B \cap B W B Z\). Further, since \(\sigma_n\) is saturated then \(W = R W B W\) and \(Z = R Z B Z\).

By the chain rule, we have:

\[
h(\sigma_n) = I_h(X_n; Y R W R Z R' B W B Z B')
\]

\[
= I_h(X_n; Y R W R Z B W Z B') + I_h(X_n; R' W Z B')
\]

\[
\geq I_h(X_n; R W R W B Z B') + I_h(X_n; Y | W Z B')
\]

\[
\geq I_h(X_n; Y | Z W B')
\]

Now, if \(B' = \emptyset\) then we are done because \(h(\sigma_n) \geq h(X_n; Y | Z W) = h(\tau_2)\) 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:

\[
A \defeq m(X_n) \cap \bigcup_{y \in Y} m(y) \cap \bigcup_{b \in B'} m(b) \cap_{X \in Z W} m^c(X)
\]

We note that \(m(\tau_2) \subseteq A\). By our assumption that \(\sigma_n = (X_n; R B W W Z B')\), then \(A \cap m(\sigma_n) = \emptyset\). Since \(\Delta_n \models \Sigma \Rightarrow \tau_2\) then by Lemma 5.3 it must hold that \(m(\Sigma') \subseteq A\).

Furthermore, since \(X_n \notin \var(\Sigma')\) then it must hold that \(m(\Sigma') \subseteq A'\) where:

\[
A' \defeq \bigcap_{y \in Y} m(y) \cap \bigcup_{b \in B'} m(b) \cap_{X \in Z W} m^c(X)
\]

Denote by \(\tau_3 \defeq (Y; B' W Z W)\), and hence \(m(\tau_3) = A'\) (see Table 1). In particular, \(m(\Sigma') \subseteq m(\tau_3)\), and by Lemma 5.1 we have that \(\Delta_n \models \Sigma' \Rightarrow \tau_3\). Since neither \(\Sigma'\) nor \(\tau_3\) mention \(X_n\), then by the induction hypothesis \([17]\), we have that \(\mu^*(m(\Sigma')) \geq \mu^*(m(\tau_3))\).

Since \(\Delta_n \models \Sigma \Rightarrow \tau_1\), and since \(X_n \notin \var(\tau_1)\), then by the argument of case 1 we have that \(\Delta_n \models \Sigma' \Rightarrow \tau_1\), and hence by Lemma 5.1 that \(m(\Sigma') \subseteq m(\tau_1)\).

Now, by the previous reasoning, we have also have that \(m(\Sigma') \subseteq m(\tau_1)\). By noting that \(m(\tau_1) \cap m(\tau_1) = \emptyset\), and applying Lemma 5.3 we get that \(m(\Sigma') \cap m(\tau_1) \subseteq m(\tau_1)\). Applying the induction hypothesis \([17]\), we get that \(\mu^*(m(\Sigma')) \cap m(\tau_1) \subseteq m(\tau_1)\).

Now, since \(\mu^*\) is set-additive, and \(m(\Sigma') \subseteq m(\tau_1)\), we get that \(\mu^*(m(\Sigma')) \subseteq m(\tau_1)\). And, by the one-to-one correspondence between Shannon’s information measures and the 1-measure (Theorem 2.2), we get that \(h(\Sigma') - h(\tau_3) \geq h(\tau_1)\).

Now, from \([19]\) we have that \(h(\sigma) \geq I_h(X_n; Y | Z W B')\).

By applying the chain rule:

\[
I_h(X_n; Y | Z W B') + I_h(Y; B' | Z W) = I_h(B' X_n; Y | W Z) \geq h(\tau_2)
\]

Overall, we get that:

\[
I_h(W; Y | Z) + I_h(X_n; Y | W Z) \leq h(\Sigma') - h(\tau_3) + h(\tau_3) = h(\Sigma)
\]

as required.

**TIGHTNESS OF BOUND**

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

\[
\Sigma = \{(X_1; X_2 X_3 \ldots X_{i-1}) : i \in \{2, \ldots, n\}\}
\]

Let \(\tau = (X_1; X_2 X_3 \ldots X_n)\). It is not hard to see that by the chain rule:

\[
I(X_1; X_2 X_3 \ldots X_n) = \sum_{i=2}^{n} I(X_1; X_i | X_2 \ldots X_{i-1}) = h(\Sigma)
\]

Hence, \(\Sigma \Rightarrow E_1 \tau\), and the bound of \([21]\) is tight.
7 APPROXIMATE IMPLICATION FOR MARGINAL CIs

In this section, we prove Theorem 4.2. Let \( \Sigma \) be a set of marginal mutual information terms, and let \( \tau = (A; B|D) \) such that \( \Gamma_n = E_1 \Sigma \vdash \tau \). Then, by the chain rule (3), \( \tau \) 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 \((X; Y)\) 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 \( \Sigma \) be a set of marginal mutual information terms, and let \( \tau = (a; b|C) \) be an elemental mutual information term. The following holds:

\[
\Gamma_n = E_1 \Sigma \vdash \tau \iff \exists (X; Y) \in \Sigma : XY \supseteq abC \text{ and } a \in X, b \in Y
\]

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

\[
t = m(a) \cap m(b) \bigcap_{y \in m(\tau \setminus ab)} m^c(y)
\]

Clearly, \(t \in m(\tau)\). 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 \(t \notin m(\sigma)\) because all the RVs in \(X\) appear in negative form in the atom \(t\). If this is the case for all \(\sigma \in \Sigma\), then \(t \notin m(\Sigma)\), and \(m(\tau) \not\subseteq m(\Sigma)\). But then, by Corollary 5.2, it cannot be that \(\Gamma_n = E_1 \Sigma \vdash \tau\), a contradiction.

So, we assume correctness for elemental terms \((a; b|C)\) where \(|C| \leq k - 1\), and prove for \(|C| = k\). Since \(\Gamma_n = E_1 \Sigma \vdash \tau\), then by Lemma 5.6 there exists a mutual information term \(\sigma = (X; Y) \in \Sigma\) such that \(XY \supseteq abC\). Hence, we denote \(C = CXC_Y\), where \(CX = X \cap C\) and \(CY = Y \cap C\). There are two cases. If \(\sigma = (a; b|C_X; C_Y)\) then, by Lemma 5.4, we have that \(h(\sigma) \geq h(\tau)\), and we are done.

Otherwise, w.l.o.g., \(\sigma = (a; C_X|Y; C_Y)\). By item 2 of Lemma 5.6, it holds that \(C_Y \neq \emptyset\).

We define:

\[
\alpha_1 \overset{\text{def}}{=} (a; C_Y|C_X) \quad \alpha_2 \overset{\text{def}}{=} (a; C_Y|b|C_X)
\]

By Lemma 5.4, we have that \(h(\sigma) \geq h(\alpha_1)\) and \(h(\sigma) \geq h(\alpha_2)\), thus \(\Gamma_n = E_1 \Sigma \vdash \{\alpha_1, \alpha_2\}\). Noting that \(\tau = (a; b|C_XC_Y)\), we have that \(\Gamma_n = E_1 \Sigma \vdash \{a; b|C_XC_Y\}\). By the chain rule (see 3) we have that \(\Sigma\) implies:

\[
(a; C_Y|C_X), (a; b|C_XC_Y) \Rightarrow (a; b|C_Y|C_X) \Rightarrow (a; b|C_X)
\]

In other words, we have that \(\Gamma_n = E_1 \Sigma \vdash (a; b|C_X)\).

By item 2 of Lemma 5.6 it holds that \(C_Y \neq \emptyset\). Hence, \(C_X \subseteq C\). Therefore, by the induction hypothesis, there exists an \(\alpha_3 \overset{\text{def}}{=} (aC_X^1, Z_1; bC_Y^2, Z_2) \in \Sigma\) where \(C_X = C_X^1C_X^2\).

In particular, by Lemma 5.4 we have that \(\alpha_3 \Rightarrow \alpha_4 \overset{\text{def}}{=} (a; b|C_X), \text{ and } h(\alpha_4) \leq h(\alpha_3)\) where \(\alpha_3 \in \Sigma\). Furthermore, by our assumption (i.e., that \(\sigma = (ab|C_X, CY)\)), then \(\alpha_3\) and \(\alpha_4\) are distinct. Consequently, we get:

\[
\begin{align*}
I(a; b|C_X) + I(a; C_Y|bC_X) & \leq h(\alpha_3) \\
\Rightarrow I(a; b|C_XC_Y) & \leq I(a; b|C_Y) = h(\tau)
\end{align*}
\]

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 \(\lambda\)-relaxation for a bounded \(\lambda\). 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 [20], 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


[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.


