Applications of Information Inequalities to Database Theory Problems *

Dan Suciu
University of Washington
Seattle, USA
suciu@cs.washington.edu

May 2, 2023

Abstract

The paper describes several applications of information inequalities to problems in database theory. The problems discussed include: upper bounds of a query's output, worst-case optimal join algorithms, the query domination problem, and the implication problem for approximate integrity constraints. The paper is self-contained: all required concepts and results from information inequalities are introduced here, gradually, and motivated by database problems.

1 Introduction

Notions and techniques from information theory have found a number of uses in various areas of database theory. For example, entropy and mutual information have been used to characterize database dependencies [Lee87a, Lee87b] and normal forms in relational and XML databases [AL02, AL05]. More recently, information inequalities were used with much success to obtain tight bounds on the size of the output of a query on a given database [AGM13, GLVV12, GM14, KNS16, KNS17], *and* to devise query plans for worst-case optimal join algorithms [KNS16, KNS17]. Information theory was also used to compare the sizes of the outputs of two queries, or, equivalently, to check query containment under bag semantics [KR11, KKNS21]. Finally, information theory has been used to reason about approximate integrity constraints in the data [KS22, KMP+20].

This paper presents some of these recent applications of information theory to databases, in a unified framework. All applications discussed here make use of information inequalities, which have been intensively studied in the information theory community [Yeu08, ZY97, ZY98, Mat07, KR13]. We will introduce gradually the concepts and results on information inequalities, motivating them with database applications.

We start by presenting in Sec. 3 a celebrated result in database theory: the AGM upper bound, which gives a tight upper bound on the query output size, given the cardinalities of the input relations. The AGM bound was first introduced by Grohe and Marx [GM06], and refined in its current form by Atserias, Grohe, and Marx [AGM13], hence the name AGM. (A related result appeared earlier in [FK98].) While the original papers already used information inequalities to prove these bounds, in this paper we provide an alternative, elementary proof, which is based on a family of inequalities due to Friedgut [Fri04], and which are of independent interest.

*Research supported in part by NSF IIS 1907997, NSF IIS 1954222, and NSF-BSF 2109922.

Next, we turn our attention in Sec. 4 to an extension of the AGM bound, by providing an upper bound on the size of the query's output using functional dependencies and statistics on degrees, in addition to cardinality statistics. The extension to functional dependencies was first studied by Gottlob et al. [GLVV12] and then by Khamis et al. [KNS16], while the general framework was introduced by Khamis et al. [KNS17]. Here, information inequalities are a necessity, and we use this opportunity to introduce entropic vectors and polymatroids, and to define information inequalities. We show simple examples of how to compute upper bounds on the query's output size using Shannon inequalities (monotonicity and submodularity, reviewed in Sec. 4.1).

A natural question is whether the upper bound on the query's output size provided by information inequalities is tight: we discuss this in Sec. 5. This question is surprisingly subtle, and it requires us to dig even deeper into information theory, and discuss non-Shannon inequalities. More than 30 years ago, Pippenger [Pip86] asserted that constraints on entropies are the "*laws of information theory*" and asked whether the basic Shannon inequalities form the complete laws of information theory, i.e., whether every constraint on entropies can be derived from the Shannon's basic inequalities. In a celebrated result published in 1998, Zhang and Yeung [ZY98] answered Pippenger's question negatively by finding a linear inequality that is satisfied by all entropic functions with 4 variables, but cannot be derived from Shannon's inequalities. Later, Matús [Mat07] proved that, for 4 variables or more, there are infinitely many, independent non-Shannon inequalities. In fact, it is an open problem whether the validity of an information inequality is decidable. We provide here a short, self-contained proof of Zhang and Yeung's result. This result has a direct consequence to our problem, computing an upper bound on the query's output size: we prove that Shannon inequalities are insufficient to compute a tight upper bound. In contrast, we show that the upper bound derived by using general information inequalities is tight, a result related to one by Gogacz and Torunczyk [GT17] (for cardinality constraints and functional dependencies only) and another one by Khamis et al. [KNS17] (for general degree constraints). The take-away of this section is that we have two upper bounds on the query's output size: one that uses Shannon inequalities, which is computable but not always tight, and another one that uses general information inequalities, which is tight but whose computability is an open problem.

This motivates us to look in Sec. 6 at a special case, when the two bounds coincide and, thus, are both tight and computable. This special case is when the statistics are restricted to cardinalities, and to degrees on a single variable. We call the corresponding class of information inequalities *simple inequalities*, and prove that they are valid for all entropic vectors iff they are provable using Shannon inequalities. Moreover, in this special case, the worst-case database instances (where the size of the query's output reaches the theoretical upper bound) have a simple yet interesting structure, called *normal database instances*, which generalize the *product database instances* that are the worst case instances for the AGM bound.

In Sec. 7 we turn to the most exciting application of upper bounds to the query's output size: the design of Worst Case Optimal Join, WCOJ, algorithms, which compute a query in a time that does not exceed the upper bound on their output size. Thus, a WCOJ algorithm is *worst-case optimal*. The vast majority of database systems today compute a conjunctive query as a sequence of binary joins, whose intermediate results may exceed the upper bound on the final output size. Therefore, database execution engines are not WCOJ algorithms. For that reason, the discovery of the first WCOJ algorithm by Ngo, Porat, Ré, and Rudra [NPRR12, NPRR18] was a highly celebrated result. While the original WCOJ algorithm was complex, some of the same authors described a very simple WCOJ, called Generic Join (GJ) in [NRR13], which, together with its refinement Leapfrog Trie Join (LFTJ) [Vel14] forms the basis of the few implementations to date [SOC16, FBS+20, MKS21, WWS23]. Looking back at these results, we observe that any concrete WCOJ algorithm also provides a proof of the upper bound of the query's output size, since the size of the output cannot exceed the runtime of the algorithm. A WCOJ algorithm can be designed in reverse: start from a *proof* of the upper bound, then convert that proof into a WCOJ algorithm. We call this paradigm *from proofs to algorithms*, and illustrate it on three different proof systems for information inequalities: we derive GJ, an algorithm we call Heavy/Light, and PANDA.

Next, in Sec. 8 we move beyond upper bounds, and consider a related problem: given two queries, check whether the size of the output of the second query is always greater than or equal to that of the first query. This problem, called the *query domination problem*, is equivalent to the query containment problem under bag semantics. The latter was introduced by Chaudhuri and Vardi [CV93], is motivated by the semantics of SQL, where queries return duplicates, hence the answer to a query is a bag rather than a set. The query containment problem is: given two queries, interpreted under bag semantics, check whether the output of the first query is always contained in that of the second query. It has been shown that the containment problem is undecidable for *unions* of conjunctive queries [IR95] and for conjunctive queries with *inequalities* [JKV06], by reduction from Hilbert's 10th problem. However, it remains an open problem to date whether the containment of two conjunctive queries is decidable. We describe in this section a surprising finding by Kopparty and Rossman [KR11], who have reduced the containment problem to information inequalities. This result was further extended in [KKNS21], and it was shown that the containment problem under bag semantics is computationally equivalent to information inequalities with max, which are inequalities that assert that the maximum of a finite number of linear expressions is ≥ 0. The decidability of either of these problems remains open to date.

Finally, we present in Sec. 9 another, quite distinct application of information inequalities: reasoning about approximate integrity constraints. The implication problem for integrity constraints asks whether a set of integrity constraints logically implies some other constraint: this is a problem in Logic, and consists of checking the validity of a sentence $\bigwedge_i \sigma_i \Rightarrow \sigma$. When the integrity constraints can be captured by some information measures, such as is the case for Functional Dependencies and Multivalued Dependencies, then an implication can be described as a *conditional information inequality*. The problem we study is whether the exact implication problem can be *relaxed* to an inequality between these information measures, $\sum_i h(\sigma_i) \geq h(\sigma)$. We review a result from [KS22] stating that every exact implication between FDs and MVDs relaxes to an inequality. However, in a surprising result, Kaced and Romashchenko [KR13] have given examples of conditional information inequalities that do *not* relax. In other words, the exact implication holds, but the tiniest violation of an integrity constraint in the premise may cause arbitrarily large violation of the integrity constraint in the consequence. Yet in another turn, [KS22] show that *every* conditional information inequality relaxes with some error term, which can be made arbitrarily small, at the cost of increasing the coefficients of the terms representing the premise. In particular, every conditional inequality could be derived from an unconditioned inequality, by having the error term tend to zero, since in the conditional inequality the premise is assumed to be zero, hence the magnitudes of their coefficients do not matter. This section leads us to our deepest dive into the space of entropic vectors and almost entropic vectors: we show that the set of entropic vectors is neither convex nor a cone, that its topological closure is a convex cone, called the set of *almost entropic functions*, and use the theory of closed convex cones to prove the relaxation-with-error theorem.

Acknowledgments I am deeply indebted to my collaborators, especially Hung Q. Ngo who introduced me to applications of information inequalities to databases and with whom I had wonderful collaborations, and also to Mahmoud Abo Khamis, Batya Kenig, and Phokion G. Kolaitis. I also thank Dan Olteanu and Andrei Romashchenko for commenting on an early version of this paper.

2 Basic Notations

For two natural numbers M, N we denote by $[M : N] \stackrel{\text{def}}{=} \{M, M+1, \ldots, N\}$; when $M = 1$ we abbreviate $[1 : N]$ by N. We will use upper case X, Y, Z for variable names, and lower case x, y, z for values of these variables. We use boldface for tuples of variables, e.g. $\boldsymbol{X}, \boldsymbol{Y}$, or tuples of values, e.g. $\boldsymbol{x}, \boldsymbol{y}$.

A **conjunctive query**, CQ, is an expression of the form:

$$Q(\boldsymbol{Y}_0) = \exists \boldsymbol{Z}(R_1(\boldsymbol{Y}_1) \wedge \cdots \wedge R_m(\boldsymbol{Y}_m)) \tag{1}$$

Each $R_j(\boldsymbol{Y}_j)$ is called an *atom*: R_j is a relation name, and $\boldsymbol{Y}_j$ are variables. We refer to $\boldsymbol{Y}_j$ interchangeably as the *variables* of R_j, or the *attributes* of R_j. The variables $\boldsymbol{Z}$ are called *existential variables*, while $\boldsymbol{Y}_0$ are called *head variables*. We denote by n the total number of variables in the query, and by $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ the set of these variables. Thus $\boldsymbol{X} = \boldsymbol{Y}_0 \cup \boldsymbol{Z}$, and $\boldsymbol{Y}_j \subseteq \boldsymbol{X}$, $\forall j$.

Fix some infinite domain Dom. If $\boldsymbol{X}$ is a set of variables, then we write $\mathsf{Dom}^{\boldsymbol{X}}$ for the set of $\boldsymbol{X}$-tuples. A *database instance* is $\boldsymbol{D} = (R_1^D, \ldots, R_m^D)$, where, for each $j = 1, m$, $R_j^D \subseteq \mathsf{Dom}^{\boldsymbol{Y}_j}$, where $\boldsymbol{Y}_j$ are the attributes of R_j. Unless otherwise stated, relations are assumed to be finite. When $\boldsymbol{D}$ is clear from the context, then we will drop the superscript and write simply R_j for the instance R_j^D, for $j = 1, m$.

We denote by $Q(\boldsymbol{D}) \subseteq \mathsf{Dom}^{\boldsymbol{Y}_0}$ the output, or answer to the query Q on the database D. The *query evaluation problem* is: given a database instance $\boldsymbol{D}$, compute the output $Q(\boldsymbol{D})$. The design and analysis of efficient query evaluation algorithms is a fundamental problem in database systems and database theory. For the complexity of the query evaluation problem, we consider only the *data complexity*, where Q is fixed, and the complexity is a function of the input database $\boldsymbol{D}$.

For a simple illustration, consider:

$$Q(X) = \exists Y \exists Z(R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)) \tag{2}$$

Q returns all nodes x that belong to an RST triangle.

A *Boolean conjunctive query* is a conjunctive query with no head variables. At the other extreme, a *full conjunctive query* is a query with no existential variables. For example, the query:

$$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X) \tag{3}$$

is a full CQ computing all triangles formed by the relations R, S, T. Full conjunctive queries are of special importance because they often occur as intermediate expressions during query evaluation. Unless otherwise stated, we will assume in this paper that the query is a full conjunctive query *without self-joins*, meaning that the relation names of the atoms $R_1, R_2, \ldots$ are distinct. Such a query is also called a natural join of the relations $R_1, \ldots, R_m$.

Fix a relation $R(\boldsymbol{X})$, with n attributes. A **functional dependency**, or FD, is an expression $\boldsymbol{U} \to \boldsymbol{V}$, where $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$. An instance R^D *satisfies* the FD, and we write $R^D \models \boldsymbol{U} \to \boldsymbol{V}$, if for any two tuples $\boldsymbol{x}_1, \boldsymbol{x}_2 \in R^D$, $\boldsymbol{x}_1.\boldsymbol{U} = \boldsymbol{x}_2.\boldsymbol{U}$ implies $\boldsymbol{x}_1.\boldsymbol{V} = \boldsymbol{x}_2.\boldsymbol{V}$. A set of functional dependencies Σ *implies* a functional dependency $\boldsymbol{U} \to \boldsymbol{V}$, in notation $\Sigma \models \boldsymbol{U} \to \boldsymbol{V}$, if, for every instance R^D, if $R^D \models \Sigma$ then $R^D \models \boldsymbol{U} \to \boldsymbol{V}$. Armstrong's axioms [AD80] form a complete axiomatization of the implication problem for FDs. The *closure* of $\boldsymbol{U} \subseteq \boldsymbol{X}$, denoted $\boldsymbol{U}^+$, is the set of all attributes X s.t. $\Sigma \models \boldsymbol{U} \to X$. The closure can be computed in polynomial time in the size of $\boldsymbol{U}$ and Σ. A set $\boldsymbol{U}$ is *closed* if $\boldsymbol{U}^+ = \boldsymbol{U}$. A *super-key* for $R(\boldsymbol{X})$ is a set $\boldsymbol{U}$ with the property that $\boldsymbol{U}^+ = \boldsymbol{X}$, and a *key* is a minimal set of attributes that is a superkey.

A finite **lattice** is a partially ordered set $(L, \preceq)$ where every two elements $x, y \in L$ have a least upper bound $x \vee y$, and a greatest lower bound $x \wedge y$. In particular the lattice has a smallest and a largest element, usually denoted by $\hat{0}, \hat{1}$. Consider now a set of variables $\boldsymbol{X}$, and a set of functional dependencies, Σ, over $\boldsymbol{X}$. We denote by $(L_\Sigma, \subseteq)$ the lattice consisting of the closed sets, $L_\Sigma = \{\boldsymbol{U} \mid \boldsymbol{U}^+ = \boldsymbol{U}\}$. One can verify that the operations in this lattice are $\boldsymbol{U} \wedge \boldsymbol{V} \stackrel{\text{def}}{=} \boldsymbol{U} \cap \boldsymbol{V}$ and $\boldsymbol{U} \vee \boldsymbol{V} \stackrel{\text{def}}{=} (\boldsymbol{U} \cup \boldsymbol{V})^+$.

The **cartesian product** of two relations $R(\boldsymbol{X}), S(\boldsymbol{Y})$ with disjoint sets of attributes is the set $R \times S \stackrel{\text{def}}{=} \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x} \in R, \boldsymbol{y} \in S\}$ with attributes $\boldsymbol{X} \cup \boldsymbol{Y}$; its size is $|R \times S| = |R| \cdot |S|$. Fix a set of attributes $\boldsymbol{X}$, and two $\boldsymbol{X}$-tuples $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{x}' = (x'_1, \ldots, x'_n)$. Their **domain product** is the $\boldsymbol{X}$-tuple $\boldsymbol{x} \otimes \boldsymbol{x}' \stackrel{\text{def}}{=} ((x_1, x'_1), \ldots, (x_n, x'_n))$; thus, the value of each attribute is a pair.

Definition 2.1. *The* domain product *of two relation instances R and S, with the same set of attributes $\boldsymbol{X}$, is $R \otimes S \stackrel{\text{def}}{=} \{\boldsymbol{x} \otimes \boldsymbol{x}' \mid \boldsymbol{x} \in R, \boldsymbol{x}' \in S\}$.*

We have $|R \otimes S| = |R| \cdot |S|$. If $\boldsymbol{D}_i = (R_1^{D_i}, \ldots, R_m^{D_i})$, $i = 1, 2$, are two database instances over the same schema, then we define their domain product $\boldsymbol{D}_1 \otimes \boldsymbol{D}_2$ as $(R_1^{D_1} \otimes R_1^{D_2}, \ldots, R_m^{D_1} \otimes R_m^{D_2})$. One can check that $Q(\boldsymbol{D}_1 \otimes \boldsymbol{D}_2) = Q(\boldsymbol{D}_1) \otimes Q(\boldsymbol{D}_2)$ for any conjunctive query Q. The domain product should not be confused with the cartesian product. It was first introduced by Fagin [Fag82] (under the name *direct product*) to prove the existence of an Armstrong relation for constraints defined by Horn clauses, and later used by Geiger and Pearl [GP93] to prove that Conditional Independence constraints on probability distributions also admit an Armstrong relation. The same construction appears under the name "fibered product" in [KR11].

3 Warmup: the AGM Bound

Consider a full conjunctive query:

$$Q(\boldsymbol{X}) = \bigwedge_{j=1,m} R_j(\boldsymbol{Y}_j) \tag{4}$$

where $\boldsymbol{X} = \{X_1, \ldots, X_n\}$. Assume we have a database $\boldsymbol{D}$, and we know the cardinality of each relation R_j^D. How large could the query output be? The answer is given by an elegant result, initially formulated by Grohe and Marx [GM06] and later refined by Atserias, Grohe, and Marx [AGM13], and is called today the AGM bound of the query Q. To state this bound, we first need to review the connection between conjunctive queries and hypergraphs.

We associate Q in (4) with the hypergraph $\mathcal{H} = (\boldsymbol{X}, E)$, where $E = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m\}$. In other words, the nodes of the hypergraph are the variables, and its hyperedges are the atoms of the query. A *fractional edge cover* of the hypergraph $\mathcal{H}$ is a tuple of non-negative weights $\boldsymbol{w} = (w_j)_{j=1,m}$, such that every variable X_i is *covered*, meaning:

$$\forall i = 1, n : \qquad \sum_{j: X_i \in \boldsymbol{Y}_j} w_j \geq 1 \tag{5}$$

A fractional edge cover of the query Q is a fractional edge cover of its associated hypergraph. The AGM bound is the following:

Theorem 3.1 (AGM Bound)**.** *For any fractional edge cover $\boldsymbol{w}$ of the query* (4)*, and every instance $\boldsymbol{D}$:*

$$|Q(\boldsymbol{D})| \leq \prod_{j=1,m} |R_j^D|^{w_j} \tag{6}$$

To reduce clutter, we will often drop $\boldsymbol{D}$ from both $Q(\boldsymbol{D})$ and R_j^D, and write the bound simply as $|Q| \leq \prod_j |R_j|^{w_j}$.

Let $\boldsymbol{B} = (B_j)_{j=1,m}$ be a non-negative vector, representing the cardinalities of the relations in the database. We define:

$$AGM(Q, \boldsymbol{B}) \stackrel{\text{def}}{=} \min_{\boldsymbol{w}} \prod_{j=1,m} B_j^{w_j} \tag{7}$$

where $\boldsymbol{w}$ ranges over all fractional edge covers of the query's hypergraph. Then Theorem 3.1 can be restated as follows: for every instance $\boldsymbol{D}$, if $|R_j^D| \leq B_j$ for $j = 1, m$, then $|Q(\boldsymbol{D})| \leq AGM(Q, \boldsymbol{B})$. When $\boldsymbol{B}$ is clear from the context, then we write the bound simply as $AGM(Q)$.

Before we prove the bound, we illustrate it with a classic example.

Example 3.2. *Consider the triangle query* (3)*, which we repeat here:* $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$*. Its associated hypergraph is a graph with three nodes* X, Y, Z *and three edges forming a triangle. A fractional edge cover is any non-negative tuple* (w_R, w_S, w_T) *satisfying:*

$$\begin{array}{llll} \textit{Cover } X: & w_R + & & w_T \quad \geq 1 \\ \textit{Cover } Y: & w_R + & w_S & \quad\quad \geq 1 \\ \textit{Cover } Z: & & w_S + & w_T \quad \geq 1 \end{array}$$

The inequality $|Q| \leq |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$ *holds for every fractional edge cover. Consider the following four fractional edge covers:* $(0, 1, 1), (1, 0, 1), (1, 1, 0), (1/2, 1/2, 1/2)$*: these are the extreme vertices of the edge-covering polytope. It follows that the AGM bound in* (7) *is achieved at one of the four extreme vertices:*

$$\begin{aligned} &AGM(Q) = \\ &\quad \min\left(|S| \cdot |T|, |R| \cdot |T|, |R| \cdot |S|, |R|^{1/2} \cdot |S|^{1/2} \cdot |T|^{1/2}\right) \end{aligned}$$

When $|R| = |S| = |T| = N$ *then* $AGM(Q) = N^{3/2}$*.*

In the rest of this section we will prove the AGM bound (6), then show that the bound is tight.

Friedgut's Inequalities While the original proof of the AGM bound used information inequalities, we postpone the discussion of information inequalities until Sec. 4, where we consider more general statistics. Instead, we give here a simple, elementary proof, based on an elegant family of inequalities introduce by Friedgut [Fri04].

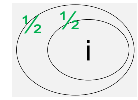
Fix a hypergraph $\mathcal{H} = (\boldsymbol{X}, E)$. Let $N > 0$ be a natural number, and for each hyperedge $\boldsymbol{Y}_j \in E$, let $r_j \in \mathbb{R}_+^{N^{|\boldsymbol{Y}_j|}}$ be a non-negative, multi-dimensional vector with $|\boldsymbol{Y}_j|$ dimensions; we will refer to r_j as a *tensor*. In what follows, we denote by $\boldsymbol{i}$ a tuple $\boldsymbol{i} = (i_1, \ldots, i_n) \in [N]^{\boldsymbol{X}}$, and by $\boldsymbol{i}_j$ its projection on $\boldsymbol{Y}_j$.

Theorem 3.3 (Friedgut's Inequality)**.** *[Fri04] For every fractional edge cover* $\boldsymbol{w}$ *of the hypergraph* $\mathcal{H}$*, the following holds:*

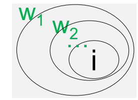
$$\sum_{\boldsymbol{i}} \prod_{j=1,m} r_j[\boldsymbol{i}_j] \leq \prod_{j=1,m} \left(\sum_{\boldsymbol{i}_j} r_j[\boldsymbol{i}_j]^{\frac{1}{w_j}} \right)^{w_j} \tag{8}$$

Fig. 1 illustrates several instances of (8). We invite the reader to check that Loomis–Whitney's inequality [LW49] is also an instance such an inequality. Using Theorem 3.3 we can prove the AGM

Cauchy-Schwartz: $\sum_i a[i] \cdot b[i] \leq \left(\sum_i a[i]^2\right)^{1/2} \cdot \left(\sum_i b[i]^2\right)^{1/2}$



Hölder: $\sum_i \prod_j a_j[i] \leq \prod_j \left(\sum_i a_j[i]^{\frac{1}{w_j}}\right)^{w_j}$ when $\sum_j w_j \geq 1$



Friedgut: $\sum_{i,j,k} a[i,j] \cdot b[j,k] \cdot c[k,i] \leq \left(\sum_{i,j} a[i,j]^2\right)^{1/2} \cdot \left(\sum_{j,k} b[j,k]^2\right)^{1/2} \cdot \left(\sum_{k,i} c[k,i]^2\right)^{1/2}$

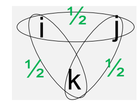


Figure 1: Examples of Friedgut's inequalities (8). In each case we show the associated hypergraph on the right.

bound as follows. Given a relational instance $\boldsymbol{D} = (R_1^D, \ldots, R_m^D)$ define the following tensors:

$$r_j[x_1, \ldots, x_{a_j}] \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } (x_1, \ldots, x_{a_j}) \in R_j^D \\ 0 & \text{otherwise} \end{cases}$$

Then the LHS of (8) is $|Q(\boldsymbol{D})|$ and the RHS is $\prod_j |R_j|^{w_j}$.

Proof. (of Theorem 3.3) While the original proof also used information inequalities, we give here a direct proof, by induction on the number n of vertices of the hypergraph $\mathcal{H}$. (This proof generalizes Loomis–Whitney's proof in [LW49].)

We replace each tensor expression $r_j[\boldsymbol{i}_j]$ with $(r_j[\boldsymbol{i}_j])^{w_j}$, then in order to prove (8) it suffices to prove:

$$\sum_{\boldsymbol{i}} \prod_j r_j[\boldsymbol{i}_j]^{w_j} \leq \prod_j \left(\sum_{\boldsymbol{i}_j} r_j[\boldsymbol{i}_j]\right)^{w_j} \tag{9}$$

We notice that the index variables $\boldsymbol{i} = (i_1, \ldots, i_n)$ used in the summation correspond one-to-one to the nodes of the hypergraph $\boldsymbol{X} = \{X_1, \ldots, X_n\}$, and the subset $\boldsymbol{i}_j$ contains the index variables corresponding to nodes in $\boldsymbol{Y}_j$. We now prove (9) by induction on n.

When $n = 1$ then this is Hölder's inequality (see Fig. 1), whose proof can be found in textbooks. Assume $n > 1$ and consider the hypergraph $\mathcal{H}'$ obtained by removing the last variable X_n: its nodes are $\{X_1, \ldots, X_{n-1}\}$ and its hyperedges are $\{\boldsymbol{Y}_j - \{X_n\} \mid j = 1, m\}$. The weights $w_1, \ldots, w_m$ continue to be a fractional edge cover for $\mathcal{H}'$. Group the LHS of Eq. (9) by factoring out the sum over the variable i_n, and apply induction hypothesis to the summation over the other variables $i_1, \ldots, i_{n-1}$, which form the hypergraph $\mathcal{H}'$:

$$\sum_{i_n} \left(\sum_{i_1, \ldots, i_{n-1}} \prod_j r_j[\boldsymbol{i}_j]^{w_j}\right) \leq \sum_{i_n} \prod_j \left(\sum_{\boldsymbol{i}_j - \{i_n\}} r_j[\boldsymbol{i}_j]\right)^{w_j}$$

We factor out the products that do not depend on the variable i_n, then use the fact that $\sum_{j: X_n \in \boldsymbol{Y}_j} w_j \geq 1$ because X_n is covered, and apply Hölder's inequality (Fig. 1) with $a_j[i_n] \stackrel{\text{def}}{=} r_j[\boldsymbol{i}_j]$.

The RHS of the expression above becomes:

$$\prod_{j:X_n \notin \boldsymbol{Y}_j} \left(\sum_{\boldsymbol{i}_j} r_j[\boldsymbol{i}_j]\right)^{w_j} \cdot \sum_{i_n} \prod_{j:X_n \in \boldsymbol{Y}_j} \left(\sum_{\boldsymbol{i}_j - \{i_n\}} r_j[\boldsymbol{i}_j]\right)^{w_j}$$
$$\leq \prod_{j:X_n \notin \boldsymbol{Y}_j} \left(\sum_{\boldsymbol{i}_j} r_j[\boldsymbol{i}_j]\right)^{w_j} \cdot \prod_{j:X_n \in \boldsymbol{Y}_j} \left(\sum_{i_n} \sum_{\boldsymbol{i}_j - \{i_n\}} r_j[\boldsymbol{i}_j]\right)^{w_j}$$

This is the RHS of (9), which completes the proof. □

The lower bound How tight is the AGM bound? One key insight in [AGM13] is that, while the upper bound is described by a linear program, a lower bound can be described using the dual linear program: tightness follows from the strong duality theorem for linear programs. They proved:

Theorem 3.4. *For any query Q with n variables, and vector $\boldsymbol{B}$ there exists a database $\boldsymbol{D}$ s.t. $|Q(\boldsymbol{D})| \geq \frac{1}{2^n} AGM(Q, \boldsymbol{B})$. We call such a database $\boldsymbol{D}$ a* worst-case instance.

Proof. The logarithm of the AGM bound (7) is the optimal value of the following *primal linear program*:

$$\begin{aligned} \text{minimize} \quad & \sum_j w_j \log B_j \\ \text{where} \quad & \forall i: \sum_{j:X_i \in \boldsymbol{Y}_j} w_j \geq 1 \\ & \forall j: w_j \geq 0 \end{aligned}$$

The *dual linear program* is:

$$\begin{aligned} \text{maximize} \quad & \sum_i v_i \\ \text{where} \quad & \forall j: \sum_{i:X_i \in \boldsymbol{Y}_j} v_i \leq \log B_j \\ & \forall i: v_i \geq 0 \end{aligned}$$

For any two feasible solutions $\boldsymbol{w}, \boldsymbol{v}$ of the primal and dual, *weak duality* holds: $\sum_j w_j \log B_j \geq \sum_i v_i$. If $\boldsymbol{w}^*, \boldsymbol{v}^*$ are the optimal solutions, then the *strong duality* theorem states that these two expressions are equal, therefore:

$$AGM(Q, \boldsymbol{B}) = 2^{\sum_j w_j^* \log B_j} = 2^{\sum_i v_i^*} = \prod_i 2^{v_i^*} \tag{10}$$

If $\boldsymbol{v}$ is any dual solution, we construct the following database instance $\boldsymbol{D}$: for each variable X_i, define the domain $V_i \stackrel{\text{def}}{=} [\lfloor 2^{v_i} \rfloor] = \{1, 2, \ldots, \lfloor 2^{v_i} \rfloor\}$, and set $R_j^D \stackrel{\text{def}}{=} \times_{i:X_i \in \boldsymbol{Y}_j} V_i$, for $j = 1, m$. We call $\boldsymbol{D}$ a *product database instance*, because each relation is a cartesian product. $\boldsymbol{D}$ satisfies the cardinality constraints $|R_j^D| \leq B_j$ because

$$|R_j^D| = \prod_{i:X_i \in \boldsymbol{Y}_j} \lfloor 2^{v_i} \rfloor \leq 2^{\sum_{i:X_i \in \boldsymbol{Y}_j} v_i} \leq 2^{\log B_j} = B_j$$

Similarly, the output to the query is the product $Q(\boldsymbol{D}) = \times_i V_i$, and its size is $\prod \lfloor 2^{v_i} \rfloor$. At optimality, when $\boldsymbol{v} = \boldsymbol{v}^*$,

$$|Q(\boldsymbol{D})| = \prod_i \lfloor 2^{v_i^*} \rfloor \tag{11}$$

Theorem 3.4 follows from (10) and (11), and observing that $\lfloor 2^{v_i^*} \rfloor \geq \frac{1}{2} 2^{v_i^*}$. □

Thus, one could say that the AGM bound is tight up to a "rounding error". The original paper [AGM13] provides an extensive discussion on tightness and proves two facts. First, they construct arbitrarily large databases $\boldsymbol{D}$ where the AGM bound is tight exactly. Second, they describe an example where the ratio between the lower and upper bound can be arbitrarily close to $1/2^n$, as n grow arbitrarily large; despite this example, the AGM bound is considered to be tight for practical purposes.

Discussion The AGM bound is elegant in that it solves completely the problem it set out to solve: find the tight upper bound when the cardinalities of all relations are known, and nothing else is known. However, the bound is limited, in that it cannot take advantage of other statistics or constraints on the input data, which are often available in practice. For example, consider the join of two relations, $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z)$, and assume that both $|R|, |S| \leq B$. The AGM bound is $|Q| \leq B^2$ (because the only fractional edge cover is $(1, 1)$), and the reader can check that this is tight, i.e. there exists relations where $|R| = |S| = B$ and $|Q| = B^2$. But, in practice, joins are often key/foreign-key joins, for example, $S.Y$ may be a key in S, and in that case $|Q| \leq B$, because every tuple in R joins with at most one tuple in S. In order to account for additional information about the data, like keys or constraints on degrees, we need to use a more powerful tool than Friedgut's inequalities: information inequalities.

4 Max-Degree Query Bounds

We describe now the general framework for computing an upper bound for the query output size, using information inequalities. We will use the cardinalities of the input relations (like in the AGM bound), keys or, more generally, functional dependencies for individual relations, and bounds on degrees, also called maximum frequencies, which generalize keys and functional dependencies. This section is based largely on [GLVV12, KNS16, KNS17]. We start with a short review of information inequalities.

4.1 Background on Information Inequalities

Consider a finite probability distribution (D, p), where $p : D \to [0, 1]$, $\sum_{x \in D} p(x) = 1$. We denote by X the random variable with outcomes in D, and define its *entropy* as:

$$H(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x) \tag{12}$$

If $N \stackrel{\text{def}}{=} |D|$, then $0 \leq H(X) \leq \log N$, the equality $H(X) = 0$ holds iff X is deterministic (i.e. $\exists x \in D$, $p(x) = 1$), and the equality $H(X) = \log N$ holds iff X is uniformly distributed (i.e. $p(x) = 1/N$ for all $x \in N$).

Consider now a finite probability distribution (R, p), where $R \subseteq \mathsf{Dom}^{\boldsymbol{X}}$ is a non-empty, finite relation with n attributes $\boldsymbol{X} = \{X_1, \ldots, X_n\}$. We will always assume w.l.o.g. that R is the support of p, otherwise we just remove from R the tuples $\boldsymbol{x}$ where $p(\boldsymbol{x}) = 0$. For each $\alpha \subseteq [n]$, define

X	Y	Z	
0	0	0	$p = 1/4$
0	1	1	$p = 1/4$
1	0	1	$p = 1/4$
1	1	0	$p = 1/4$

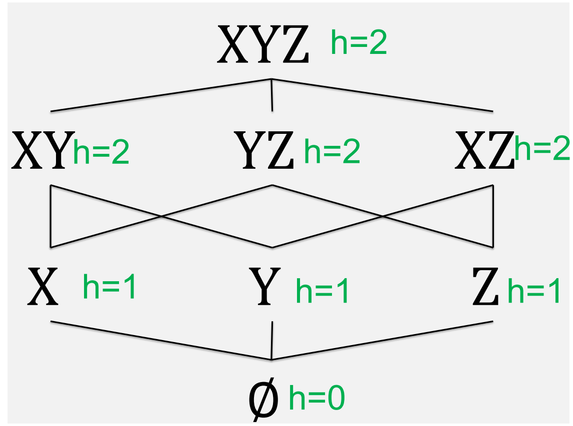


Figure 2: A relation defining the parity entropy $\boldsymbol{h}$. The marginal distribution of X is $p(X = 0) = p(X = 1) = 1/2$, hence its entropy is $h(X) = 1$, and similarly for the others values.

$X_\alpha \stackrel{\text{def}}{=} (X_i)_{i\in\alpha}$ the joint random variable obtained as follows: draw randomly a tuple $\boldsymbol{x} \in R$ with probability $p(\boldsymbol{x})$, then return $\boldsymbol{x}_\alpha$. We associate the probability space (R, p) with a vector $\boldsymbol{h} \in \mathbb{R}_+^{2^{[n]}}$, by defining $\boldsymbol{h}_\alpha \stackrel{\text{def}}{=} H(\boldsymbol{X}_\alpha)$, and call $\boldsymbol{h}$ an *entropic vector*. For any vector $\boldsymbol{h}$ (entropic or not) we will write $h(\boldsymbol{X}_\alpha)$ for $\boldsymbol{h}_\alpha$. In other words, we will blur the distinction between a vector in $\mathbb{R}_+^{2^{[n]}}$, a vector in $\mathbb{R}_+^{2^{\boldsymbol{X}}}$, and a function $2^{\boldsymbol{X}} \to \mathbb{R}_+$.

In analyzing properties of queries, we often examine the entropic vector derived from a uniform distribution.

Definition 4.1. *The* uniform *probability space associated to a non-empty, finite relation $R \subseteq \mathsf{Dom}^{\boldsymbol{X}}$ is (R, p) where $p(\boldsymbol{x}) = 1/|R|$ for every tuple $\boldsymbol{x} \in R$. We will call its entropic vector $\boldsymbol{h}$* uniform *and say that it is* associated *to R.*

If $\boldsymbol{h}$ is associated to R, then $h(\boldsymbol{X}) = \log |R|$, and $h(\boldsymbol{U}) \leq \log |\Pi_{\boldsymbol{U}}(R)|$ for every subset $\boldsymbol{U} \subseteq \boldsymbol{X}$. Fig. 2 illustrates the entropic vector h associated to a relational instance R with attributes X, Y, Z; we call h the *parity function*, because the relation R contains all triples (x, y, z) that have an even number of 1.

Any entropic vector $\boldsymbol{h} \in \mathbb{R}_+^{2^{\boldsymbol{X}}}$ satisfies the following *basic Shannon inequalities*:

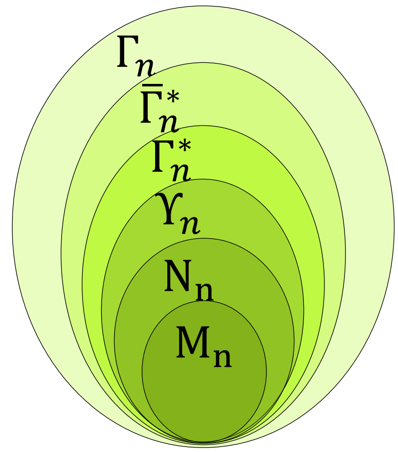
$$h(\emptyset) = 0 \tag{13}$$

$$h(\boldsymbol{U} \cup \boldsymbol{V}) \geq h(\boldsymbol{U}) \tag{14}$$

$$h(\boldsymbol{U}) + h(\boldsymbol{V}) \geq h(\boldsymbol{U} \cup \boldsymbol{V}) + h(\boldsymbol{U} \cap \boldsymbol{V}) \tag{15}$$

The last two inequalities are called called *monotonicity* and *submodularity* respectively. A *Shannon inequality* is a positive linear combination of basic Shannon inequalities.

Any vector $h : 2^{\boldsymbol{X}} \to \mathbb{R}_+$ that satisfies the basic Shannon inequalities is called a *polymatroid*. The set of entropic vectors is denoted by Γ_n^* and the set of polymatroids is denoted by Γ_n, where $n = |\boldsymbol{X}|$ is the number of variables. The following holds: $\Gamma_n^* \subsetneq \Gamma_n \subsetneq \mathbb{R}_+^{2^{[n]}}$. In particular, not every polymatroid is entropic, as we will see shortly (in Fig. 5). Fig. 3 represents these two sets, as well as other sets, defined later in this paper. In some of the literature the entropic vectors and the polymatroids are defined as $(2^n - 1)$-dimensional vectors, by dropping the $\emptyset$-dimension, because, in that case, both sets Γ_n^* and Γ_n have a non-empty interior. We prefer to use 2^n dimensions since this simplifies most of the discussion, and postpone dealing with the non-empty interior until Section 9.6.



Γ_n: polymatroids
$\bar{\Gamma}_n^*$: almost-entropic
Γ_n^*: entropic
Υ_n: group realizable
N_n: normal polymatroids
M_n: modular polymatroids

Figure 3: Landscape of polymatroids

An information inequality is an assertion stating that a linear expression of entropic terms is ≥ 0.

Definition 4.2. *We associate to any vector* $\boldsymbol{c} \in \mathbb{R}^{2^{[n]}}$ *the following* information inequality*:*

$$\sum_{\alpha \subseteq [n]} c_\alpha h(\boldsymbol{X}_\alpha) \geq 0 \tag{16}$$

By using the dot-product notation, we can write the inequality as $\boldsymbol{c} \cdot \boldsymbol{h} \geq 0$. *If the inequality holds for all* $h \in K$*, where* $K \subseteq \mathbb{R}_+^{2^{[n]}}$ *is some set, then we say that it is valid for* K*, and write* $K \models \boldsymbol{c} \cdot \boldsymbol{h} \geq 0$.

Thus, we will talk about inequalities valid for entropic vectors, or valid for polymatroids, and the latter are precisely the Shannon inequalities (this is implicit in the proof of Th. 5.2 below). Any Shannon inequality is also valid for entropic vectors; however, we will see in Th. 5.7 below a non-Shannon inequality, which is valid for entropic vectors, but not for polymatroids. In analogy with mathematical logic, one should think the vectors $\boldsymbol{h}$ as *models*, inequalities $\boldsymbol{c} \cdot \boldsymbol{h} \geq 0$ as *formulas*, and sets $K \subseteq \mathbb{R}_+^{2^{[n]}}$ as *classes of models*.

Example 4.3. *The following is a Shannon inequality, called* Shearer's inequality*:*

$$h(XY) + h(YZ) + h(ZX) - 2h(XYZ) \geq 0 \tag{17}$$

To prove it, we apply submodularity twice, underlining the affected terms:

$$\begin{aligned}
&\underline{h(XY) + h(YZ)} + h(ZX) \\
&\quad \geq h(XYZ) + \underline{h(Y) + h(ZX)} \\
&\quad \geq h(XYZ) + h(XYZ) + h(\emptyset) = 2h(XYZ)
\end{aligned}$$

Equivalently, we observe that (17) *is the sum of the following basic Shannon inequalities:*

$$\begin{aligned}
h(XY) + h(YZ) - h(Y) - h(XYZ) &\geq 0 \\
h(Y) + h(ZX) - h(\emptyset) - h(XYZ) &\geq 0 \\
h(\emptyset) &= 0
\end{aligned}$$

We will prove shortly (Theorem 5.2 below) that one can decide in time exponential in n whether an inequality is valid for all polymatroids. In contrast, it is an open problem whether entropic validity is decidable.

4.2 The Entropic Bound and the Polymatroid Bound

The general framework for computing a bound on a query's output uses degree constraints, which, in turn, correspond to conditional entropies. We define these two notions first.

We write $\boldsymbol{UV}$ for set $\boldsymbol{U} \cup \boldsymbol{V}$. Given $\boldsymbol{h} \in \mathbb{R}_+^{2^{[n]}}$, define:

$$h(\boldsymbol{V}|\boldsymbol{U}) \stackrel{\text{def}}{=} h(\boldsymbol{UV}) - h(\boldsymbol{U}) \tag{18}$$

$\boldsymbol{U}, \boldsymbol{V}$ need not be disjoint, and $h(\boldsymbol{V}|\boldsymbol{U}) = h(\boldsymbol{V} - \boldsymbol{U}|\boldsymbol{U})$; for example, $h(XY|X) = h(Y|X)$. If $h(\boldsymbol{V}|\boldsymbol{U}) = 0$ then we say that $\boldsymbol{h}$ satisfies the functional dependency $\boldsymbol{U} \to \boldsymbol{V}$, and we write $\boldsymbol{h} \models \boldsymbol{U} \to \boldsymbol{V}$. Lee [Lee87a] proved that, if R is a relation instance with attributes $\boldsymbol{X}$, $p : R \to [0, 1]$ is a probability distribution, and $\boldsymbol{h}$ is its entropic vector, then $R \models \boldsymbol{U} \to \boldsymbol{V}$ iff $\boldsymbol{h} \models \boldsymbol{U} \to \boldsymbol{V}$. For a simple illustration, referring to Fig. 2, both R and its entropy $\boldsymbol{h}$ satisfy the FDs $XY \to Z$, $XZ \to Y$, and $YZ \to X$: for example XY is a key (all 4 tuples have distinct values XY) and $h(Z|XY) = h(XYZ) - h(XY) = 2 - 2 = 0$.

Fix $\boldsymbol{U}$, and denote by $h(-|\boldsymbol{U}) : 2^{\boldsymbol{X}-\boldsymbol{U}} \to \mathbb{R}_+$ the function $\boldsymbol{V} \mapsto h(\boldsymbol{V}|\boldsymbol{U})$. If $\boldsymbol{h}$ is a polymatroid, then $h(-|\boldsymbol{U})$ is also a polymatroid, called the conditional polymatroid. If $\boldsymbol{h}$ is an entropic vector, then, surprisingly, $\boldsymbol{h}(-|\boldsymbol{U})$ is not necessarily entropic (as we will see later in Sec. 9.3), yet the name *conditional entropy* is justified by the following. Suppose $\boldsymbol{h}$ is associated to (R, p). Fix an outcome $\boldsymbol{u} \in \mathsf{Dom}^{\boldsymbol{U}}$, consider the random variable $\boldsymbol{V}$ conditioned on $\boldsymbol{U} = \boldsymbol{u}$, and denote its entropy by $h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})$. Then:

$$h(\boldsymbol{V}|\boldsymbol{U}) = \mathop{\mathbb{E}}_{\boldsymbol{u}} [h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})] \tag{19}$$

In other words, $h(\boldsymbol{V}|\boldsymbol{U})$ equals the *expectation* over the outcomes $\boldsymbol{u}$ of the (standard) entropy of the random variable $\boldsymbol{V}$ conditioned on $\boldsymbol{U} = \boldsymbol{u}$. The proof of identity (19) consists of applying directly the definition of the entropy given in Eq. (12).

When proving Shannon inequalities it is sometimes convenient to write the submodularity inequality as $h(\boldsymbol{V}|\boldsymbol{U}) \geq h(\boldsymbol{V}|\boldsymbol{UW})$.[1] In other words, conditioning on more variables can only decrease the entropy.

Example 4.4. *We illustrate a simple Shannon inequality with conditionals:*

$$\begin{aligned}
\underline{h(XY)+h(YZ)} + h(ZU) + h(U|XZ) + h(X|YU) \geq& \\
\geq & h(XYZ) + \underline{h(Y) + h(ZU)} + h(U|XZ) + h(X|YU) \\
\geq & h(XYZ) + h(YZU) + \underline{h(U|XZ)} + \underline{h(X|YU)} \\
\geq & h(XYZ) + h(YZU) + h(U|XYZ) + h(X|YZU) \\
= & 2h(XYZU)
\end{aligned}$$

Next, we define degrees of a relation instance $R \subseteq \mathsf{Dom}^{\boldsymbol{X}}$. Given subsets $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$, and $\boldsymbol{u} \in \mathsf{Dom}^{\boldsymbol{U}}$, the *$\boldsymbol{V}$-degree of $\boldsymbol{U} = \boldsymbol{u}$ in R* is the number of distinct values $\boldsymbol{v}$ that occur in R together with $\boldsymbol{u}$; the *max-$\boldsymbol{V}$-degree of $\boldsymbol{U}$* is the maximum degree over all values $\boldsymbol{u}$. Formally:

$$\begin{aligned}
\mathtt{deg}_R(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}) \stackrel{\text{def}}{=} & |\{\boldsymbol{v} \mid (\boldsymbol{u}, \boldsymbol{v}) \in \Pi_{\boldsymbol{UV}}(R)\}| \\
\mathtt{deg}_R(\boldsymbol{V}|\boldsymbol{U}) \stackrel{\text{def}}{=} & \max_{\boldsymbol{u}} (\mathtt{deg}_R(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}))
\end{aligned}$$

We note that $\mathtt{deg}_R(\boldsymbol{V}|\boldsymbol{U}) \geq 1$ (since we assumed $R \neq \emptyset$), and equality holds iff R satisfies the functional dependency $\boldsymbol{U} \to \boldsymbol{V}$.

[1] This is equivalent to $h(\boldsymbol{UV}) - h(\boldsymbol{U}) \geq h(\boldsymbol{UVW}) - h(\boldsymbol{UW})$; when $\boldsymbol{V} \cap \boldsymbol{W} = \emptyset$ then this is a submodularity inequality.

Definition 4.5. *Fix a relation $R(\boldsymbol{X})$. A* degree statistics*, or a* statistics *in short, σ, is an expression of the form $\sigma = (\boldsymbol{V}|\boldsymbol{U})$ where $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$; when $\boldsymbol{U} = \emptyset$ then we call σ a* cardinality statistics*, and write it as $(\boldsymbol{V})$. If Σ is a set of statistics, then we call $\boldsymbol{B} = (B_\sigma)_{\sigma\in\Sigma}$, where $B_\sigma \geq 1$,* statistics values *associated to Σ. The* log-statistics values *are $\boldsymbol{b} = \log \boldsymbol{B} = (b_\sigma := \log B_\sigma)_{\sigma\in\Sigma}$.*

We abbreviate $h(\boldsymbol{V}|\boldsymbol{U})$ and $\texttt{deg}_R(\boldsymbol{V}|\boldsymbol{U})$ with $h(\sigma)$ and $\texttt{deg}_R(\sigma)$ respectively. If Σ is a set of statistics, then a Σ-*inequality* is an inequality of the following form:

$$\sum_{\sigma\in\Sigma} w_\sigma h(\sigma) \geq h(\boldsymbol{X}) \tag{20}$$

where $\boldsymbol{w} = (w_\sigma)_{\sigma\in\Sigma}$ are nonnegative weights.

Fix a hypergraph $\mathcal{H} = (\boldsymbol{X}, E)$. We say that Σ is *guarded by* $\mathcal{H}$ if, for every $\sigma = (\boldsymbol{V}|\boldsymbol{U})$ in Σ, there exists a hyperedge $\boldsymbol{Y}_\sigma \in E$ such that $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{Y}_\sigma$; we call $\boldsymbol{Y}_\sigma$ the *guard* of σ. When $\mathcal{H}$ is the hypergraph of a query Q, then we say that Σ is guarded by Q, and that R_σ is the guard of σ. The following theorem establishes the key connection between information inequalities and query output size. The proof relies on a method originally introduced by Chung et al. for a combinatorial problem [CGFS86], and adapted by Grohe and Marx for constraint satisfaction [GM14], then by Atserias, Grohe, and Marx for their AGM bound [AGM13].

Theorem 4.6. *Assume Σ is guarded by Q. If the Σ-inequality* (20) *is valid for entropic vectors then:*

$$|Q| \leq \prod_{\sigma\in\Sigma} \texttt{deg}_{R_\sigma}^{w_\sigma}(\sigma) \tag{21}$$

where R_σ is the guard of $\sigma \in \Sigma$.

Proof. Fix a database instance $\boldsymbol{D}$, and let $\boldsymbol{h}$ be the entropic vector associated to the relation $Q(\boldsymbol{D})$; by uniformity, $h(\boldsymbol{X}) = \log|Q(\boldsymbol{D})|$. If $\sigma = (\boldsymbol{V}|\boldsymbol{U}) \in \Sigma$ has guard R_σ, then:

$$\begin{aligned} h(\sigma) =& \mathop{\mathbb{E}}_{\boldsymbol{u}}[h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})] \leq \max_{\boldsymbol{u}} h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}) \\ \leq& \max_{\boldsymbol{u}} \log \deg_{R_\sigma}(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}) \\ =& \log \deg_{R_\sigma}(\boldsymbol{V}|\boldsymbol{U}) = \log \deg_{R_\sigma}(\sigma) \end{aligned}$$

Using (20) we derive:

$$\sum_\sigma w_\sigma \log \deg_{R_\sigma}(\sigma) \geq \sum_\sigma w_\sigma h(\sigma) \geq h(\boldsymbol{X}) = \log|Q|$$

□

Inequality (20) is similar to the AGM inequality (6). Next, we proceed as we did for the AGM bound: fix numerical values for the statistics, then minimize the bound over all valid $\boldsymbol{w}$'s. We say that a database instance $\boldsymbol{D}$ *satisfies* the statistics $\Sigma, \boldsymbol{B}$, in notation $\boldsymbol{D} \models (\Sigma, \boldsymbol{B})$, if $\texttt{deg}_{R_\sigma^{\boldsymbol{D}}}(\sigma) \leq B_\sigma$ for all $\sigma \in \Sigma$. Similarly, we say that a vector $\boldsymbol{h}$ satisfies the log-statistics $\Sigma, \boldsymbol{b}$ if $h(\sigma) \leq b_\sigma$ for all σ. We define:

Definition 4.7 (Query Upper Bound)**.** *Let $\Sigma, \boldsymbol{B}$ be statistics values, guarded by the query Q. Fix some set $K \subseteq \mathbb{R}^{2^{[n]}}$. The* Upper Bound *w.r.t. K of the query Q is:*

$$\textit{U-Bound}_K(Q, \Sigma, \boldsymbol{B}) \stackrel{def}{=} \inf_{\boldsymbol{w}:K\models Eq.(20)} \prod_{\sigma\in\Sigma} B_\sigma^{w_\sigma}$$

The entropic upper bound *is $\textit{U-Bound}_{\Gamma_n^*}$ and the* polymatroid upper bound *is $\textit{U-Bound}_{\Gamma_n}$.*

Sometimes it is more convenient to use the logarithm and define:

$$\text{Log-U-Bound}_K(Q, \Sigma, \boldsymbol{b}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{w}: K \models \text{Eq.}(20)} \sum_{\sigma \in \Sigma} w_\sigma b_\sigma \tag{22}$$

Corollary 4.8. *The following hold:*

- *U-Bound*$_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{B}) \leq$ *U-Bound*$_{\Gamma_n}(Q, \Sigma, \boldsymbol{B})$, *and*
- *If* $\boldsymbol{D} \models (\Sigma, \boldsymbol{B})$ *then* $|Q(\boldsymbol{D})| \leq$ *U-Bound*$_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{B})$.

The first item is by $\Gamma_n^* \subseteq \Gamma_n$, the second is by Th. 4.6.

Let's compare these bound with the AGM bound (7). There, we had to minimize an expression where $\boldsymbol{w}$ ranged over the fractional edge covers of the query's hypergraph. In our new setting, $\boldsymbol{w}$ ranges over valid Σ-inequalities, a much more difficult task. To compute the polymatroid bound, $\boldsymbol{w}$ ranges over Σ-inequalities valid for polymatroids, and we will show in Th. 5.2 that this bound can be computed in time exponential in n. However, in order to compute the entropic bound, $\boldsymbol{w}$ needs to define a valid entropic inequality, and it is currently open whether this bound is computable. On the other hand, we will prove that the entropic bound is asymptotically tight, while the polymatroid bound is not. Thus, we are faced with a difficult choice, between and exact but non-computable bound, or a computable but inexact bound. This justifies examining non-trivial special cases of statistics Σ when these two bounds agree. We illustrate the entropic upper bound with an example.

Example 4.9. *Consider the following conjunctive query:*

$$\begin{aligned} Q(X, Y, Z, U) = & R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \\ & \wedge A(X, Z, U) \wedge B(X, Y, U) \end{aligned}$$

Suppose that we are given the following set of statistics $\Sigma = \{(XY), (YZ), (ZU), (U|XZ), (X|YU)\}$. *In other words, we have bounds on the cardinalities of* R, S, T, *but not of* A, B, *hence we can assume that* $|A| = |B| = \infty$. *Instead, we have the statistics* $\texttt{deg}_A(U|XZ)$ *and* $\texttt{deg}_B(X|YU)$. *The AGM bound (6) is* $|Q| \leq |R| \cdot |T|$, *because the only fractional edge cover whose bound is* $< \infty$ *is* $w_R = w_T = 1$ *and* $w_S = w_A = w_B = 0$.

The polymatroid bound follows from these Σ*-inequalities:*

$$\begin{aligned} & h(XY) + h(YZ) + h(ZU) + h(U|XZ) + h(X|YU) \geq \\ & \qquad\qquad\qquad\qquad \geq 2h(XYZU) \\ & h(XY) + h(ZU) \geq h(XYZU) \\ & h(XY) + h(YZ) + h(U|XZ) \geq h(XYZU) \\ & h(YZ) + h(ZU) + h(X|YU) \geq h(XYZU) \end{aligned}$$

We proved the first inequality in Example 4.4, while the other three are immediate. They imply:

$$\begin{aligned} |Q| \leq & \big(|R| \cdot |S| \cdot |T| \cdot \texttt{deg}_A(U|XZ) \cdot \texttt{deg}_B(X|YU)\big)^{1/2} \\ |Q| \leq & |R| \cdot |T| \\ |Q| \leq & |R| \cdot |S| \cdot \texttt{deg}_B(X|YU) \\ |Q| \leq & |S| \cdot |T| \cdot \texttt{deg}_A(U|XZ) \end{aligned}$$

The AGM bound is the second inequality. We show in Appendix .1 that the entropic bound is the minimum over all four expressions above. (This requires proving that there is no "better" inequality

$$R = \begin{array}{|ccccccc|c} \hline X_1 & \dots & X_{i-1} & X_i & X_{i+1} & \dots & X_n & p \\ \hline 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1/2 \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 & 1/2 \\ \hline \end{array} \qquad h^{X_i}(\boldsymbol{U}) = \begin{cases} 1 & \text{if } X_i \in \boldsymbol{U} \\ 0 & \text{if } X_i \notin \boldsymbol{U} \end{cases}$$

Figure 4: A relation R with two tuples that agree on all attributes, except X_i. Its entropic vector is called the *basic modular function*, $\boldsymbol{h}^{X_i}$; it is used in Th. 4.10, and discussed in more detail in Sec. 6.

that gives us a smaller bound.) Since all four inequalities are Shannon inequalities, it follows that, in this case, the entropic bound is equal to the polymatroid bound.

When XZ is a key in A, and YU is a key in B, then the polymatroid bound simplifies to:

$$|Q| \leq \min((|R| \cdot |S| \cdot |T|)^{1/2}, |R| \cdot |T|, |R| \cdot |S|, |S| \cdot |T|)$$

Special Case: Cardinality Constraints We show next that the AGM bound is a special case where the polymatroid and entropic bounds coincide; we will see a more general setting when this happens in Sec. 6. Assume Σ is restricted to cardinality constraints, and assume for simplicity that Σ has exactly one cardinality constraint $(\boldsymbol{Y}_j)$ for each relation $R_j(\boldsymbol{Y}_j)$ in the query; $\Sigma = \{(\boldsymbol{Y}_1), \dots, (\boldsymbol{Y}_m)\}$. The Σ-inequality (20) becomes:

$$\sum_{j=1,m} w_j h(\boldsymbol{Y}_j) \geq h(\boldsymbol{X}) \tag{23}$$

When $w_1 = \dots = w_m$ then (23) is called Shearer's inequality.

Theorem 4.10. *The following are equivalent:*

(i) Inequality (23) *is valid for polymatroids.*

(ii) Inequality (23) *is valid for entropic functions.*

(iii) The weights $\boldsymbol{w}$ form a fractional edge cover of the hypergraph $(\boldsymbol{X}, \{\boldsymbol{Y}_j \mid j = 1, m\})$.

Proof. (i) $\Rightarrow$ (ii) is immediate. For (ii) $\Rightarrow$ (iii), assume that (23) holds for all entropic vectors, and consider any variable $X_i \in \boldsymbol{X}$, for $i = 1, n$. Consider the *basic modular entropic function* $\boldsymbol{h}^{X_i}$ shown in Fig. 4. Since $\boldsymbol{h}^{X_i}$ satisfies inequality (23), it follows that $\sum_{j=1,m: X_i \in \boldsymbol{Y}_j} w_j \geq 1$ (because $h^{X_i}(\boldsymbol{Y}_j) = 1$ iff $X_i \in \boldsymbol{Y}_j$), proving that $\boldsymbol{w}$ is a fractional edge cover.

It remains to prove the implication (iii) $\Rightarrow$ (i). This is a well known result, and it admits multiple proofs (we give a second proof in Sec. 7.2). Here, we will prove the inequality by induction on n:

- Partition the set of indices j into J_0 and J_1:
  $J_0 \stackrel{\text{def}}{=} \{j \mid X_n \notin \boldsymbol{Y}_j\}$, $J_1 \stackrel{\text{def}}{=} \{j \mid X_n \in \boldsymbol{Y}_j\}$.
- If $j \in J_1$ then write $h(\boldsymbol{Y}_j) = h(X_n) + h(\boldsymbol{Y}_j - X_n | X_n)$. Note that $\sum_{j \in J_1} w_j \geq 1$ because X_n is covered.
- If $j \in J_0$ then write $h(\boldsymbol{Y}_j) \geq h(\boldsymbol{Y}_j - X_n | X_n)$.

Using the steps above we obtain:

$$\begin{aligned}\sum_j w_j h(\boldsymbol{Y}_j) \geq & h(X_n) + \sum_j w_j h(\boldsymbol{Y}_j - X_n|X_n) \\ \geq & h(X_n) + h(\boldsymbol{X} - X_n|X_n) = h(\boldsymbol{X})\end{aligned}$$

The last line used induction on the polymatroid $h(-|X_n)$. □

It follows immediately that the AGM bound, the entropic bound, and the polymatroid bound coincide in the simple case when the statistics are restricted to the cardinalities of the input relations.

Discussion Degree constraints occur often and naturally in database applications. For example, if a relation $R_j(\boldsymbol{Y}_j)$ has a key $\boldsymbol{U}$, then $\texttt{deg}_{R_j}(\boldsymbol{Y}_j|\boldsymbol{U}) = 1$. In practice almost every relation has a key, so this case is very common. In other cases some cardinality constraints can be obtained directly from the application. For example, suppose that in a database of customers we require that no customer may have more than 10 credit cards, which naturally leads to a max-degree constraint. Such constraints are used in some modern systems, for example in *scale-independent query processing* [AFP+09, ACK+11, ALK+13].

5 The Worst-Case Instance

Informally, we call a database instance $\boldsymbol{D}$ a *worst-case instance* if it satisfies the given statistics, and the query's output is as large as, or approaches asymptotically (in a sense to be made precise), the entropic upper bound. We will show that such a worst-case instance exists, proving that the entropic bound is *asymptotically tight*, which is a weaker notion of tightness than for the AGM bound in Th. 3.4. We will also show that, in general, the polymatroid bound is not tight, even for this weaker notion of tightness.

To construct the worst-case instance we need a *dual* definition of the entropic and polymatroid bounds. We define them directly using log-version:

Definition 5.1 (Query Lower Bound). *Fix log-statistics* $\Sigma, \boldsymbol{b}$. *For any set* $K \subseteq \mathbb{R}^{2^{[n]}}$, *the* Log Lower Bound *w.r.t.* K *is:*

$$\textit{Log-L-Bound}_K(Q, \Sigma, \boldsymbol{b}) \stackrel{def}{=} \sup_{\boldsymbol{h} \in K : \boldsymbol{h} \models (\Sigma, \boldsymbol{b})} h(\boldsymbol{X}) \tag{24}$$

As before, the entropic log-lower bound *is* $\textit{Log-L-Bound}_{\Gamma_n^*}$, *and the* polymatroid log-lower bound *is* $\textit{Log-L-Bound}_{\Gamma_n}$.

The log-lower bound asks us to find a vector $\boldsymbol{h}$ that satisfies all log-statistics $\Sigma, \boldsymbol{b}$, and where $\boldsymbol{h}(\boldsymbol{X})$ is as large as possible. We call $\boldsymbol{h}$ a *worst case* entropic vector, or the worst case polymatroid respectively. Using $\boldsymbol{h}$, we would like to construct a worst-case database instance $\boldsymbol{D}$, that satisfies $\Sigma, \boldsymbol{B}$, and $\log |Q(\boldsymbol{D})| = h(\boldsymbol{X})$. The difficulty lies in the fact that, when $\boldsymbol{h}$ is a polymatroid then such a database may not exists general, and when $\boldsymbol{h}$ is an entropic vector, then it may be realized by a probability space that is non-uniform, hence we cannot use it to construct $\boldsymbol{D}$. We start by observing the following, which is easy to check:

$$\text{Log-L-Bound}_K(Q, \Sigma, \boldsymbol{b}) \leq \text{Log-U-Bound}_K(Q, \Sigma, \boldsymbol{b}) \tag{25}$$

Indeed, if $\boldsymbol{h} \in K$ satisfies $\boldsymbol{h} \models (\Sigma, \boldsymbol{b})$, and $\boldsymbol{w}$ satisfies $\forall \boldsymbol{h} \in K, \sum_\sigma w_\sigma h(\sigma) \geq h(\boldsymbol{X})$, then $h(\boldsymbol{X}) \leq \sum_\sigma w_\sigma h(\sigma) \leq \sum_\sigma w_\sigma b_\sigma$, and the claim follows from $\text{Log-L-Bound}_K = \sup_{\boldsymbol{h}} h(\boldsymbol{X}) \leq \text{Log-U-Bound}_K = \inf_{\boldsymbol{w}} \sum_\sigma w_\sigma b_\sigma$.

When $K = \Gamma_n$, then [KNS17] showed that the two bounds are equal. We prove a slightly more general statement:

Theorem 5.2. *Suppose the set K is defined by linear constraints: $K = \{\boldsymbol{h} \in \mathbb{R}_+^{2^{[n]}} \mid \boldsymbol{M} \cdot \boldsymbol{h} \geq 0\}$, where $\boldsymbol{M}$ is some matrix.*[2] *Then, Log-L-Bound$_K$ and Log-U-Bound$_K$ are defined by a pair of primal/dual linear programs, with a number of variables exponential in n; the expressions* $\inf, \sup$ *in* (22), (24) *can be replaced by* $\min, \max$*; and Eq.* (25) *becomes an equality, $h^*(\boldsymbol{X}) = \sum_\sigma w^*_\sigma b_\sigma$, where $\boldsymbol{h}^*$, $\boldsymbol{w}^*$ are the optimal solutions of the primal and dual program respectively.*

Proof. Denote $s = |\Sigma|$, and let $\boldsymbol{A}$ be the $s \times 2^n$ matrix that maps $\boldsymbol{h}$ to the vector $\boldsymbol{A} \cdot \boldsymbol{h} = (h(\sigma))_{\sigma \in \Sigma} \in \mathbb{R}^s$. Let $\boldsymbol{c} \in \mathbb{R}^{2^n}$ be the vector $c_{\boldsymbol{X}} = 1$, $c_{\boldsymbol{U}} = 0$ for $\boldsymbol{U} \neq \boldsymbol{X}$. The two bounds are the optimal solutions to the following pair of primal/dual linear programs:

Log-L-Bound$_K$	Log-U-Bound$_K$
Maximize $\boldsymbol{c}^T \cdot \boldsymbol{h}$	Minimize $\boldsymbol{w}^T \cdot \boldsymbol{b}$
where $\boldsymbol{A} \cdot \boldsymbol{h} \leq \boldsymbol{b}$	where $\boldsymbol{w}^T \cdot \boldsymbol{A} - \boldsymbol{c}^T \geq \boldsymbol{u}^T \cdot \boldsymbol{M}$
$-\boldsymbol{M} \cdot \boldsymbol{h} \leq 0$	

where the primal variables are $\boldsymbol{h} \geq 0$, and the dual variables are $\boldsymbol{w}, \boldsymbol{u} \geq 0$; the reader may check that the two programs above form indeed a primal/dual pair. Log-L-Bound$_K$ is by definition the optimal value of the program above. We prove that Log-U-Bound$_K$ is the value of the dual. First, observe that the Σ-inequality (20) is equivalent to $(\boldsymbol{w}^T \cdot \boldsymbol{A} - \boldsymbol{c}^T) \cdot h \geq 0$. We claim that this inequality holds $\forall \boldsymbol{h} \in K$ iff there exists $\boldsymbol{u}$ s.t. $(\boldsymbol{w}, \boldsymbol{u})$ is a feasible solution to the dual. For that consider the following primal/dual programs with variables $\boldsymbol{h} \geq 0$ and $\boldsymbol{u} \geq 0$ respectively:

Minimize $(\boldsymbol{w}^T \cdot \boldsymbol{A} - \boldsymbol{c}^T) \cdot \boldsymbol{h}$	Maximize 0
where $\boldsymbol{M} \cdot \boldsymbol{h} \geq 0$	where $\boldsymbol{u}^T \cdot \boldsymbol{M} \leq \boldsymbol{w}^T \boldsymbol{A} - \boldsymbol{c}^T$

The primal (left) has optimal value 0 iff the inequality $(\boldsymbol{w}^T \cdot \boldsymbol{A} - \boldsymbol{c}^T) \cdot \boldsymbol{h} \geq 0$ holds forall $\boldsymbol{h} \in K$; otherwise its optimal is $-\infty$. The dual (right) has optimal value 0 iff there exists a feasible solution $\boldsymbol{u}$; otherwise its optimal is $-\infty$. Strong duality proves our claim. $\square$

When $K = \Gamma_n^*$, then the two terms in (25) may no longer be equal in general, but we prove that they are equal asymptotically. Call a *k-amplification* of a set of log-statistics $\Sigma, \boldsymbol{b}$ the log-statistics $\Sigma, k\boldsymbol{b}$, where k is a natural number. Observe that the entropic log-upper bound increases linearly with the k-amplification:

$$\text{Log-U-Bound}_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b}) = k\text{Log-U-Bound}_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{b}) \tag{26}$$

The lower bound increases at least linearly, $\text{Log-L-Bound}_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b}) \geq k\text{Log-L-Bound}_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{b})$, because of the following proposition:

Proposition 5.3. *If $\boldsymbol{h}_1, \boldsymbol{h}_2$ are two entropic vectors, then so is $\boldsymbol{h}_1 + \boldsymbol{h}_2$.*

Proof. Suppose $\boldsymbol{h}_1, \boldsymbol{h}_2$ are realized by two finite probability spaces $(R_1, p_1), (R_2, p_2)$. Then their sum is realized by $(R_1 \otimes R_2, p)$ (see Def. 2.1), where $p(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2) \stackrel{\text{def}}{=} p(\boldsymbol{x}_1) \cdot p(\boldsymbol{x}_2)$. $\square$

We prove:

[2] Equivalently: K is a polyhedral cone, reviewed in Sec. 9.3.

Theorem 5.4. *Fix any $Q, \Sigma, \boldsymbol{b}$. The entropic upper and lower bounds are* asymptotically equal*, in the following sense:*

$$\sup_k \frac{\textit{Log-L-Bound}_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b})}{\textit{Log-U-Bound}_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b})} = 1 \tag{27}$$

The proof of this result, which appears to be novel, requires a discussion of the set of *almost-entropic functions*, and we defer this to Sec. 9.

Finally, we can answer the central question in this section: the entropic bound is tight asymptotically, while the polymatroid bound is not.

Theorem 5.5. *(1) For any query Q and statistics $\Sigma, \boldsymbol{B}$, the entropic bound is asymptotically tight, in the following sense:*

$$\sup_k \frac{\sup_{\boldsymbol{D}:\boldsymbol{D}\models \boldsymbol{B}^k} \log |Q(\boldsymbol{D})|}{\textit{Log-L-Bound}_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b})} = 1 \tag{28}$$

(2) The polymatroid bound is not asymptotically tight: there exists a query Q and statistics $\Sigma, \boldsymbol{B}$ such that:

$$\sup_k \frac{\sup_{\boldsymbol{D}:\boldsymbol{D}\models \boldsymbol{B}^k} \log |Q(\boldsymbol{D})|}{\textit{Log-L-Bound}_{\Gamma_n}(Q, \Sigma, k\boldsymbol{b})} \leq \frac{43}{44} \tag{29}$$

Here $\boldsymbol{B}^k \stackrel{def}{=} (B_\sigma^k)_{\sigma\in\Sigma}$, and $\boldsymbol{b} \stackrel{def}{=} \log \boldsymbol{B}$. Moreover, this property holds even if $\Sigma, \boldsymbol{B}$ consists only of cardinality constraints and functional dependencies, i.e. $\forall \sigma \in \Sigma$ either $\sigma = (\boldsymbol{V})$ or $B_\sigma = 1$.

Equivalently, Eq. (28) says that, for any statistics $\Sigma, \boldsymbol{B}$, if we allow a sufficiently large amplification $\Sigma, \boldsymbol{B}^k$, then there exists a worst-case instance $\boldsymbol{D}$ satisfying the amplified statistics such that $\log |Q(\boldsymbol{D})|$ approaches $\text{L-Bound}_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{B}^k)$. Notice that this is a weaker notion of tightness than for the AGM bound in Theorem 3.4. There, tightness referred to the ratio betwen the lower and upper bound, while here tightness refers to the ratio of their logarithms, a weaker notion. Eq. (29) says that even this weaker tightness fails for the polymatroid bound.

The first proof of asymptotic tightness was given by Gogacz and Torunczyk [GT17], for the restricted case when the statistics are either cardinalities, or functional dependencies. The general case was proven in [KNS17]. Both results were stated slightly differently from ours, by using almost-entropic functions. We prefer to state our result in terms of the entropic functions, since it is more natural, and defer the discussion of almost entropic functions to Sec. 9, where they have a very natural justification.

In the rest of this section we prove Theorem 5.5. We use this opportunity to continue our dive into the fascinating world of entropic functions, and non-Shannon inequalities, which are needed for the proof. However, the rest of this section is rather technical, and readers not interested in this background may safely skip the rest of this section, since we do not need it, except for the short introduction of mutual information.

5.1 Background: Non-Shannon Inequalities, Lattices, Groups

Mutual Information Given a vector $\boldsymbol{h} \in \mathbb{R}^{2^{[n]}}$ and three disjoint sets of variables $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W} \subseteq \boldsymbol{X}$, we denote by:

$$I_h(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U}) \stackrel{\text{def}}{=} h(\boldsymbol{U}\boldsymbol{V}) + h(\boldsymbol{U}\boldsymbol{W}) - h(\boldsymbol{U}) - h(\boldsymbol{U}\boldsymbol{V}\boldsymbol{V}) \tag{30}$$

When h is clear from the context, then we will drop the index h from I_h and write simply I. When h is an entropic vector, then $I_h(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U})$ is called the *mutual information of* $\boldsymbol{V},\boldsymbol{W}$ *conditioned on* $\boldsymbol{U}$. In that case, $I_h(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) = 0$ iff the probability space realizing h satisfies $\boldsymbol{V} \perp \boldsymbol{W}|\boldsymbol{U}$, meaning that $\boldsymbol{V},\boldsymbol{W}$ are independent conditioned on $\boldsymbol{U}$. With some abuse, we will call I_h a conditional mutual information even when $\boldsymbol{h}$ is a polymatroid. The following properties hold:

Proposition 5.6. *For any polymatroid h:*

- $I_h(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) \geq 0$ *(this is a submodularity inequality).*
- *The* chain rule *holds:*
  $I_h(\boldsymbol{U}\boldsymbol{V};\boldsymbol{W}|\boldsymbol{Z}) = I_h(\boldsymbol{U};\boldsymbol{W}|\boldsymbol{Z}) + I_h(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}\boldsymbol{Z})$.
- *An* elemental mutual information term *is an expression of the form $I_h(X_i;X_j|\boldsymbol{U})$, where $X_i \neq X_j \notin \boldsymbol{U}$ are single variables. Every mutual information $I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U})$ is the sum of elemental terms.*
- *For any subsets $\boldsymbol{U}_0 \subseteq \boldsymbol{U}$ and $\boldsymbol{V}_0 \subseteq \boldsymbol{V}$:*

$$I_h(\boldsymbol{U}_0;\boldsymbol{V}_0|\boldsymbol{Z}) \leq I_h(\boldsymbol{U};\boldsymbol{V}|\boldsymbol{Z}) \tag{31}$$

  If $I_h(\boldsymbol{U};\boldsymbol{V}|\boldsymbol{Z}) = 0$, then $I_h(\boldsymbol{U}_0;\boldsymbol{V}_0|\boldsymbol{Z}) = 0$.

Non-Shannon inequalities The first non-Shannon inequality was proven by Zhang and Yeung [ZY98]. We review it here, following the simplified presentation by Romashchenko [Rom22] (see also Csirmaz [Csi22]).

Zhang and Yeung [ZY98] proved the following:

Theorem 5.7. *The following is a non-Shannon inequality:*

$$\begin{aligned} I_h(X;Y) \leq & I_h(X;Y|A) + I_h(X;Y|B) + I_h(A;B) \\ & + I_h(X;Y|A) + I_h(A;Y|X) + I_h(A;X|Y) \end{aligned} \tag{32}$$

In other words, this inequality is valid for entropic vectors, but it cannot be proven using the basic Shannon inequalities, hence the term *non-Shannon*. The proof of this inequality, and that of several other non-Shannon inequalities proven after 1998, relies on the following *copy lemma*.

Lemma 5.8 (Copy Lemma)**.** *Let $\boldsymbol{X},\boldsymbol{Y}$ be two disjoint sets of variables, and let $\boldsymbol{h}$ be an entropic vector with variables $\boldsymbol{X}\boldsymbol{Y}$. Let $\boldsymbol{Y}'$ be fresh copies of the variables $\boldsymbol{Y}$. Thus, each variable $Y \in \boldsymbol{Y}$ has a copy $Y' \in \boldsymbol{Y}'$. Then there exists an entropic vector $\boldsymbol{h}'$ over variables $\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Y}'$ such that the following hold:*

$$\begin{aligned} & I_{h'}(\boldsymbol{Y};\boldsymbol{Y}'|\boldsymbol{X}) = 0 \\ \forall \boldsymbol{U} \subseteq \boldsymbol{X}\boldsymbol{Y}: \ & h'(\boldsymbol{U}) = h(\boldsymbol{U}) \\ \forall \boldsymbol{U}' \subseteq \boldsymbol{X}\boldsymbol{Y}': \ & h'(\boldsymbol{U}') = h(\boldsymbol{U}) \end{aligned}$$

We say that $\boldsymbol{Y}'$ is a copy of $\boldsymbol{Y}$ over $\boldsymbol{X}$*.*

The first inequality asserts $\boldsymbol{Y} \perp \boldsymbol{Y}'|\boldsymbol{X}$. The second asserts that $\boldsymbol{h},\boldsymbol{h}'$ agree on $\boldsymbol{X}\boldsymbol{Y}$. And the last equality asserts that $\boldsymbol{h}'$ on $\boldsymbol{X}\boldsymbol{Y}'$ is identical to $\boldsymbol{h}$ on $\boldsymbol{X}\boldsymbol{Y}$ up to the renaming of variables from $\boldsymbol{Y}'$ to $\boldsymbol{Y}$ (assuming $X' = X$ for $X \in \boldsymbol{U}$).

Proof. (of Lemma 5.8) Let p be a probability distribution of random variables $\boldsymbol{XY}$ that realizes the entropic vector $\boldsymbol{h}$. Define the following probability distribution p' of random variables $\boldsymbol{XYY'}$: the domains of the variables $\boldsymbol{Y'}$ is the same as that of $\boldsymbol{Y}$, and for all outcomes $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{y'})$, $p'(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{Y'} = \boldsymbol{y'}) \stackrel{\text{def}}{=} \frac{p(\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Y}=\boldsymbol{y}) p(\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Y}=\boldsymbol{y'})}{p(\boldsymbol{X}=\boldsymbol{x})}$. The claims in the lemma are easily verified. □

In general, the copy lemma does not hold for polymatroids, as we will see shortly. We prove now Zhang and Yeung's inequality (32). Start from the following Shannon inequality over 5 variables, X, Y, A, B, A':

$$\begin{aligned} I_h(X;Y) \leq & I_h(X;Y|A) + I_h(X;Y|B) + I_h(A;B) + \\ & + I_h(X;Y|A') + I_h(A';Y|X) + I_h(A';X|Y) + \\ & + 3I_h(A';AB|XY) \end{aligned} \tag{33}$$

While this is "only" a Shannon inequality, it is surprisingly difficult to prove; we invite the readers to try it themselves, but, for completeness, we give the proof in Appendix .2. Consider now an entropic vector $\boldsymbol{h}$ over four variables, X, Y, A, B. We apply the copy lemma, and copy AB over XY, resulting in an entropic vector h' with variables X, Y, A, B, A', B'. In particular, h' satisfies (33) (we don't use B'). Now we observe that (a) $I(A';AB|XY) = 0$, and (b) every occurrence of A' in the second line can be replaced by A; for example $I(X;Y|A') = I(X;Y|A)$, because $I(X;Y|A')$ is expressed in terms of $h(A'), h(XA'), h(YA'), h(XYA')$, which are equal to their copies $h(A), h(XA), h(YA), h(XYA)$. Thus, inequality (33) becomes (32), proving that (32) is valid for all entropic functions $\boldsymbol{h}$.

It remains to prove that (32) is not a Shannon inequality, and for that it suffices to describe one polymatroid that fails the inequality. To "see" this polymatroid, it is best to view it as being defined over a lattice. We take this opportunity to discuss another important concept: polymatroids on lattices.

Polymatroids on lattices A *polymatroid* on a lattice $(L, \preceq)$ is a function $h : L \to \mathbb{R}_+$ satisfying:

$$\begin{aligned} h(\hat{0}) &= 0 & & \\ h(x \vee y) &\geq h(x) & & \text{monotonicity} \\ h(x) + h(y) &\geq h(x \vee y) + h(x \wedge y) & & \text{submodularity} \end{aligned}$$

Let $\boldsymbol{X}$ be a set of n variables, and Σ a set of functional dependencies for $\boldsymbol{X}$. Recall from Sec. 2 that $(L_\Sigma, \subseteq)$ is the lattice of closed sets. If a (standard) polymatroid $h \in \Gamma_n$ satisfies the functional dependencies Σ, then it is not hard to see that its restriction to L_Σ is a polymatroid on L_Σ. Conversely, any polymatroid $\boldsymbol{h}$ on the lattice $(L_\Sigma, \subseteq)$ can be extended to a standard polymatroid $\bar{h} : 2^{\boldsymbol{X}} \to \mathbb{R}_+$ by setting $\bar{h}(\boldsymbol{U}) \stackrel{\text{def}}{=} h(\boldsymbol{U}^+)$, and, furthermore, $\boldsymbol{h}$ satisfies Σ. In short, there is a one-to-one correspondence between polymatroids satisfying a set of functional dependencies, and polymatroids defined on the associated lattice.

We now complete the proof of Theorem 5.7, by showing that inequality (32) does not hold for the polymatroid $\boldsymbol{h}$ in Fig. 5. To read the figure, recall that $h(\boldsymbol{U}) = h(\boldsymbol{U}^+)$ for any set $\boldsymbol{U}$. For example, $ABX^+ = ABXY$, therefore $h(ABX) = h(ABX^+) = h(ABXY) = 4$, and, also, $h(AB) = h(AB^+) = h(ABXY) = 4$. We check now that $\boldsymbol{h}$ violates the inequality (32):

$$\begin{aligned} & I(X;Y) = 1 \\ & I(X;Y|A) + I(X;Y|B) + I(A;B) = 0 + 0 + 0 \\ & I(X;Y|A) + I(A;Y|X) + I(A;X|Y) = 0 + 0 + 0 \end{aligned}$$

The LHS of (32) is 1, while the RHS is 0. This completes the proof of Theorem 5.7.

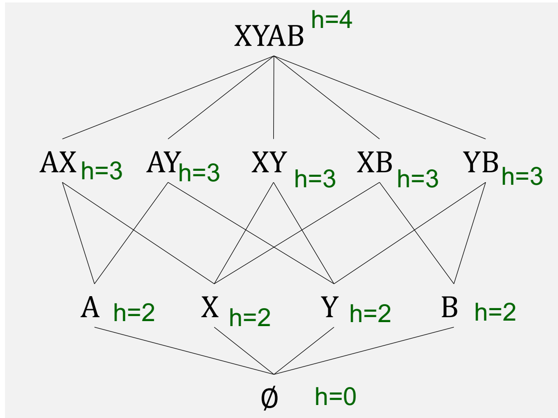


Figure 5: A lattice, and the polymatroid from [ZY98] defined on the lattice.

As a final comment, we note that it is instructive to check directly that the polymatroid in Fig. 5 fails to satisfy the copy lemma, without using Zhang and Yeung's inequality; we provide a direct proof in Appendix .2.

Group-theoretic characterization of information inequalities Chan and Yeung [CY02] described an elegant characterization of information inequalities in terms of group inequalities. Given a finite group G and a subgroup $G_1 \subseteq G$, a *left coset* is a set of the form aG_1, for some $a \in G$. By Lagrange's theorem, the set of left cosets, denoted G/G_1, forms a partition of G, and $|G/G_1| = |G|/|G_1|$. Fix n subgroups $G_1, \ldots, G_n$, and consider the relational instance:

$$R = \{(aG_1, \ldots, aG_n) \mid a \in G\} \tag{34}$$

whose set of attributes we identify, as usual, with $\boldsymbol{X} = \{X_1, \ldots, X_n\}$. Notice that $|R| = |G|/|\bigcap_{i=1,n} G_i|$. The entropic vector $\boldsymbol{h}$ associated to the relation R (Def. 4.1) is called a *group realizable entropic vector*, and the set of group realizable entropic vectors is denoted by $\Upsilon_n \subseteq \Gamma_n^*$, see Fig. 3. One can check that, for any subset of variables $\boldsymbol{U} \subseteq \boldsymbol{X}$, $h(\boldsymbol{U}) = \log |G|/|\bigcap_{X_i \in \boldsymbol{U}} G_i|$. The following was proven in [CY02]:

Theorem 5.9. *For any $\boldsymbol{h} \in \Gamma_n^*$ there exists a sequence $\boldsymbol{h}^{(r)} \in \Upsilon_n$, such that $\lim_{r\to\infty} \frac{1}{r}\boldsymbol{h}^{(r)} = \boldsymbol{h}$.*

It follows easily from the original proof that, if $\boldsymbol{h}$ satisfies a set of functional dependencies, then so do all functions $\boldsymbol{h}^{(r)}$, for $r \geq 0$; for completeness, we will include the argument in Appendix .3.

Open Problems Characterizing the valid entropic information inequalities is a major open problem. Matús [Mat07] proved that, for $n \geq 4$, there are infinitely many independent non-Shannon inequalities. Currently, the only techniques known for proving such inequalities consists of repeated applications of Shannon inequalities and the Copy Lemma.

A related open problem is the complexity of deciding Shannon inequalities: what is the complexity of checking $\Gamma_n \models \boldsymbol{c} \cdot \boldsymbol{h} \geq 0$, as a function of $||\boldsymbol{c}||_1$? It is implicit in the proof of Theorem 5.2 that this can be decided in time exponential in n, but the complexity in terms of $||\boldsymbol{c}||_1$ is open. More discussion can be found in [KKNS20]

5.2 The Entropic Bound Is Asymptotically Tight

We prove here Theorem 5.5 item (1). The plan is the following. We need to find a database $\boldsymbol{D}$ such that $\log |Q(D)|$ comes close to Log-L-Bound$_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{b})$. By definition, there exists $\boldsymbol{h} \in \Gamma_n^*$

s.t. $h(\boldsymbol{X})$ is close to Log-L-Bound$_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{b})$. We can't construct a database $\boldsymbol{D}$ out of $\boldsymbol{h}$, because the probability distribution realizing $\boldsymbol{h}$ may be non-uniform, instead we use Chan and Yeung's theorem to approximate $r\boldsymbol{h}$ by a group realizable vector $\boldsymbol{h}^{(r)}$, which is by definition associated to a relation instance. Hence, the need to amplify by the factor r. However, if we amplify, we don't know how Log-L-Bound$_{\Gamma_n^*}(Q, \Sigma, r\boldsymbol{b})$ grows. Here we use Theorem 5.4, showing that Log-L-Bound$_{\Gamma_n^*}$ and Log-U-Bound$_{\Gamma_n^*}$ are asymptotically equal, then use the fact that Log-U-Bound$_{\Gamma_n^*}$ is linear, see Eq. (26). We give the details next.

By Corollary 4.8, for all $k \in \mathbb{N}$:

$$\frac{\sup_{\boldsymbol{D}:\boldsymbol{D}\models\boldsymbol{B}^k} \log|Q(\boldsymbol{D})|}{\text{Log-U-Bound}_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b})} \leq 1$$

Together with Theorem 5.4 (Eq. (27)) this implies:

$$\sup_k \frac{\sup_{\boldsymbol{D}:\boldsymbol{D}\models\boldsymbol{B}^k} \log|Q(\boldsymbol{D})|}{\text{Log-L-Bound}_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b})} \leq 1$$

To prove equality, it suffices to show that, $\forall \varepsilon > 0$, $\exists k \in \mathbb{N}$ such that:

$$\log|Q(\boldsymbol{D})| \geq (1-\varepsilon)^4 \text{Log-L-Bound}_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b}) \tag{35}$$

Let $U \stackrel{\text{def}}{=}$ Log-U-Bound$_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{b})$. We will assume that Log-U-Bound$_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{b})$ is finite; otherwise, we let U be an arbitrarily large number and the proof below requires only minor adjustments, which we omit. We will assume w.l.o.g. that $U > 0$. Recall that Log-U-Bound$_{\Gamma_n^*}$ is linear (26). We prove:

Claim 1. *For all $\varepsilon > 0$, there exists $k \in \mathbb{N}$, and a database $\boldsymbol{D}$ such that $\boldsymbol{D} \models (\Sigma, \boldsymbol{B}^k)$ and $\log|Q(\boldsymbol{D})| \geq (1-\varepsilon)^4 kU$*

Eq. (35) follows from $kU =$ Log-U-Bound$_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b}) \geq$ Log-L-Bound$_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b})$. It remains to prove Claim 1. Since Log-L-Bound$_{\Gamma_n^*}$ and Log-U-Bound$_{\Gamma_n^*}$ are asymptotically equal (27), there exists $k_0 \in \mathbb{N}$ such that

$$\begin{aligned} &\text{Log-L-Bound}_{\Gamma_n^*}(Q, \Sigma, k_0\boldsymbol{b}) \geq \\ &\geq (1-\varepsilon)\text{Log-U-Bound}_{\Gamma_n^*}(Q, \Sigma, k_0\boldsymbol{b}) = (1-\varepsilon)k_0 U \end{aligned} \tag{36}$$

By the definition of Log-L-Bound$_{\Gamma_n^*}(Q, \Sigma, k_0\boldsymbol{b})$ in (24), there exists $\boldsymbol{h} \in \Gamma_n^*$ such that:

$$\begin{aligned} h(\boldsymbol{X}) \geq& (1-\varepsilon)\text{Log-L-Bound}_{\Gamma_n^*}(Q, \Sigma, k_0\boldsymbol{b}) \geq (1-\varepsilon)^2 k_0 U \\ h(\sigma) \leq& k_0 b_\sigma, \quad \forall \sigma \in \Sigma \end{aligned}$$

At this point we need the following Slack Lemma:

Lemma 5.10 (Slack Lemma)**.** *For every $\boldsymbol{h} \in \Gamma_n^*$ and every $\varepsilon \in [0,1]$, there exists $k \in \mathbb{N}$ and $\boldsymbol{h}' \in \Gamma_n^*$ such that:*

$$\begin{aligned} \boldsymbol{h}' \geq& (1-\varepsilon)k\boldsymbol{h} \\ \forall \boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}: \quad h'(\boldsymbol{V}|\boldsymbol{U}) \leq& (1-\varepsilon/2)kh(\boldsymbol{V}|\boldsymbol{U}) \end{aligned}$$

Proof. Assume w.l.o.g. that $\varepsilon > 0$, and set $k \stackrel{\text{def}}{=} \lceil \frac{1}{\varepsilon} \rceil$ and $\boldsymbol{h}' \stackrel{\text{def}}{=} (k-1)\boldsymbol{h}$. Then $\frac{1}{\varepsilon} \leq k \leq 1 + \frac{1}{\varepsilon} \leq \frac{2}{\varepsilon}$ which implies $\varepsilon/2 \leq \frac{1}{k} \leq \varepsilon$. We have:

$$\boldsymbol{h}' = \left(1 - \frac{1}{k}\right) k\boldsymbol{h} \geq (1-\varepsilon)k\boldsymbol{h}$$
$$\boldsymbol{h}'(\boldsymbol{V}|\boldsymbol{U}) = \left(1 - \frac{1}{k}\right) k\boldsymbol{h}(\boldsymbol{V}|\boldsymbol{U}) \leq (1-\varepsilon/2)kh'(\boldsymbol{V}|\boldsymbol{U})$$

□

We apply the Slack Lemma to $\boldsymbol{h}$ and obtain a number k_1 and an entropic vector $\boldsymbol{h}'$ such that:

$$\begin{aligned} h'(\boldsymbol{X}) \geq &(1-\varepsilon)k_1 h(\boldsymbol{X}) \geq (1-\varepsilon)^3 k_0 k_1 U \\ h'(\sigma) \leq &(1-\varepsilon/2)k_1 h(\sigma) \leq (1-\varepsilon/2)k_0 k_1 b_\sigma, \ \forall \sigma \in \Sigma \end{aligned} \tag{37}$$

Let $g \stackrel{\text{def}}{=} \min_{\boldsymbol{U},\boldsymbol{V}\subseteq \boldsymbol{X}: h'(\boldsymbol{V}|\boldsymbol{U})>0} h'(\boldsymbol{V}|\boldsymbol{U})$ be the smallest non-zero value of $h'(\boldsymbol{V}|\boldsymbol{U})$. By Chan and Yeung's theorem 5.9, there exists a group realizable entropic vector $\boldsymbol{h}^{(r)}$ that satisfies all the FDs satisfied by $\boldsymbol{h}'$, and $||\boldsymbol{h}' - \frac{1}{r}\boldsymbol{h}^{(r)}||_\infty \leq \varepsilon g/4$. Since $U > 0$ we have $h'(\boldsymbol{X}) > 0$ hence $h'(\boldsymbol{X}) \geq g$ and we derive from (37):

$$\begin{aligned} \frac{1}{r}h^{(r)}(\boldsymbol{X}) \geq &h'(\boldsymbol{X}) - \varepsilon g/4 \geq (1-\varepsilon/4)h'(\boldsymbol{X}) \\ \geq &(1-\varepsilon)^4 k_0 k_1 U \end{aligned}$$

On the other hand, $\frac{1}{r}h^{(r)}(\boldsymbol{V}|\boldsymbol{U}) \leq h(\boldsymbol{V}|\boldsymbol{U}) + \varepsilon g/2$, for all sets $\boldsymbol{U},\boldsymbol{V}$. We use $\boldsymbol{h}' \models (1-\varepsilon/2)k_0 k_1 \boldsymbol{b}$ to prove $\boldsymbol{h}^{(r)} \models rk_0 k_1 \boldsymbol{b}$. Consider a statistics $\sigma \in \Sigma$. If $h'(\sigma) = 0$, then $\boldsymbol{h}'$ satisfies the FD σ, and therefore $\boldsymbol{h}^{(r)}$ also satisfies this FD, thus $h^{(r)}(\sigma) = 0 \leq k_0 k_1 b_\sigma$. If $h'(\sigma) > 0$ then $h'(\sigma) \geq g$ and the claim follows from:

$$\begin{aligned} \frac{1}{r}h^{(r)}(\sigma) \leq &h'(\sigma) + \varepsilon g/2 \\ \leq &h'(\sigma) + (\varepsilon/2)h'(\sigma) = (1+\varepsilon/2)h'(\sigma) \\ \leq &(1+\varepsilon/2)(1-\varepsilon/2)k_0 k_1 b_\sigma \leq k_0 k_1 b_\sigma \end{aligned}$$

So far, we have:

$$h^{(r)}(\boldsymbol{X}) \geq (1-\varepsilon)^4 r k_0 k_1 U \qquad\qquad \boldsymbol{h}^{(r)} \models rk_0 k_1 \boldsymbol{b} \tag{38}$$

To complete the proof of Claim 1, we construct the database $\boldsymbol{D}$ as follows. Let the relation R be the group realization of $\boldsymbol{h}^{(r)}$ (Eq. (34)). For each relation $R_j(\boldsymbol{Y}_j)$, define $R_j^{\boldsymbol{D}} \stackrel{\text{def}}{=} \Pi_{\boldsymbol{Y}_j}(R)$. By construction, $Q(\boldsymbol{D}) = R$, and $\log|Q(\boldsymbol{D})| = h^{(r)}(\boldsymbol{X}) \geq (1-\varepsilon)^4 rk_0 k_1 U$ by (38). Furthermore, since $\boldsymbol{h}^{(r)}$ is group-realized, for every statistics $\sigma \in \Sigma$, with guard R_σ, we have $\log \texttt{deg}_{R_\sigma}(\sigma) = h^{(r)}(\sigma) \leq rk_0 k_1 b_\sigma$; thus, $\boldsymbol{D} \models (\Sigma, \boldsymbol{B}^{rk_0 k_1})$. This implies:

$$\sup_{\boldsymbol{D}:\boldsymbol{D}\models(\Sigma,\boldsymbol{B}^{rk_0k_1})} \log|Q(\boldsymbol{D})| \geq (1-\varepsilon)^4 rk_0 k_1 U$$

proving Claim 1 for $k = rk_0 k_1$.

5.3 The Polymatroid Bound Is Not Asymptotically Tight

We prove now Theorem 5.5 item (2).

Proposition 5.11. *The following is a non-Shannon inequality:*

$$\begin{aligned}
11h(ABXYC) \leq & \\
& 3h(XY) + 3h(AX) + 3h(AY) \\
& + h(BX) + h(BY) + 5h(C) \\
& + (h(XYC|AB) + 4h(BC|AXY) + h(AC|BXY)) \\
& + (h(BXY|AC) + 2h(ABY|XC) + 2h(ABX|YC))
\end{aligned} \tag{39}$$

Proof. Consider the following five inequalities:

$$\begin{aligned}
0 \leq & 3h(AX) + 3h(AY) - 4h(AXY) - h(A) \\
& + h(BX) + h(BY) - h(BXY) \\
& - h(AB) + 3h(XY) - 2h(X) - 2h(Y) \\
0 \leq & h(A) + h(C) - h(AC) \\
0 \leq & 2(h(X) + h(C) - h(XC)) \\
0 \leq & 2(h(Y) + h(C) - h(YC)) \\
11h(ABXYC) = & 11h(ABXYC)
\end{aligned}$$

The first inequality holds because it is inequality Eq. (32), expanded and re-arranged. The next three inequalities are basic Shannon inequalities. The last line is an identity. A tedious but straightforward calculation shows that if we add the five (in)equalities above, then we obtain (39), proving the claim. □

Consider the following query, derived from inequality (39):

$$\begin{aligned}
Q(A, B, X, Y, C) = & R_1(X, Y) \wedge R_2(A, X) \wedge R_3(A, Y) \\
& \wedge R_4(B, X) \wedge R_5(B, Y) \wedge R_6(C) \\
& \wedge R_7(A, B, X, Y, C)
\end{aligned}$$

and the following statistics:

$$\begin{aligned}
\Sigma = & \{(XY), (AX), (AY), (BX), (BY), (C), \\
& (XYC|AB), (BC|AXY), h(AC|BXY), \\
& (BXY|AC), (ABY|XC), (ABX|YC)\} \\
\boldsymbol{b} = & \{b_{XY} = b_{AX} = b_{AY} = b_{BX} = b_{BY} = 3, b_C = 2, \\
& b_{XYC|AB} = b_{BC|AXY} = b_{AC|BXY} = 0, \\
& b_{(BXY|AC)} = b_{ABY|XC} = b_{ABX|YC} = 0\}
\end{aligned}$$

In other words, we are given the cardinalities of $R_1, \ldots, R_6$, but are not given the cardinality of R_7, instead we are told that it satisfies the 6 FD's corresponding to the 6 conditional terms in inequality (39). Consider any scale factor $k > 0$, and the scaled log-statistics $k\boldsymbol{b}$. Inequality (39) and the definition (22) imply:

$$\begin{aligned}
& \text{Log-U-Bound}_{\Gamma_n^*}(Q, \Sigma, k\boldsymbol{b}) \leq \\
& k\frac{3b_{XY} + 3b_{AX} + 3b_{AY} + b_{BX} + b_{BY} + 5b_C}{11} = \frac{43k}{11}
\end{aligned}$$

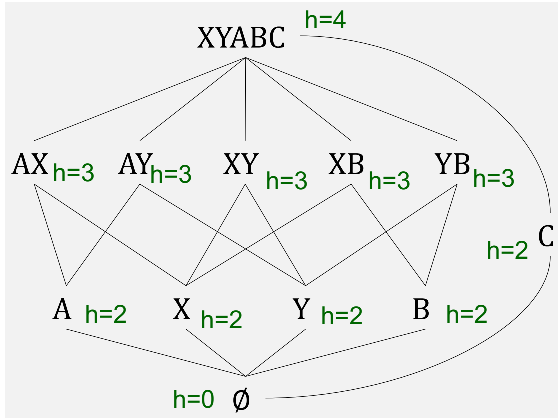


Figure 6: A polymatroid proving that the polymatroid bound is not tight.

By Corollary 4.8, for any database $\boldsymbol{D}$, if $\boldsymbol{D} \models (\Sigma, \boldsymbol{B}^k)$ then:

$$\log |Q(\boldsymbol{D})| \leq \frac{43k}{11}$$

On the other hand, consider the polymatroid $k\boldsymbol{h}$, where $\boldsymbol{h}$ is the polymatroid in Fig. 6. Since $h(ABXYC) = 4$ and $\boldsymbol{h} \models (\Sigma, \boldsymbol{b})$, it follows that $kh(ABXYC) = 4k$, and $k\boldsymbol{h} \models k\boldsymbol{b}$, therefore:

$$\frac{\sup_{\boldsymbol{D}:\boldsymbol{D}\models\boldsymbol{B}^k} \log |Q(\boldsymbol{D})|}{\text{Log-L-Bound}_{\Gamma_n}(Q, \Sigma, k\boldsymbol{b})} \leq \frac{43}{44}$$

This implies Theorem 5.5 item (2).

6 Simple Inequalities

We have a dilemma: the entropic bound is asymptotically tight, but it is open whether it is computable, while the polymatroid bound is computable, but is provably not tight in general. We show in this section that, under a reasonable syntactic restriction on the statistics Σ, these two bounds are equal. We do this by describing a similar syntactic restriction for information inequalities, which we call *simple inequalities*. In that case validity over entropic functions coincides with validity over polymatroids, and we recover the stronger notion of tightness that we had for the AGM bound.

6.1 Background: Subclasses of Polymatroids

A polymatroid $\boldsymbol{h}$ is called *modular* if the submodularity inequality (15) is an equality. Equivalently, $\boldsymbol{h}$ is modular if $\boldsymbol{h} \geq 0$ and for every subset $\alpha \subseteq [n]$, $h(\boldsymbol{X}_\alpha) = \sum_{i\in\alpha} h(X_i)$. We will denote by M_n the set of modular polymatroids, see Fig. 3. For each $i = 1, n$, we call the function $\boldsymbol{h}^{X_i}$ in Fig. 4 a *basic modular function*; recall that $h^{X_i}(\boldsymbol{U}) = 1$ when $X_i \in \boldsymbol{U}$ and $= 0$ otherwise. The following holds:

Proposition 6.1. *(1) A function $\boldsymbol{h}$ is modular iff it is a positive linear combination of basic modular functions, $\boldsymbol{h} = \sum_i a_i \boldsymbol{h}^{X_i}$, where $a_i \geq 0$ for all i. (2) Every modular function is entropic.*

Proof. Item (1) is straightforward, but item (2) requires some thought. It suffices to prove that $a\boldsymbol{h}^{X_i}$ is entropic for all real numbers $a \geq 0$. For that purpose we need to describe one random variable

X_i, whose entropy is $h(X_i) = a$. Let N be a natural number such that $\log N \geq a$, and consider the uniform probability space where X_i has N outcomes with the same probabilities, $p_i = 1/N$, $i = 1, N$. Replace p_1 by $p_1 + \theta$, and replace each p_j with $j > 1$ by $p_j - \theta/(N-1)$, for $\theta \in [0, 1 - \frac{1}{N}]$. When $\theta = 0$ then the distribution is uniform and $h(X_j) = \log N$; when $\theta = 1 - \frac{1}{N}$ then the distribution is deterministic, $p_1 = 1, p_2 = \cdots = p_N = 0$, and $h(X_j) = 0$. By continuity, there exists some θ where $h(X_j) = a$. □

Fix a set of variables $\boldsymbol{W} \subseteq \boldsymbol{X}$. The *step function at* $\boldsymbol{W}$ is:

$$h_{\boldsymbol{W}}(\boldsymbol{U}) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } \boldsymbol{U} \subseteq \boldsymbol{W} \\ 1 & \text{otherwise} \end{cases} \tag{40}$$

There are $2^n - 1$ non-zero step functions (since $\boldsymbol{h}_{\boldsymbol{X}} = 0$). $\boldsymbol{h}_{\boldsymbol{W}}$ is the entropy of the (uniform distribution of the) following relation with 2 tuples:

$$R_{\boldsymbol{W}} \stackrel{\text{def}}{=} R^{\boldsymbol{X}-\boldsymbol{W}} \stackrel{\text{def}}{=} \begin{array}{|cc|c} \boldsymbol{W} & \boldsymbol{X}-\boldsymbol{W} & p \\ \hline 0\cdots0 & 0\cdots0 & 1/2 \\ 0\cdots0 & 1\cdots1 & 1/2 \\ \hline \end{array} \tag{41}$$

Sometimes it is convenient to use an alternative notation. For a set of variables $\boldsymbol{V} \subseteq \boldsymbol{X}$, define:

$$h^{\boldsymbol{V}}(\boldsymbol{U}) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } \boldsymbol{U} \cap \boldsymbol{V} = \emptyset \\ 1 & \text{otherwise} \end{cases} \tag{42}$$

Then $\boldsymbol{h}^{\boldsymbol{V}} = \boldsymbol{h}_{\boldsymbol{X}-\boldsymbol{V}}$. A basic modular function $\boldsymbol{h}^{X_i}$ is the same as the step function $\boldsymbol{h}^{\{X_i\}}$; if $|\boldsymbol{V}| \geq 2$ then $\boldsymbol{h}^{\boldsymbol{V}}$ is not modular.

Definition 6.2. *A* normal polymatroid *is a positive linear combination of step functions,*

$$\boldsymbol{h} = \sum_{\boldsymbol{V} \subseteq \boldsymbol{X}, \boldsymbol{V} \neq \emptyset} a_{\boldsymbol{V}} \boldsymbol{h}^{\boldsymbol{V}} \tag{43}$$

where $a_{\boldsymbol{V}} \geq 0$ *for all* $\boldsymbol{V}$*.*

We denote by N_n the set of normal polymatroids, see Fig. 3. Normal polymatroids are the same as *polymatroids with a non-negative I-measure* described in [Yeu08, KKNS21].

Proposition 6.3. *The* $2^n - 1$ *non-zero step functions* $\boldsymbol{h}^{\boldsymbol{V}}$*,* $\boldsymbol{V} \neq \emptyset$ *form a basis of the vector space* $\{\boldsymbol{h} \in \mathbb{R}^{2^{[n]}} \mid h(\emptyset) = 0\}$*. More precisely, every such vector* $\boldsymbol{h}$ *satisfies* $\boldsymbol{h} = \sum_{\boldsymbol{V}} a_{\boldsymbol{V}} \boldsymbol{h}^{\boldsymbol{V}}$*, where:*

$$a_{\boldsymbol{U}} \stackrel{def}{=} - \sum_{\boldsymbol{V} \subseteq \boldsymbol{U}} (-1)^{|\boldsymbol{V}|} h(\boldsymbol{V} | \boldsymbol{X} - \boldsymbol{U}) \tag{44}$$

The proof follows by solving the following system of linear equations with unknowns $a_{\boldsymbol{V}}$:

$$\forall \boldsymbol{U} \neq \emptyset: \quad h(\boldsymbol{U}) = \sum_{\boldsymbol{V}: \boldsymbol{V} \cap \boldsymbol{U} \neq \emptyset} a_{\boldsymbol{V}} \tag{45}$$

The solution is obtained by using Möbius' inversion formula (we prove this in Appendix .4) and consists of the expression (44). Expression (44) is called *conditional interaction information*, and denoted by $I(X_{i_1}; X_{i_2}; \cdots | \boldsymbol{X} - \boldsymbol{U})$, where $\boldsymbol{U} = \{X_{i_1}, X_{i_2}, \ldots\}$. The following holds (the proof is immediate and omitted):

Proposition 6.4. *(1) A function $\boldsymbol{h}$ is a normal polymatroid iff, for every set $\boldsymbol{U} \subseteq \boldsymbol{X}$, $\boldsymbol{U} \neq 0$, the conditional interaction information* (44) *is* ≥ 0. *(2) Every normal polymatroid is entropic.*

Example 6.5. *The parity function $\boldsymbol{h}$ Fig. 2 is the simplest example of a polymatroid that is not normal. The coefficients can be derived using* (44)*, or, we can check directly that:*

$$\boldsymbol{h} = \boldsymbol{h}^{X,Y} + \boldsymbol{h}^{X,Z} + \boldsymbol{h}^{Y,Z} - \boldsymbol{h}^{X,Y,Z}$$

The coefficient of $\boldsymbol{h}^{X,Y,Z}$ is negative, hence $\boldsymbol{h}$ is not normal.

6.2 Special Inequalities

We describe here a class of information inequalities, called *simple inequalities*, were Γ_n^*-validity coincides with Γ_n-validity. The modular and normal polymatroids turn out to be the key tools to study these inequalities.

The following is sometimes referred in the literature as the *modularization lemma*:

Lemma 6.6. *For any polymatroid $\boldsymbol{h}$ there exists a modular polymatroid $\boldsymbol{h}'$ such that (a) $\boldsymbol{h}' \leq \boldsymbol{h}$ and (b) $\boldsymbol{h}'(\boldsymbol{X}) = \boldsymbol{h}(\boldsymbol{X})$.*

Proof. Order the variables arbitrarily $X_1, \ldots, X_n$ and define $h'(X_i) \stackrel{\text{def}}{=} h(X_i|\boldsymbol{X}_{[1:i-1]})$, where $\boldsymbol{X}_{[1:i-1]} \stackrel{\text{def}}{=} \{X_1, \ldots, X_{i-1}\}$. We check condition (a): for $\alpha \subseteq [n]$,

$$\begin{aligned} h'(\boldsymbol{X}_\alpha) =& \sum_{i\in\alpha} h(X_i|\boldsymbol{X}_{[1:i-1]}) \\ \leq& \sum_{i\in\alpha} h(X_i|\boldsymbol{X}_{[1:i-1]\cap\alpha}) = h(\boldsymbol{X}_\alpha) \end{aligned}$$

We check (b): $h'(\boldsymbol{X}) = \sum_{i=1,n} h(X_i|\boldsymbol{X}_{[1:i-1]}) = h(\boldsymbol{X})$. □

The modularization lemma gives us an alternative, and more general proof of Theorem 4.10:

Corollary 6.7. *Consider an inequality of the form $\sum_i w_i h(\boldsymbol{V}_i) \geq h(\boldsymbol{X})$, where $w_i \geq 0$ and $\boldsymbol{V}_i$ are subsets of $\boldsymbol{X}$. The following conditions are equivalent:*

(1) The inequality is valid for polymatroids.

(2) The inequality is valid for entropic functions.

(3) The inequality is valid for modular functions.

Proof. The implications (1) ⇒ (2) ⇒ (3) are immediate. We prove (3) ⇒ (1), by contradiction: if the inequality fails on some polymatroid $\boldsymbol{h}$, $\sum_i w_i h(\boldsymbol{V}_i) < h(\boldsymbol{X})$, and $\boldsymbol{h}'$ is the modular function in Lemma 6.6, then, $\sum_i w_i h'(\boldsymbol{V}_i) \leq \sum_i w_i h(\boldsymbol{V}_i) < h(\boldsymbol{X}) = h'(\boldsymbol{X})$ contradicting (3). □

We prove in Appendix .5 the following extension of the Modularization Lemma:

Lemma 6.8. *For any polymatroid $\boldsymbol{h}$ there exists a normal polymatroid $\boldsymbol{h}'$ such that (a) $\boldsymbol{h}' \leq \boldsymbol{h}$, (b) $h'(\boldsymbol{X}) = h(\boldsymbol{X})$, and (c) $h'(X_i) = h(X_i)$ for every variable $X_i \in \boldsymbol{X}$.*

Definition 6.9. *We call a set of statistics Σ* simple *if, for all $(\boldsymbol{V}|\boldsymbol{U}) \in \Sigma$, $|\boldsymbol{U}| \leq 1$. A* simple information inequality *is a Σ-inequality where Σ is simple:*

$$\sum_{\sigma\in\Sigma} w_\sigma h(\Sigma) \geq h(\boldsymbol{X}) \tag{46}$$

We immediately derive:

Corollary 6.10. *Given a simple inequality* (46)*, the following are equivalent:*

(1) The inequality is valid for polymatroids.

(2) The inequality is valid for entropic functions.

(3) The inequality is valid for normal polymatroids

The proof is identical to that of Corollary 6.7 and omitted.

6.3 Special Databases

When the statistics Σ are simple, then we show here that the polymatroid and the entropic bound coincide. We also show that the bound is tight, using a similar notion of tightness as in the AGM bound, where the ratio between the lower and upper bound depends only on the query; also, there is no need to amplify the statistics values. Moreover, like in the AGM bound, the worst-case database instance has a special structure, which we call a *normal database*. We start by showing:

Theorem 6.11. *If Σ is simple, then:*

$$\begin{aligned} &\textit{Log-U-Bound}_{N_n}(Q, \Sigma, \boldsymbol{b}) = \\ &\quad = \textit{Log-U-Bound}_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{b}) = \textit{Log-U-Bound}_{\Gamma_n}(Q, \Sigma, \boldsymbol{b}) \end{aligned}$$

Proof. Since $N_n \subseteq \Gamma_n^* \subseteq \Gamma_n$ we have inequalities above: $\cdots \leq \cdots \leq \cdots$ Corollary 6.10 implies $\text{Log-U-Bound}_{N_n}(Q, \Sigma, \boldsymbol{b}) = \text{Log-U-Bound}_{\Gamma_n}(Q, \Sigma, \boldsymbol{b})$, hence all three quantities are equal. □

We describe now normal relational instances, and normal databases. Start with a single relation $R(\boldsymbol{X})$ with n attributes $\boldsymbol{X}$. Recall that an instance R is a *product relation* if $R = S_1 \times \cdots \times S_n$, for n sets $S_1, \ldots, S_n$: the worst-case instance of the AGM bound consisted of product relations. We generalize this concept:

Definition 6.12. *A relation instance T with n attributes is a* normal relation *if there exists m finite sets $S_1, \ldots, S_m$ and a function $\psi : [n] \to 2^{[m]}$ such that*

$$T = \{(\boldsymbol{s}_{\psi(1)}, \boldsymbol{s}_{\psi(2)}, \ldots, \boldsymbol{s}_{\psi(n)}) \mid \boldsymbol{s} \in S_1 \times \cdots \times S_m\}$$

In a normal relation the values of an attribute can be tuples themselves. Every product relation is a normal relation, but not vice versa. A database instance is *normal* if each of its relations is normal. A *basic normal relation of size N* is the following:

$$T_N^{\boldsymbol{V}} \stackrel{\text{def}}{=} \{(k \cdot \mathbf{1}_{X_1 \in \boldsymbol{V}}, \cdots, k \cdot \mathbf{1}_{X_n \in \boldsymbol{V}}) \mid k = 0, N-1\} \tag{47}$$

Here $\mathbf{1}_{X_i \in \boldsymbol{V}}$ is an indicator variable that is 1 when $X_i \in \boldsymbol{V}$ and 0 otherwise; thus, if an attribute X_i is in $\boldsymbol{V}$ then it takes the values $0, 1, \ldots, N-1$ in the relation $T_N^{\boldsymbol{V}}$, otherwise it has constant values 0. The entropic vector of $T_N^{\boldsymbol{V}}$ is $(\log N)\boldsymbol{h}^{\boldsymbol{V}}$. In particular, the relation $R^{\boldsymbol{V}}$ in (41) is $T_2^{\boldsymbol{V}}$.

Example 6.13. *We give three examples of normal relations with $n = 3$ attributes:*

$$\begin{aligned} A =& \{(i, j, k) \mid i, j, k \in [0 : N-1]\} && \textit{product relation} \\ B =& \{(i, i, i) \mid i \in [0 : N-1]\}, && \textit{normal relation} \\ C =& \{(i, (i, j), j) \mid i, j \in [0 : N-1]\} && \textit{normal relation} \end{aligned}$$

Their cardinalities are $|A| = N^3$, $|B| = N$, $|C| = N^2$. We also notice:

$$A = T_N^X \otimes T_N^Y \otimes T_N^Z \qquad B = T_N^{XYZ} \qquad C = T_N^{X,Y} \otimes T_N^{Y,Z}$$

We prove that the lower bound for simple statistics is tight.

Theorem 6.14. *Let Σ be a set of simple statistics for a query Q and let $\boldsymbol{B}$ be statistics values. Then there exists a worst-case instance $\boldsymbol{D}$ such that $|Q(\boldsymbol{D})| \geq \frac{1}{2^{2^n-1}}$ U-Bound$_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{B})$.*

Proof. We use the following, whose proof is immediate:

Proposition 6.15. *Let $R(\boldsymbol{X})$, $R'(\boldsymbol{X})$ be relations over the same attributes $\boldsymbol{X}$.*

- *If R, R' are normal relations, then $R \otimes R'$ is normal.*
- $\mathtt{deg}_{R\otimes R'}(\sigma) = \mathtt{deg}_R(\sigma) \cdot \mathtt{deg}_{R'}(\sigma)$, *for all $\sigma = (\boldsymbol{V}|\boldsymbol{U})$.*

Denote by $U \stackrel{\text{def}}{=} \text{U-Bound}_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{B})$, $\boldsymbol{b} \stackrel{\text{def}}{=} \log \boldsymbol{B}$, then $\log U = \text{Log-U-Bound}_{N_n}(Q, \Sigma, \boldsymbol{b}) = \text{Log-L-Bound}_{N_n}(Q, \Sigma, \boldsymbol{b})$, by Theorem 6.11 and Theorem 5.2 respectively. Let $\boldsymbol{h}^* \in N_n$ be the optimal solution to the linear program defining $\text{Log-L-Bound}_{N_n}(Q, \Sigma, \boldsymbol{b})$ (see Theorem 5.2), then $\boldsymbol{h}^* \models (\Sigma, \boldsymbol{b})$ and $h^*(\boldsymbol{X}) = \log U$. Since $\boldsymbol{h}^*$ is normal, it can be written as:

$$\boldsymbol{h}^* = \sum_{\boldsymbol{V}\neq\emptyset} a_{\boldsymbol{V}} \boldsymbol{h}^{\boldsymbol{V}}, \qquad a_{\boldsymbol{V}} \geq 0$$

Then, $\log U = h^*(\boldsymbol{X}) = \sum_{\boldsymbol{V}\neq\emptyset} a_{\boldsymbol{V}}$, and $U = \prod_{\boldsymbol{V}} 2^{a_{\boldsymbol{V}}}$.

For each set $\boldsymbol{V}$, $\emptyset \neq \boldsymbol{V} \subseteq \boldsymbol{X}$ we define:

$$b_{\boldsymbol{V}} \stackrel{\text{def}}{=} \lfloor 2^{a_{\boldsymbol{V}}} \rfloor, \quad P^{\boldsymbol{V}} \stackrel{\text{def}}{=} T^{\boldsymbol{V}}_{b_{\boldsymbol{V}}} \qquad \text{basic normal relation (47)}$$

$$R \stackrel{\text{def}}{=} \bigotimes_{\boldsymbol{V}} P^{\boldsymbol{V}} \qquad \text{normal relation}$$

Define the worst-case instance as $\boldsymbol{D} = (R_1^D, \ldots, R_m^D)$, where $R_j^D = \Pi_{\boldsymbol{Y}_j}(R)$. We first check that $\boldsymbol{D}$ satisfies the constraints, and for that let $\sigma \in \Sigma$ have witness R_σ, then:

$$\log \mathtt{deg}_{R_\sigma}(\sigma) = \log \mathtt{deg}_R(\sigma) = \sum_{\boldsymbol{V}} \log \mathtt{deg}_{P^{\boldsymbol{V}}}(\sigma)$$
$$= \sum_{\boldsymbol{V}} (\log b_{\boldsymbol{V}}) h^{\boldsymbol{V}}(\sigma) \leq \sum_{\boldsymbol{V}} a_{\boldsymbol{V}} h^{\boldsymbol{V}}(\sigma) = h^*(\sigma) \leq b_\sigma$$

Finally, we check the query's output size:

$$\log |Q(\boldsymbol{D})| = \log |R| = \log \prod_{\boldsymbol{V}} |P^{\boldsymbol{V}}| = \sum_{\boldsymbol{V}} (\log b_{\boldsymbol{V}}) h^{\boldsymbol{V}}(\boldsymbol{X})$$

Since $h^{\boldsymbol{V}}(\boldsymbol{X}) = 1$, this implies $|Q(\boldsymbol{D})| = \prod_{\boldsymbol{V}} b_{\boldsymbol{V}} = \prod_{\boldsymbol{V}} \lfloor 2^{a_{\boldsymbol{V}}} \rfloor \geq \frac{1}{2^{2^n-1}} U$, because $\lfloor 2^{a_{\boldsymbol{V}}} \rfloor \geq \frac{1}{2} 2^{a_{\boldsymbol{V}}}$. □

The reader may want to check the analogy with the worst-case instance of the AGM bound: the optimal solution $\boldsymbol{v}^*$ there became here $\boldsymbol{h}^*$, and the domain $V_i = [\lfloor 2^{v_i^*} \rfloor]$ defined for the variable X_i became here the normal relation $P^{\boldsymbol{V}}$. As before, we constructed the worst-case instance $\boldsymbol{D}$ without amplifying the statistics, and $|Q(\boldsymbol{D})|$ is within a constant, which depends only on the query, of $\text{U-Bound}_{\Gamma_n^*}(Q, \Sigma, \boldsymbol{B})$.

Discussion The restriction to simple statistics occurs naturally in many applications. Databases are often designed with simple keys (consisting of a single attribute), and applications that use degrees often consider only simple degrees. The restriction to simple statistics is often acceptable.

It remains open where one can extend this definition to richer classes of statistics, or inequalities, while still preserving the property that validity for entropic vectors is the same as validity for polymatroids. The set of statistics in Example 4.9 is not "simple", yet the entropic bound coincides with the polymatroid bound. This (and other examples) suggests that other non-trivial syntactic classes may exist where these two bounds agree.

7 Query Evaluation

The query evaluation problem is: given a conjunctive query Q, evaluate it on a (usually large) database $\boldsymbol{D}$. In this paper we consider only the *data complexity*, where the query is fixed, and the runtime is given as a function of the statistics of $\boldsymbol{D}$. Database systems compute queries using a sequence of *binary joins*, of the form $C(X,Y,Z) = A(X,Y) \wedge B(Y,Z)$, which are written as $C = A \bowtie B$. Assuming all relations are pre-sorted, the time complexity of the join is $\tilde{O}(|A| + |B| + |A \bowtie B|)$. A *semi-join*, denoted $C = A \ltimes B$, is a join followed by the projection on the attributes of the first relation, meaning $C(X,Y) = \exists Z(A(X,Y) \wedge B(Y,Z))$. A semijoin can be computed in time $\tilde{O}(|A|)$.

A *Worst Case Optimal Join* (WCOJ) is an algorithm that evaluates Q in time no larger than its theoretical upper bound. A sequence of binary joins is usually not a WCOJ, because intermediate results may be larger than the theoretical upper bound of the query. For example the upper bound for the triangle query in Example 3.2 is $N^{3/2}$, but if we evaluate it as $(R \bowtie S) \bowtie T$, the join $R \bowtie S$ can have size N^2.

Any WCOJ algorithm represents an indirect proof of the query's upper bound, since the size of the output cannot exceed the time complexity of the algorithm. For example, if we are given an algorithm for the triangle query, together with a proof that its runtime is $\tilde{O}(N^{3/2})$, then we have a proof that the size of the output is also $\tilde{O}(N^{3/2})$. This means that proving an upper bound on the query's output is *inevitable* for designing a WCOJ. We show in this section that one can proceed in reverse: given a proof of the upper bound, convert it into a WCOJ. We call this paradigm *From Proofs to Algorithms*. Thus, the question to ask in designing a WCOJ algorithm is: how do we *prove* an upper bound on the query's output? And how do we convert it into an algorithm?

7.1 Generic Join

Consider the setting of the AGM bound: we are given only cardinality statistics on the base relations. In that case, a proof of the upper bound is a proof of $\sum_j w_j h(\boldsymbol{Y}_j) \geq h(\boldsymbol{X})$, since it implies $|Q| \leq \prod_j |R_j|^{w_j}$. We gave a proof of this inequality in Theorem 4.10; the proof consists of conditioning on the last variable X_n, then applying induction on the remaining variables. We convert that proof into an algorithm: iterate X_n over its domain, and compute recursively the residual query. This algorithm is called *Generic Join*, or GJ, and was introduced by Ngo, Ré, and Rudra [NRR13]. We describe it in detail next.

Fix a full conjunctive query with variables $\boldsymbol{X}$, which we write as $Q = \Join_{j=1,m} R_j$. As usual, $\boldsymbol{Y}_j$ are the variables of R_j. Generic Join computes Q as follows:

- Let X_n be an arbitrary variable.
- Partition the set of indices j into J_0 and J_1:
$$J_0 \stackrel{\text{def}}{=} \{j \mid X_n \notin \boldsymbol{Y}_j\}, \qquad J_1 \stackrel{\text{def}}{=} \{j \mid X_n \in \boldsymbol{Y}_j\}.$$
- Compute the set $D = \bigcap_{j \in J_1} \Pi_{X_n}(R_j)$.
- For each value $x \in D$, do:
  - Compute $R_j[x] := \Pi_{\boldsymbol{Y}_j - X_n}(\sigma_{X_n = x}(R_j))$, for $j \in J_1$.
  - Denote $R_j[x] := R_j$ for $j \in J_0$.
  - Compute the residual query $\Join_{j=1,m} R_j[x]$.

We invite the reader to check how the algorithm can be "read off" the proof of Theorem 4.10. To compute the runtime of the algorithm, assume that the relations are given in listing representation, sorted lexicographically using the attribute order $X_n, X_{n-1}, \ldots, X_1$. Then, the runtime, T_n, is:

$$T_n(R_1, \ldots, R_m) = T_{\text{intersection}} + \sum_x T_{n-1}(R_1[x], \ldots, R_m[x])$$

By induction hypothesis:

$$T_{n-1}(R_1[x], \ldots, R_m[x]) = \tilde{O}\left(\prod_j |R_j[x]|^{w_j}\right)$$

which leads to:

$$\begin{aligned}
T_n = T_{\text{intersection}} + \tilde{O}\left(\prod_{j \in J_0} |R_j|^{w_j} \sum_x \prod_{j \in J_1} |R_j[x]|^{w_j}\right) \\
\le T_{\text{intersection}} + \tilde{O}\left(\prod_{j \in J_0} |R_j|^{w_j} \prod_{j \in J_1} \left(\sum_x |R_j[x]|\right)^{w_j}\right) \\
= T_{\text{intersection}} + \tilde{O}\left(\prod_j |R_j|^{w_j}\right)
\end{aligned}$$

We used Hölder's inequality in Fig. 1 (since $\sum_{j \in J_1} w_j \ge 1$, because X_n is covered), and the fact that $\sum_x |R_j[x]| = |R_j|$ for $j \in J_1$. The crux of the algorithm is the intersection: its runtime should not exceed $\prod_j |R_j|^{w_j}$, and for that it suffices to iterate over the smallest set $\Pi_{X_n}(R_j)$, and probe in the others: the runtime is $\tilde{O}(\min_{j \in J_1} |R_j|) \le \tilde{O}(\prod_{j \in J_1} |R_j|^{w_j})$, since $\sum_{j \in J_1} w_j \ge 1$.

Example 7.1. *Using the variable order X, Y, Z, GJ computes the triangle query $R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ as follows:*

For $x \in \Pi_X(R) \cap \Pi_X(T)$ ***do:***
  For $y \in \Pi_Y(R[X=x]) \cap \Pi_Y(S)$ ***do:***
    For $z \in \Pi_Z(S[Y=y]) \cap \Pi_Z(T[X=x])$ ***do:***
      output(x, y, z)

The choice of algorithm for computing the intersection is critical for GJ. To see this, consider the simplest query, $Q(X) = R(X) \wedge S(X)$, that *is* an intersection. The AGM bound is $\min(|R|, |S|)$, corresponding to the edge covers $(1,0)$ and $(0,1)$, and GJ must compute the query in time $\tilde{O}(\min(|R|, |S|))$. By assumption, R, S are already sorted, but we cannot run a standard merge algorithm, since its runtime is $O(|R| + |S|)$; instead, we iterate over the smaller relation and do a binary search in the larger.

Because of its simplicity and ease of implementation, GJ is the poster child of WCOJ algorithms. One remarkable property of GJ is that its runtime is always bounded by the AGM bound, no matter what variable order we choose. Before GJ, Veldhuizen [Vel14] described an algorithm called Leapfrog Triejoin (LFTJ), which uses a similar logic as GJ, but also specifies in the details of the required trie data structure. Several implementations of GJ/LFTJ exists today [SOC16, AtCG+15, FBS+20, MKS21, WWS23].

7.2 The Heavy/Light Algorithm

Balister and Bollobás [BB12] provided the following alternative proof of an inequality of the form (23), which we write in an equivalent form using integer coefficients:

$$E \stackrel{\text{def}}{=} \sum_{j=1,m} k_j h(\boldsymbol{Y}_j) \geq k_0 h(\boldsymbol{X}) \tag{48}$$

where $k_i \in \mathbb{N}$, for $i = 0, m$. View the expression E as a bag of terms $h(\boldsymbol{Y}_j)$ where each term $h(\boldsymbol{Y}_j)$ occurs k_j times. A *compression step* consists of the following:

- Choose two terms $h(\boldsymbol{U}), h(\boldsymbol{V}) \in E$ such that $\boldsymbol{U} \not\subseteq \boldsymbol{V}$ and $\boldsymbol{V} \not\subseteq \boldsymbol{U}$.
- Replace $h(\boldsymbol{U}) + h(\boldsymbol{V})$ with $h(\boldsymbol{U} \cup \boldsymbol{V}) + h(\boldsymbol{U} \cap \boldsymbol{V})$.

Theorem 7.2. *[BB12] Any sequence of compression steps eventually leads to:*

$$E = \ell_0 h(\boldsymbol{Z}_0) + \ell_1 h(\boldsymbol{Z}_1) + \cdots, \qquad \text{where } \boldsymbol{Z}_0 \supset \boldsymbol{Z}_1 \supset \cdots \tag{49}$$

Furthermore, if each variable X_i is covered at least $k_0 \geq 1$ times[3] *by the original expression E in* (48)*, then $\boldsymbol{Z}_0 = \boldsymbol{X}$ and $\ell_0 \geq k_0$; in particular, the inequality* (48) *is valid.*

Proof. Each compression step strictly increases the quantity $\sum_{h(\boldsymbol{Z})\in E} |\boldsymbol{Z}|^2$. To see this, write $\boldsymbol{U} = \boldsymbol{A} \cup \boldsymbol{C}$, $\boldsymbol{V} = \boldsymbol{B} \cup \boldsymbol{C}$ where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ are disjoint sets, then $|\boldsymbol{U}|^2 + |\boldsymbol{V}|^2 = (|\boldsymbol{A}| + |\boldsymbol{C}|)^2 + (|\boldsymbol{B}| + |\boldsymbol{C}|)^2$, while $|\boldsymbol{U} \cup \boldsymbol{V}|^2 + |\boldsymbol{U} \cap \boldsymbol{V}|^2 = (|\boldsymbol{A}| + |\boldsymbol{B}| + |\boldsymbol{C}|)^2 + |\boldsymbol{C}|^2$, and the latter is strictly larger when $|\boldsymbol{A}| \cdot |\boldsymbol{B}| > 0$. This quantity cannot exceed $(\sum_j k_j) n^2$, therefore compression needs to terminate, and this happens when for any two sets in E one contains the other. Then, E must have the form (49). Finally, we observe that compression preserves the number of times each variable X_i is covered by E, because the number of sets in $\{\boldsymbol{U}, \boldsymbol{V}\}$ containing X_i is the same as the number of sets in $\{\boldsymbol{U} \cup \boldsymbol{V}, \boldsymbol{U} \cap \boldsymbol{V}\}$ containing X_i. Therefore, if each variable is covered by E at least k_0 times, then $\boldsymbol{Z}_0 = \boldsymbol{X}$ and $\ell_0 \geq k_0$. □

Call a sequence of compression steps that converts an expression E in (48) to (49) a *BB-proof sequence*. To derive an algorithm, we need to impose an additional restriction. Call a BB-proof sequence *divergent* if, after each compression step $h(\boldsymbol{U}) + h(\boldsymbol{V}) \to h(\boldsymbol{U} \cup \boldsymbol{V}) + h(\boldsymbol{U} \cap \boldsymbol{V})$, we can split E into $E' + E''$, such that E' contains $h(\boldsymbol{U} \cup \boldsymbol{V})$ and covers every variable at least k_0' times, E'' contains $h(\boldsymbol{U} \cap \boldsymbol{V})$ and covers every variable at least k_0'' times, and $k_0 = k_0' + k_0''$.

We convert a divergent BB-proof sequence into an algorithm called the *Heavy/Light Algorithm*. Let $\boldsymbol{B}$ be the statistics values, $\boldsymbol{b} \stackrel{\text{def}}{=} \log \boldsymbol{B}$, and set $B \stackrel{\text{def}}{=} \max_j B_j$. Assume w.l.o.g. that the inequality $E \stackrel{\text{def}}{=} \sum_j k_j h(\boldsymbol{Y}_j) \geq k_0 h(\boldsymbol{X})$ is optimal, meaning that $\sum_j (k_j/k_0) b_j = \text{Log-U-Bound}_{\Gamma_n}(Q, \Sigma, \log(\boldsymbol{B}))$ (see Eq. (22)). Denote by $\boldsymbol{h}^*$ an optimal solution to the dual, meaning $\boldsymbol{h}^* \models (\Sigma, \boldsymbol{b})$ and $h^*(\boldsymbol{X}) = \text{Log-L-Bound}_{\Gamma_n}(Q, \Sigma, \boldsymbol{b})$ (see Eq. (24)). These two quantities are the same by Thm. 5.2: $\sum_j (k_j/k_0) b_j = h^*(\boldsymbol{X})$. The algorithm uses a working memory which stores, for each term $h(\boldsymbol{Z})$ in E, a temporary relation $S(\boldsymbol{Z})$, called the *guard* of $h(\boldsymbol{Z})$, and maintains the invariant: $\log |S(\boldsymbol{Z})| \leq h^*(\boldsymbol{Z})$. Initially, the working memory is $\{R_j \mid k_j > 0\}$: by complementary slackness, if $k_j > 0$ then the dual constraint constraint is tight, $h^*(\boldsymbol{Y}_j) = b_j$, and the invariant holds because $\log |R_j| \leq \log B_j = b_j = h^*(\boldsymbol{Y}_j)$.

The algorithm repeatedly processes a compression step $h(\boldsymbol{U}) + h(\boldsymbol{V}) \to h(\boldsymbol{U} \cup \boldsymbol{V}) + h(\boldsymbol{U} \cap \boldsymbol{V})$ of the BB-sequence, as follows. If the two guards are $S(\boldsymbol{U})$ and $S'(\boldsymbol{V})$, let $\boldsymbol{C} \stackrel{\text{def}}{=} \boldsymbol{U} \cap \boldsymbol{V}$, write

[3] Meaning: $\sum_{j: X_i \in \boldsymbol{Y}_j} k_j \geq k_0$.

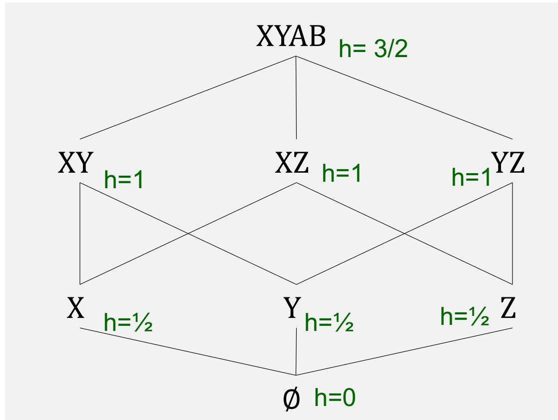


Figure 7: A simple polymatroid used in Example 7.3.

$\boldsymbol{U} = \boldsymbol{AC}$, $\boldsymbol{V} = \boldsymbol{BC}$, and define $M \stackrel{\text{def}}{=} 2^{h^*(\boldsymbol{A}|\boldsymbol{C})}$. Partition the guard $S(\boldsymbol{AC})$ into two subsets:

$$\begin{aligned} S_{\text{light}}(\boldsymbol{A}, \boldsymbol{C}) =& \{(\boldsymbol{a}, \boldsymbol{c}) \in S \mid \texttt{deg}_S(\boldsymbol{A}|\boldsymbol{C} = \boldsymbol{c}) \leq M\} \\ S_{\text{heavy}}(\boldsymbol{C}) =& \{\boldsymbol{c} \in \Pi_{\boldsymbol{C}}(S) \mid \texttt{deg}_S(\boldsymbol{A}|\boldsymbol{C} = \boldsymbol{c}) > M\} \end{aligned}$$

Compute new guards using a join and a semijoin:

$$\begin{aligned} S''(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) :=& S_{\text{light}}(\boldsymbol{A}, \boldsymbol{C}) \bowtie S'(\boldsymbol{B}, \boldsymbol{C}) \\ S'''(\boldsymbol{C}) :=& S_{\text{heavy}}(\boldsymbol{C}) \ltimes S'(\boldsymbol{B}, \boldsymbol{C}) \end{aligned}$$

The invariant holds because $|S''| \leq M \cdot |S|$ implies $\log |S''| \leq h^*(\boldsymbol{A}|\boldsymbol{C}) + h^*(\boldsymbol{BC}) \leq h^*(\boldsymbol{ABC})$, and because $|S_{\text{heavy}}| \leq |S|/M$ (since every $\boldsymbol{c} \in S_{\text{heavy}}$ occurs $\geq M$ times in S) implies $\log |S'''| \leq \log |S| - \log M = h^*(\boldsymbol{AC}) - h^*(\boldsymbol{A}|\boldsymbol{C}) = h^*(\boldsymbol{C})$. The runtime of the join and semijoin is $\leq \tilde{O}(2^{h^*(\boldsymbol{X})}) = \tilde{O}(\text{U-Bound}_{\Gamma_n}(Q, \Sigma, \boldsymbol{B}))$. Next, the algorithm proceeds recursively, by processing independently E' and E'', semi-joins the result of E' with the relations missing from E', similarly semi-joins the result of E'' with the relations missing from E'', then returns the union of these two results. Correctness is easily checked.

Example 7.3. *Consider the triangle query $R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$, and the following divergent proof: $\underline{h(XY) + h(YZ)} + h(ZX) \to h(XYZ) + \underline{h(Y) + h(ZX)} \to h(XYZ) + h(XYZ)$. Assume for simplicity that the three relations have the same cardinalities $|R| = |S| = |T| = B$. The optimal polymatroid is $\boldsymbol{h}^* \stackrel{def}{=} \log B \cdot \boldsymbol{h}$, for $\boldsymbol{h}$ in Fig. 7.*

The Heavy/Light Algorithm proceeds as follows. For the first compression step it partitioning R into:

$$\begin{aligned} R_{light} :=& \{(x, y) \mid \texttt{deg}_R(X|Y = y) \leq 2^{h^*(Y|X)} = B^{1/2}\} \\ R_{heavy} :=& \{y \mid \texttt{deg}_R(X|Y = y) > 2^{h^*(Y|X)} = B^{1/2}\} \end{aligned}$$

then it computes:

$$\begin{aligned} \texttt{Temp}_1(X, Y, Z) :=& R_{light}(X, Y) \bowtie S(Y, Z) \\ \texttt{Temp}_2(Y) :=& R_{heavy}(Y) \ltimes S(Y, Z) \end{aligned}$$

At this point the BB-proof diverged into two branches, $h(XYZ) \geq h(XYZ)$, and $h(Y) + h(ZX) \geq h(XYZ)$, and we perform a recursive call for each branch. The first branch immediately returns

$\texttt{Temp}_1$, which we semi-join with T: $Q'(X, Y, Z) = \texttt{Temp}_1(X, Y, Z) \ltimes T(Z, X)$. The other branch applies the second compression step which corresponds to the following join operation:

$$\texttt{Temp}_3(X, Y, Z) := \texttt{Temp}_2(Y) \times T(Z, X)$$

which we semi-join with R and S: $Q''(X, Y, Z) := \texttt{Temp}_3(X, Y, Z) \ltimes R(X, Y) \ltimes S(Y, Z)$. Finally, we return the union $Q' \cup Q''$. The reader may verify that the runtime is $\tilde{O}(B^{3/2})$.

An advantage of the Heavy/Light Algorithm over GJ is that it reuses existing join operators, which already have very efficient implementations in database systems. However, the algorithm only works for divergent BB-proofs. This raises the question: does every inequality (48) have a divergent proof? The answer is negative, as provided by the following example due to Yilei Wang [Wan22].

Example 7.4. *The following has no divergent BB-proof:*

$$E = h(XYZ) + h(ZUV) + h(VWX) + h(YUW) \geq 2h(XYZUVW)$$

Assume w.l.o.g. that we start by compressing $h(XYZ) + h(ZUV) \to h(XYZUV) + h(Z)$ (by symmetry, all other choices are equivalent). Then we need to partition E into $E' + E''$. Suppose E' contains $h(Z)$; since E' covers every variable, it must contain both remaining terms $h(VWX) + h(YUW)$, which means that E'' can only contain $h(XYZUV)$ alone, and it does not cover W.

7.3 PANDA

Both Generic Join and the Heavy/Light Algorithm are restricted to cardinality statistics, in other words they only work in the framework of the AGM bound. PANDA, introduced in [KNS17], is a WCOJ algorithm that works for general statistics. While it runs in time given by the theoretical query upper bound, it also includes a polylogarithm factor in the size of the database, with a rather large exponent. We describe PANDA here at a high level, and refer the reader to [KNS17] for details.

Let Σ be a set of statistics, and consider a Σ-inequality with integer coefficients:

$$E \stackrel{\text{def}}{=} \sum_{\sigma \in \Sigma} k_\sigma h(\sigma) \geq k_0 h(\boldsymbol{X}) \tag{50}$$

A *CD-proof sequence* for the inequality (50) is a sequence of steps that convert the LHS to the RHS, where each step is one of the following:

- Composition: $h(\boldsymbol{U}) + h(\boldsymbol{V}|\boldsymbol{U}) \to h(\boldsymbol{UV})$.
- Decomposition: $h(\boldsymbol{UV}) \to h(\boldsymbol{U}) + h(\boldsymbol{V}|\boldsymbol{U})$.
- Submodularity: $h(\boldsymbol{V}|\boldsymbol{U}) \to h(\boldsymbol{V}|\boldsymbol{UW})$.
- No-Op: $h(\boldsymbol{U}) \to 0$.

We say that the CD-proof sequence *proves* the inequality (50) if its starts from the LHS and ends with the RHS. The following was proven in [KNS17]:

Lemma 7.5. *Inequality* (50) *is valid for polymatroids iff it admits a CD-proof sequence.*

PANDA converts a CD-proof sequence into an algorithm, similarly to the way we converted a BB-sequence to an algorithm. Given statistics $\Sigma, \boldsymbol{B}$, guarded by the query Q (Sec. 4.2), assume that inequality (50) is the optimal solution to Log-U-Bound$_{\Gamma_n}(Q, \Sigma, \log(\boldsymbol{B}))$; otherwise, choose a better inequality. Denote by h^* an optimal solution to Log-L-Bound$_{\Gamma_n}(Q, \Sigma, \log(\boldsymbol{B}))$. The algorithm has a working memory consisting of a guard, call it $S_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{Z})$, for every term $h(\boldsymbol{V}|\boldsymbol{U})$ in E, satisfying the following invariant: $\boldsymbol{V} \subseteq \boldsymbol{Z}$ and there exists a subset $\boldsymbol{U}_0 \subseteq \boldsymbol{U} \cap \boldsymbol{Z}$, such that:

$$\log \mathtt{deg}_{S_{\boldsymbol{V}|\boldsymbol{U}}}(\boldsymbol{V}|\boldsymbol{U}_0) \leq h^*(\boldsymbol{V}|\boldsymbol{U})$$

The guard need not have all variables $\boldsymbol{U}$, but only a subset $\boldsymbol{U}_0$ that is sufficient to prove the bound on the max-degree.

Initially, the working memory consists of all guards R_σ of the statistics $\sigma = (\boldsymbol{V}|\boldsymbol{U}) \in \Sigma$, where $k_\sigma > 0$. By complementary slackness, if $k_\sigma > 0$, then the corresponding constraint on $\boldsymbol{h}^*$ is tight, $h^*(\sigma) = b_\sigma (\stackrel{\text{def}}{=} \log B_\sigma)$, therefore $\log \mathtt{deg}_{R_\sigma}(\sigma) \leq b_\sigma = h^*(\sigma)$ because the input database satisfies the statistics. PANDA performs the following action for each step of the CD-proof sequence:

Composition $h(\boldsymbol{U}) + h(\boldsymbol{V}|\boldsymbol{U}) \rightarrow h(\boldsymbol{U}\boldsymbol{V})$. Compute the new guard as:

$$S_{\boldsymbol{U}\boldsymbol{V}} := \Pi_{\boldsymbol{U}}(S_{\boldsymbol{U}}) \bowtie \Pi_{\boldsymbol{U}_0\boldsymbol{V}}(S_{\boldsymbol{V}|\boldsymbol{U}})$$

Since $|S_{\boldsymbol{U}\boldsymbol{V}}| \leq |S_{\boldsymbol{U}}| \cdot \mathtt{deg}_{S_{\boldsymbol{V}|\boldsymbol{U}}}(\boldsymbol{V}|\boldsymbol{U}_0)$, we have:

$$\begin{aligned}\log |S_{\boldsymbol{U}\boldsymbol{V}}| \leq & \log \mathtt{deg}_{S_{\boldsymbol{U}|\emptyset}}(\boldsymbol{U}|\emptyset) + \log \mathtt{deg}_{S_{\boldsymbol{V}|\boldsymbol{U}}}(\boldsymbol{V}|\boldsymbol{U}_0) \\ \leq & h^*(\boldsymbol{U}) + h^*(\boldsymbol{V}|\boldsymbol{U}) = h^*(\boldsymbol{V})\end{aligned}$$

Thus, the invariant holds, and the runtime does not exceed the polymatroid bound, whose log is $h^*(\boldsymbol{X})$.

Submodularity $h(\boldsymbol{V}|\boldsymbol{U}) \rightarrow h(\boldsymbol{V}|\boldsymbol{U}\boldsymbol{W})$. Here PANDA only records that the new term $h(\boldsymbol{V}|\boldsymbol{U}\boldsymbol{W})$ has the same guard as the old term $h(\boldsymbol{V}|\boldsymbol{U})$.

Decomposition $h(\boldsymbol{U}\boldsymbol{V}) \rightarrow h(\boldsymbol{U}) + h(\boldsymbol{V}|\boldsymbol{U})$. Here PANDA first projects out the extra variables in the guard of $h(\boldsymbol{U}\boldsymbol{V})$ and obtains a relation $S := \Pi_{\boldsymbol{U}\boldsymbol{V}}(S_{\boldsymbol{U}\boldsymbol{V}})$ whose size $N \stackrel{\text{def}}{=} |S|$ satisfies $\log N \leq h^*(\boldsymbol{U}\boldsymbol{V})$. Next, it performs *regularization*: partition S into $\log N$ fragments $S = \bigcup_{i=1,\log N} S_i$, where:

$$S_i(\boldsymbol{U}, \boldsymbol{V}) \stackrel{\text{def}}{=} \{(\boldsymbol{u}, \boldsymbol{v}) \in S \mid \mathtt{deg}_S(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}) \in [2^{i-1}, 2^i]\}$$

PANDA then continues with $\log B$ recursive calls. The i'th recursive call replaces S with S_i in the query, adds two new statistics $(\boldsymbol{V}|\boldsymbol{U})$ and $(\boldsymbol{U})$ to Σ, and two log-statistics values, $b_{\boldsymbol{V}|\boldsymbol{U}} \stackrel{\text{def}}{=} i$ and $b_{\boldsymbol{U}} \stackrel{\text{def}}{=} N/2^{i-1}$, both with guard $S_i(\boldsymbol{U}\boldsymbol{V})$. Then, PANDA computes a *new* optimal primal/dual solutions to the polymatroid bound, resulting in a new inequality (50) and a new polymatroid $\boldsymbol{h}^*$. It uses these to compute the residual query where S is replaced by S_i. Finally, it returns the union of all $\log N$ results from all recursive calls.

We leave out several details of PANDA, including the proof of termination, and refer the reader to [KNS17]. We also note that PANDA was extended from computing full conjunctive queries, to computing Boolean conjunctive queries, with a runtime given by the *submodular width* of the query, a notion introduced by Marx [Mar13].

8 The Domination Problem

We now move beyond the query upper bound problem, and consider a related question, called the *domination problem*: given two queries Q, Q', check if, for any database $\boldsymbol{D}$, $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$. The queries Q and Q' need not have the same number of variables. In this section we consider full conjunctive queries that may have self-joins, i.e. the same relation name may occur several times in the query; for example in $R(X,Y) \wedge R(Y,Z)$ the same relation R occurs twice.

Definition 8.1. *Given two conjunctive queries $Q(\boldsymbol{X})$, $Q'(\boldsymbol{Y})$ we say that Q'* dominates *Q, and write $Q \preceq Q'$, if for every database instance $\boldsymbol{D}$, $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$.*

The original motivation for the domination problems comes from the query containment problem under bag semantics. Given a (not necessarily full) conjunctive query $Q(\boldsymbol{Y}_0)$, as in (1), its value under *bag semantics* is a bag of tuples, where each tuple $\boldsymbol{y}_0$ occurs as many times as the number of homomorphisms from Q to $\boldsymbol{D}$ that map $\boldsymbol{Y}_0$ to $\boldsymbol{y}_0$. SQL uses bag semantics. Chaudhuri and Vardi [CV93] were the first to study the *query containment problem under bag semantics*: given Q, Q', check whether $Q(\boldsymbol{D}) \subseteq Q'(\boldsymbol{D})$ for every $\boldsymbol{D}$, where both $Q(\boldsymbol{D})$, $Q'(\boldsymbol{D})$ are bags of tuples. This problem has been intensively studied in the last thirty years. It has been shown that the containment problem under bag semantics is undecidable for *unions of conjunctive queries* [IR95], and for *conjunctive queries with inequalities* [JKV06]; both used reduction from Hilbert's 10th Problem. It should be noted that, under set semantics, the containment problem for these two classes of queries is decidable.

When $Q() = \ldots$ is a Boolean query, then under standard set-semantics it returns either $\{\}$ or $\{()\}$, representing FALSE and TRUE. Under bag semantics it may return a bag $\{(), (), \ldots, ()\}$, representing a number, and this number is equal to the size of the output of the full query, $Q(\boldsymbol{X}) = \ldots$. Based on this discussion, the domination problem $Q \preceq Q'$ for full conjunctive queries is the same as the query containment problem under bag semantics for Boolean queries.

Kopparty and Rossman [KR11] were the first to establish the connection between the domination problem and information theory. We describe this connection, following their example.

Example 8.2. *This example from [KR11] is attributed to Eric Vee. Consider the following queries:*

$$\begin{aligned} Q(X,Y,Z) =& R(X,Y) \wedge R(Y,Z) \wedge R(Z,X) \\ Q'(U,V,W) =& R(U,V) \wedge R(U,W) \end{aligned}$$

We will show that $Q \preceq Q'$. Chaudhuri and Vardi [CV93] already noted that, if there exists a surjective homomorphism $Q' \to Q$, then $Q \preceq Q'$. In our example we have three homomorpisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$, but none of them is surjective.

Consider the following linear expression in entropic terms, defined over the variables U, V, W in Q':

$$E \stackrel{def}{=} h(UV) + h(UW) - h(U) = h(UV) + h(W|U)$$

(The expression is derived from the tree decomposition of Q', as we explain below.) For each of the three homomorphism φ_i, denote by $E \circ \varphi_i$ the result of substituting the variables U, V, W in E with $\varphi_i(U), \varphi_i(V), \varphi_i(W)$.

Claim 2. *The following inequality holds for all polymatroids:*

$$h(XYZ) \leq \max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \tag{51}$$

Proof. We expand:

$$\begin{aligned}
&\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) = \\
&\quad = \max(h(XY) + h(Y|X), h(YZ) + h(Z|Y), h(XZ) + h(X|Z)) \\
&\quad \geq \frac{1}{3}\left(h(XY) + h(YZ) + h(ZX) + h(Y|X) + h(Z|Y) + h(X|Z)\right) \\
&\quad = \frac{1}{3}\left((h(XY) + h(Z|Y)) + (h(YZ) + h(X|Z)) + (h(ZX) + h(Y|X))\right) \\
&\quad \geq h(XYZ)
\end{aligned}$$

where the last inequality follows from $h(XY) + h(Z|Y) \geq h(XY) + h(Z|XY) = h(XYZ)$ and similarly for the other two terms. $\square$

To prove $Q \preceq Q'$*, consider a database instance* $\boldsymbol{D}$*, let* $N \stackrel{def}{=} |Q(\boldsymbol{D})|$*, and consider the uniform probability distribution* $(Q(\boldsymbol{D}), p)$*. Its entropy* $\boldsymbol{h}$ *satisfies inequality* (51)*: assume w.l.o.g. that* $h(XYZ) \leq E \circ \varphi_1 = h(XY) + h(Y|X)$ *(the other two cases are similar). We use* φ_1 *to define a probability space* $(Q'(\boldsymbol{D}), p')$*: for every three constants* u, v, w *in the instance* $\boldsymbol{D}$ *s.t.* $p(X = u) \neq 0$*, define*

$$p'(U = u, V = v, W = w) \stackrel{def}{=} \frac{p(X = u, Y = v)p(X = u, Y = w)}{p(X = u)}$$

Thus, $V \perp W|U$*, the distribution of* UV *is the same as that of* XY*, and the distribution of* UW *is also the same as that of* XY*. (This is similar to the Copy Lemma 5.8.) Denoting by* $\boldsymbol{h}'$ *the entropic vector associated to* p'*, we derive:*

$$\begin{aligned}
\log|Q'(\boldsymbol{D})| \geq & h'(UVW) = h'(VW|U) + h(U) \\
= & h'(V|U) + h'(W|U) + h'(U) \quad \textit{because } V \perp W|U \\
= & h(Y|X) + h(Y|X) + h(X) = h(XY) + h(Y|X) \\
= & E \circ \varphi_1 \geq h(XYZ) = \log|Q(\boldsymbol{D})|
\end{aligned}$$

We generalize Example 8.2. A *tree decomposition* of a query $Q(\boldsymbol{X}) = \bigwedge_j R_j(\boldsymbol{Y}_j)$ is a pair (T, χ), where T is a tree and $\chi : \text{Nodes}(T) \to 2^{\boldsymbol{X}}$ such that every atom $R_j(\boldsymbol{Y}_j)$ is *covered*, meaning $\exists n$, $\boldsymbol{Y}_j \subseteq \chi(n)$, and for any variable $X \in \boldsymbol{X}$, the set of nodes $\{n \in \text{Nodes}(T) \mid X \in \chi(n)\}$ induces a connected subgraph of T. Each set $\chi(n)$ is called a *bag*. Q is *chordal* if it admits a tree decomposition where every bag $\chi(n)$ induces a clique in the Gaifman graph of Q; equivalently, for any two variables $X, Y \in \chi(n)$ the query has a predicate that contains both X, Y. A chordal query has a canonical tree decomposition where the bags are the maximal cliques. Q is called *acyclic*[4] if there exists a tree decomposition where each bag is precisely one atom of the query, $\chi(n) = \boldsymbol{Y}_j$ for some j. An acyclic query is, in particular, chordal.

Fix a query $Q(\boldsymbol{U})$ with variables $\boldsymbol{U}$ and a tree decomposition T. We define the following expression of entropic terms:

$$E_T \stackrel{\text{def}}{=} \sum_{n \in \text{Nodes}(T)} h(\chi(n)) - \sum_{(n,n') \in \text{Edges}(T)} h(\chi(n) \cap \chi(n')) \tag{52}$$

[4]More precisely, it is called α-acyclic [Fag83].

Equivalently, choose a root node for T and orient all edges to point away from the root. Then:

$$E_T = \sum_{n \in \text{Nodes}(T)} h(\chi(n)|\chi(n) \cap \chi(\text{Parent}(n)))$$

where we set $\chi(\text{Parent}(\text{Root})) \stackrel{\text{def}}{=} \emptyset$. The following holds:

Theorem 8.3. *Let $Q(\boldsymbol{X}), Q'(\boldsymbol{U})$ be two full conjunctive queries, over variables $\boldsymbol{X}$ and $\boldsymbol{U}$ respectively.*

- *[KR11, KKNS21] Let T be a tree decomposition for Q'. If the following inequality holds for all entropic vectors $\boldsymbol{h}$:*

$$h(\boldsymbol{X}) \leq \max_{\varphi:\varphi\in\text{hom}(Q',Q)} E_T \circ \varphi \tag{53}$$

  then Q' dominates Q, $Q \preceq Q'$.

- *[KKNS21] If Q' is chordal and $Q \preceq Q'$, then inequality (53) holds for all entropic vectors, where T is the canonical tree decomposition of Q' consisting of its maximal cliques. In other words, (53) is a necessary and sufficient condition for dominance.*

Let's call a tree decomposition T *simple* if for every edge $(n, n') \in \text{Edges}(T)$, $|\chi(n) \cap \chi(n')| \leq 1$. If Q' admits a simple tree decomposition, then condition (53) is decidable; the proof follows immediately from Lemma 6.8. This implies:

Corollary 8.4. *Assume that Q' is chordal and admits a simple tree decomposition. Then it is decidable whether $Q \preceq Q'$. Moreover, if $Q \not\preceq Q'$, then there exists a normal database instance (Sec. 6.3) such that $|Q(\boldsymbol{D})| > |Q'(\boldsymbol{D})|$.*

Finally, we remark that the connection between the query domination problem and information inequalities is very tight. The following was proven in [KKNS21]:

Theorem 8.5. *The following problems are computationally equivalent. (1) Check if an inequality of the form $\max_{j=1,p}(\boldsymbol{c}^{(j)} \cdot \boldsymbol{h}) \geq 0$ is valid for all entropic vectors $\boldsymbol{h}$, where $\boldsymbol{c}^{(1)}, \ldots \boldsymbol{c}^{(p)} \in \mathbb{R}^{2^{[n]}}$ are p vectors. (2) Given two queries Q, Q' where Q' is acyclic, check whether $Q \preceq Q'$.*

It is currently open whether these problems are decidable.

9 Conditional Inequalities and Approximate Implication

Our last application of information inequalities is for the approximate implication problem, which can be described informally as follows. Let $\sigma_1, \ldots, \sigma_p$ be some constraints on the database (we will define shortly what constraints we consider), and suppose we have a proof of the implication $\bigwedge_i \sigma_i \Rightarrow \sigma_0$. The question is, if the database $\boldsymbol{D}$ satisfies the constraints σ_i only approximatively, is it the case that that σ_0 also holds approximatively? We will show here that this question is related to *conditional information inequalities*, whose study requires us to do another deep dive into the space of polymatroids and entropic functions. We start by defining a conditional inequality:

Definition 9.1. *A* conditional information inequality *is an assertion of the following form:*

$$\boldsymbol{c}_1 \cdot \boldsymbol{h} \geq 0 \wedge \cdots \boldsymbol{c}_p \cdot \boldsymbol{h} \geq 0 \Rightarrow \boldsymbol{c}_0 \cdot \boldsymbol{h} \geq 0 \tag{54}$$

where $\boldsymbol{c}_i \in \mathbb{R}^{2^{[n]}}$ for $i = 0, p$ are vectors.

Sometimes it will be more convenient to replace $\boldsymbol{c}_i$ by $-\boldsymbol{c}_i$, and write the implication as $\bigwedge_i \boldsymbol{c}_i \cdot \boldsymbol{h} \leq 0 \Rightarrow \boldsymbol{c}_0 \cdot \boldsymbol{h} \leq 0$. As before, the validity of a conditional inequality depends on the domain of $\boldsymbol{h}$, e.g. it can be valid for polymatroids, or entropic functions, etc.

The first non-Shannon inequality discovered was a conditional inequality [ZY97], predating the first unconditioned Shannon inequality [ZY98]. Kaced and Romashchenko [KR13] showed the first examples of essentially conditional inequalities (explained below). We start by describing the connection between conditional inequalities and the constraint implication problem in databases, then study the *relaxation problem*, a technique for transferring exact inferences to approximate judgments. We end with the proof of Theorem 5.4, which we have postponed until we developed sufficient technical machinery.

9.1 The Constraint Implication Problem

An *integrity constraint*, σ, is an assertion about a relation $R(\boldsymbol{X})$ that is required to hold strictly. The constraints considered here are Functional Dependencies (FD), already reviewed in Sec. 2, and Multivalued Dependencies (MVD). An *MVD* is a statement $\sigma = (\boldsymbol{U} \twoheadrightarrow \boldsymbol{V}|\boldsymbol{W})$ where $\boldsymbol{U} \cup \boldsymbol{V} \cup W$ form a partition of $\boldsymbol{X}$. A relation instance R *satisfies* the MVD, $R \models \sigma$, if $R = \Pi_{\boldsymbol{UV}}(R) \bowtie \Pi_{\boldsymbol{UW}}(R)$.

The *implication problem* asks whether a set of FDs and/or MVDs σ_i, $i = 1, p$, implies another FD or MVD σ_0:

$$\sigma_1 \wedge \cdots \wedge \sigma_p \Rightarrow \sigma_0 \tag{55}$$

Armstrong's axioms [AD80] are complete for the implication problem for FDs, while Beeri et al. [BFH77] gave a complete axiomatization for both FDs and MVDs, and showed that the implication problem is decidable. In contrast, Herrmann [Her06] showed that the implication problem of *Embedded MVDs* is undecidable; we do not discuss EMVDs here.

Lee [Lee87a] showed the following connection between information theory and constraints. Fix a relational instance R, and let $\boldsymbol{h}$ be its associated (uniform) entropic vector. Then $R \models \boldsymbol{U} \rightarrow \boldsymbol{V}$ iff $h(\boldsymbol{V}|\boldsymbol{U}) = 0$, and $R \models \boldsymbol{U} \twoheadrightarrow \boldsymbol{V}|\boldsymbol{W}$ iff $I_h(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U}) = 0$. Therefore, every implication problem for FDs and MVDs can be stated as a conditional information inequality. For example, the *augmentation axiom* [BFH77] states

$$(A \twoheadrightarrow B|CD) \Rightarrow (AC \twoheadrightarrow B|D)$$

and is equivalent to the following conditional inequality:[5]

$$I(B; CD|A) = 0 \Rightarrow I(B; D|AC) = 0 \tag{56}$$

This can be proven immediately by observing that the identity $I(B; CD|A) = I(B; C|A) + I(B; D|AC)$ implies:

$$I(B; D|AC) \leq I(B; CD|A)$$

Since both terms are ≥ 0, the implication (56) follows.

Beyond database applications, Conditional Independencies (CI) are commonly used in AI, Knowledge Representation, and Machine Learning. A CI is an assertions of the form $X \perp Y \mid Z$, where X, Y, Z are three random variables, stating that X is independent of Y conditioned on Z. The AI community has extensively studied the implication problem for CIs. It was shown that the implication problem is decidable and finitely axiomatizable for *saturated* CIs [GP93] (where $XYZ =$ all variables), but not finitely axiomatizable in general [Stu90].

[5] This has the form in Def. 9.1 once we write it as $-I(B; CD|A) \geq 0$ implies $-I(B; D|AC) \geq 0$.

9.2 The Relaxation Problem

How can we prove a conditional inequality (54)? One approach is as follows. Find p non-negative real numbers $\lambda_1, \ldots, \lambda_p$ for which the following inequality is valid:

$$\boldsymbol{c} \cdot \boldsymbol{h} \geq \left(\sum_{i=1,p} \lambda_i \boldsymbol{c}_i \right) \cdot \boldsymbol{h} \tag{57}$$

Then, observe that (57) implies (54). A natural question is whether every conditional information inequality can be derived in this way, from an unconditional inequality: when that is the case, then we say that the conditional inequality (54) *relaxes* to (57), or that it is *essentially unconditional*. Otherwise we say that it is *essentially conditional*. For example, the augmentation axiom above can be relaxed, hence it is essentially unconditioned.

Besides offering an important proof technique, relaxation is important in modern database applications, because often the integrity constraints don't hold *exactly*, but only *approximatively*, especially when they are mined from a given dataset [GR04, SBHR06, CIPY14, BBF$^+$16, KN18, SGS18, PD11, FD16]. The relaxation problem allows us to transfer proofs over exact constraints to approximate constraints.

Every implication problem for FDs and MVDs relaxes [KS22]:

Theorem 9.2. *Consider the statement* (55) *asserting the implication between a set of FDs/MVDs. Consider the associated conditional information inequality:*

$$\left(\bigwedge_i h(\sigma_i) = 0 \right) \Rightarrow (h(\sigma_0) = 0) \tag{58}$$

where each expression $h(\sigma_i)$ represents either $I_h(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U})$ or $h(\boldsymbol{V}|\boldsymbol{U})$. Then the following are equivalent:

- *The implication* (55) *holds for FDs/MVDs.*
- *The implication* (58) *holds for all polymatroids.*
- *The implication* (58) *holds for all entropic functions.*
- *The implication* (58) *holds for all normal polymatroids.*
- *The inequality $n^2/4(\sum_i h(\sigma_i)) \geq h(\sigma_0)$ holds for all polymatroids. Furthermore, if σ_0 is an FD (rather than MVD), then $n^2/4$ can be replaced by 1.*

Thus, the implication problem for FDs and MVDs relaxes. A consequence of the theorem is that, if (58) fails, then there exists a relation with only two tuples falsifying the implication, namely one of the relations $R_{\boldsymbol{W}}$ in (41) associated to a step function.

Next, we examine whether all conditional inequalities relax. To do this, we need another (last) deep dive into the structure of entropic functions, and introduce *almost-entropic functions*.

9.3 Background: Almost Entropic Functions

We start with a brief review of cones, following [Sch03, BV04]. A set $K \subseteq \mathbb{R}^n$ is called a *cone*, if $\boldsymbol{x} \in K$ and $\theta \geq 0$ implies $\theta \boldsymbol{x} \in K$. The cone is *convex* if $\boldsymbol{x}_1, \boldsymbol{x}_2 \in K$ and $\theta \in [0, 1]$ implies $\theta \boldsymbol{x}_1 + (1 - \theta)\boldsymbol{x}_2 \in K$. For any set $K \subseteq \mathbb{R}^n$, we denote by $\bar{K}$ its *topological closure*, and by

$K^* \stackrel{\text{def}}{=} \{\boldsymbol{y} \mid \forall \boldsymbol{x} \in K, \boldsymbol{y}^T \cdot \boldsymbol{x} \geq 0\}$, it's *dual*. The dual is always a closed, convex cone, $K \subseteq K^{**}$ and, when K is a closed, convex cone, then $K = K^{**}$.

A cone is *polyhedral* if it has the form $K = \{\boldsymbol{x} \mid \boldsymbol{M} \cdot \boldsymbol{x} \geq 0\}$, for some matrix $\boldsymbol{M} \in \mathbb{R}^{m \times n}$. Any polyhedral cone is closed and convex, and its dual is also polyhedral.

The set of polymatroids Γ_n and of entropic vectors Γ_n^* are subsets of $\mathbb{R}^{2^n}$. The superscript $*$ in Γ_n^* is an unfortunate notation, since it does not represent a dual, but this notation is already widely used. Valid inequalities (Def. 4.2) are the dual cones, $(\Gamma_n^*)^*$, and $(\Gamma_n)^*$ respectively. Clearly, Γ_n is polyhedral, and therefore $\Gamma_n = \Gamma_n^{**}$. What about Γ_n^*?

It turns out that, when $n \geq 3$, then Γ_n^* is neither a cone nor convex. This may come as a surprise, so we take the opportunity to briefly review here the elegant proof by Zhang and Yeung [ZY97]. Consider the parity function h, shown in Fig. 2, and observe that $h(X) = 1$, $h(Z|XY) = 0$, and $I_h(X;Y) = 0$. If c is a natural number, the vector $c \cdot \boldsymbol{h}$ is also entropic, by Prop. 5.3; but in general, $c \cdot \boldsymbol{h}$ is entropic only if there exists a natural number N such that $c = \log N$, which implies Γ_n^* is neither a cone, nor convex. To prove this, assume that $c \cdot \boldsymbol{h}$ is entropic, and realized by a probability distribution $p(X, Y, Z)$. Choose any two values x, y such that $p(X = x) > 0$ and $p(Y = y) > 0$. Since $I_{c\boldsymbol{h}}(X;Y) = c \cdot I_h(X;Y) = 0$, it holds that $X \perp Y$, hence $p(X = x)p(Y = y) = p(X = x, Y = y) > 0$. Furthermore, $c \cdot h(Z|XY) = 0$, therefore p satisfies the functional dependency $XY \to Z$, and there exists a unique value z s.t. $p(X = x, Y = y, Z = z) > 0$. We have obtained $p(X = x)p(Y = y) = p(X = x, Y = y, Z = z)$ and, by symmetry, it also holds that $p(X = x)p(Z = z) = p(X = x, Y = y, Z = z)$. This implies $p(Y = y) = p(Z = z)$. Since y was arbitrary, it follows that $p(Y = y) = p(Y = y')$ for all y, y' in the support of Y. Therefore, the marginal distribution of Y is uniform, and its entropy is $c \cdot h(Y) = \log N$, where N is the size of the support, proving $c = \log N$, since $h(Y) = 1$. Recall that in Sec. 4.2 we stated that the conditional entropy $h(-|\boldsymbol{U})$ is not always an entropic vector: we invite the reader to give such an example.

While Γ_n^* is neither a cone nor convex, Yeung [Yeu08] proved:

Theorem 9.3. *The topological closure $\bar{\Gamma}_n^*$ of Γ_n^* is a closed, convex cone. A vector $\boldsymbol{h} \in \bar{\Gamma}^*$ is called* almost-entropic*.*

The complete picture of all sets of polymatroids discussed in this paper is shown in Fig. 3.

If an inequality is valid for Γ_n^*, then it is also valid for $\bar{\Gamma}_n^*$, by continuity. However, this no longer holds for *conditional* information inequalities: Kaced and Romashchenko [KR13] gave an example of a conditional inequality that is valid for Γ_n^*, but not for $\bar{\Gamma}_n^*$. Since we are interested in the relaxation problem, we will consider only validity for $\bar{\Gamma}_n^*$.

When $n = 3$ then one can show that $\bar{\Gamma}_3^* = \Gamma_3$, hence it is polyhedral. However, Matúš [Mat07] showed that, for $n \geq 4$, $\bar{\Gamma}_n^*$ is not polyhedral. This explains the difficulties in understanding the non-Shannon inequalities, and also in reasoning about conditional inequalities.

9.4 A Conditional Inequality that Does Not Relax

Kaced and Romashchenko [KR13] gave four examples of essentially conditional inequalities. This is a very surprising result: it means that a proof of the implication (55) becomes useless if one of the assumptions has even a tiny violation in the data. We describe here one of their examples, following the adaptation in [KS22].

Theorem 9.4. *[KR13] The following conditional inequality is valid for $\bar{\Gamma}_n^*$, and is essentially conditional:*

$$I(X;Y|A) = I(X;Y|B) = I(A;B) = I(A;X|Y) = 0 \Rightarrow I(X;Y) = 0$$

Notice that none of the constraints is an FD or an MVD, so this does not contradict Theorem 9.2. We prove the theorem, and start by proving the conditional inequality. For that we use the following non-Shannon inequality, by Matús [Mat07]. For every $k \geq 1$,

$$\begin{aligned} I(X;Y) \leq & \frac{k+3}{2} I(X;Y|A) + I(X;Y|B) + I(A;B) \\ & + \frac{k+1}{2} I(A;X|Y) + \frac{1}{k} I(A;Y|X) \end{aligned} \tag{59}$$

When $k = 1$, this is Zhang and Yeung's inequality (32). Matús proved (59) by induction on k, by applying the Copy Lemma 5.8 at each induction step; we omit the proof.

Since (59) holds for Γ_n^*, it also holds for $\bar{\Gamma}_n^*$. To check the conditional inequality in Thm. (9.4), let $\boldsymbol{h} \in \bar{\Gamma}_n^*$ such that $I(X;Y|A) = I(X;Y|B) = I(A;B) = I(A;X|Y) = 0$: then inequality (59) becomes $I(X;Y) \leq \frac{1}{k} I(A;Y|X)$ and, since k is arbitrary, it follows that $I(X;Y) = 0$.

Finally, we show that the conditional inequality does not relax, by describing, for each $\lambda > 0$, an entropic vector $\boldsymbol{h}$ s.t.

$$I(X;Y) \geq \lambda(I(X;Y|A) + I(X;Y|B) + I(A;B) + I(A;X|Y)) \tag{60}$$

Let $\boldsymbol{h}$ be the entropy of the following distribution:

A	B	X	Y	p
0	0	0	0	$1/2 - \varepsilon$
1	0	0	1	$1/2 - \varepsilon$
0	1	1	0	ε
1	1	0	0	ε

If ε is small enough then one can check:[6]

$$\begin{aligned} I(X;Y) =& \varepsilon + O(\varepsilon^2), \; I(A;X|Y) = O(\varepsilon^2), \\ I(X;Y|A) =& I(X;Y|B) = I(A;B) = 0 \end{aligned}$$

which proves (60) for ε is small enough.

9.5 Conditional Inequalities Relax with Error Terms

However, in another twist, it turns out that every conditional inequality relaxes, if we admit a small error term. The following was proven in [KS22]:

Theorem 9.5. *Suppose that the following holds:*

$$\forall \boldsymbol{h} \in \bar{\Gamma}_n^*, \quad \left(\bigwedge_{i=1,p} \boldsymbol{c}_i \cdot \boldsymbol{h} \leq 0 \right) \Rightarrow \boldsymbol{c}_0 \cdot \boldsymbol{h} \leq 0 \tag{61}$$

Then, for every $\varepsilon > 0$ there exists $\lambda_1, \ldots, \lambda_p \geq 0$ such that:

$$\forall \boldsymbol{h} \in \bar{\Gamma}_n^*, \quad \boldsymbol{c}_0 \cdot \boldsymbol{h} \leq \left(\sum_{i=1,p} \lambda_i \boldsymbol{c}_i \cdot \boldsymbol{h} \right) + \varepsilon h(\boldsymbol{X}) \tag{62}$$

where $\boldsymbol{X}$ is the set of all n variables.

[6] Complete calculations are included in [KS22]; note that here we have swapped the roles of A and B, in order to better draw the connection to Zhang and Yeung's inequality (32).

Even with the error term, condition (62) still implies (61), because, if $\boldsymbol{c}_i \cdot \boldsymbol{h} \leq 0$ for all i, then (62) implies $\boldsymbol{c}_0 \cdot \boldsymbol{h} \leq \varepsilon h(\boldsymbol{X})$ and, since ε is arbitrary, we obtain $\boldsymbol{c}_0 \cdot \boldsymbol{h} \leq 0$. In fact, Matús' inequality (59) can be seen as a relaxation, with an error term, of the conditional inequality in Theorem 9.4. Theorem 9.5 shows that this was not accidental: *every* conditional inequality follows from an unconditional with an error term that tends to 0. In the next section we will show that Theorem 9.5 has a surprising application, to the proof of Theorem 5.4. Before that, we prove Theorem 9.5 by showing:

Lemma 9.6. *[KS22] Let $K \subseteq \mathbb{R}^n$ be a closed, convex cone, and $\boldsymbol{c}_i \in \mathbb{R}^n$, $i = 0, p$ be vectors such that the following holds:*

$$\forall \boldsymbol{x} \in K : \quad \left(\bigwedge_{i=1,p} \boldsymbol{c}_i \cdot \boldsymbol{x} \leq 0 \right) \Rightarrow \boldsymbol{c}_0 \cdot \boldsymbol{x} \leq 0 \tag{63}$$

Then, for every $\varepsilon > 0$, there exists $\lambda_1, \ldots, \lambda_p \geq 0$ such that:

$$\forall \boldsymbol{x} \in K : \quad \boldsymbol{c}_0 \cdot \boldsymbol{x} \leq \left(\sum_{i=1,p} \lambda_i \boldsymbol{c}_i \right) \cdot \boldsymbol{x} + \varepsilon ||\boldsymbol{x}||_\infty \tag{64}$$

The lemma implies the theorem, because $\bar{\Gamma}_n^*$ is a closed, convex cone, and $||\boldsymbol{h}||_\infty = h(\boldsymbol{X})$.

Proof. (of Lemma 9.6) Let $L \stackrel{\text{def}}{=} \{-\boldsymbol{c}_i \mid i = 1, p\}$. Then condition (63) says that $\boldsymbol{x} \in K \cap L^*$ implies $\boldsymbol{c}_0 \cdot \boldsymbol{x} \leq 0$, or, equivalently, $-\boldsymbol{c}_0 \in (K \cap L^*)^*$. The following holds for any closed, convex cones K_1, K_2 (see [KS22, Sec.5.3]):

$$(K_1 \cap K_2)^* = \overline{\textbf{conhull}(K_1^* \cup K_2^*)}$$

where $\textbf{conhull}(A)$ is the conic hull of a set A, i.e. the set of positive, linear combinations of vectors in A. Also, if L is finite, then $L^{**} = \textbf{conhull}(L)$. Therefore, $-\boldsymbol{c}_0$ belongs to the following set:

$$\begin{aligned}(K \cap L^*)^* =& \overline{\textbf{conhull}(K^* \cup L^{**})} \\ =& \overline{\textbf{conhull}(K^* \cup \textbf{conhull}(L))} \\ =& \overline{\textbf{conhull}(K^* \cup L)}\end{aligned}$$

For any $\varepsilon > 0$, there exists $\boldsymbol{e} \in \mathbb{R}^n$ such that $||\boldsymbol{e}||_1 < \varepsilon$ and:

$$-\boldsymbol{c}_0 + \boldsymbol{e} \in \textbf{conhull}(K^* \cap L)$$

By the definition of the conic hull, and the fact that K^* is a convex cone, we obtain that there exists $\boldsymbol{d} \in K^*$ and $\lambda_i \geq 0$, for $i = 1, p$ such that:

$$-\boldsymbol{c}_0 + \boldsymbol{e} = \boldsymbol{d} - \sum_{i=1,p} \lambda_i \boldsymbol{c}_i$$

We prove (64). Let $\boldsymbol{x} \in K$, and observe that $\boldsymbol{d} \cdot \boldsymbol{x} \geq 0$, then:

$$\left(\sum_{i=1,p} \lambda_i \boldsymbol{c}_i \right) \cdot \boldsymbol{x} - \boldsymbol{c}_0 \cdot \boldsymbol{x} + \boldsymbol{e} \cdot \boldsymbol{x} = \boldsymbol{d} \cdot \boldsymbol{x} \geq 0$$

and (64) follows from $\boldsymbol{e} \cdot \boldsymbol{x} \leq ||\boldsymbol{e}||_1 \cdot ||\boldsymbol{x}||_\infty \leq \varepsilon ||\boldsymbol{x}||_\infty$. $\square$

We end this section with a brief discussion of why Theorem 9.5 only holds for the set of almost entropic functions $\bar{\Gamma}^*_n$, and fails for Γ^*_n. Kaced and Romashchenko [KR13] gave an example of a conditional inequality that is valid for Γ^*_n, but not for $\bar{\Gamma}^*_n$. If Theorem 9.5 were to hold for that inequality, then the relaxed inequality (62) holds for $\bar{\Gamma}^*_n$, and then we could prove that the conditional inequality also holds for $\bar{\Gamma}^*_n$.

9.6 Proof of Theorem 5.4

We have now the machinery needed to prove Theorem 5.4, which states that $\text{Log-L-Bound}_{\Gamma^*_n}$ and $\text{Log-U-Bound}_{\Gamma^*_n}$ are asymptotically equal. What makes the proof a little difficult is the ill-behaved nature of the entropic functions Γ^*_n. To cope with that, use some help from the almost entropic functions $\bar{\Gamma}^*_n$, and prove that, here, the two bounds are equal (not just asymptotically). We state and prove the theorem for an arbitrary closed, convex cone K:

Theorem 9.7. *Fix $Q, \Sigma, \boldsymbol{b}$, and let K be a closed, convex cone s.t. $N_n \subseteq K \subseteq \Gamma_n$. (Recall that N_n is the set of normal polymatroids, Def. 6.2.) Then:*

$$\textit{Log-L-Bound}_K(Q, \Sigma, \boldsymbol{b}) = \textit{Log-U-Bound}_K(Q, \Sigma, \boldsymbol{b})$$

In particular, the lower and upper bounds are equal for $K = \bar{\Gamma}^*_n$, while they are not necessarily equal for Γ^*_n. It shows that the set $\bar{\Gamma}^*_n$ is better behaved that Γ^*_n, and that explains why it was used in prior work [GT17, KNS17] to study lower bounds.

Before we prove the theorem, we show how to use it to prove Theorem 5.4. Its proof follows from Theorem 9.7 and two identities, (65) and (66), which we prove below. The first is very simple:

$$\text{Log-U-Bound}_{\bar{\Gamma}^*_n}(Q, \Sigma, \boldsymbol{b}) = \text{Log-U-Bound}_{\Gamma^*_n}(Q, \Sigma, \boldsymbol{b}) \tag{65}$$

and follows directly from the fact that $\bar{\Gamma}^*_n$ and Γ^*_n have the same dual cone, $(\bar{\Gamma}^*_n)^* = (\Gamma^*_n)^*$; in other words, they define the same set of valid inequalities $\boldsymbol{c} \cdot \boldsymbol{h} \geq 0$.

The second equality requires a proof, and we state it as a lemma:

Lemma 9.8. *The following holds:*

$$\sup_k \frac{\textit{Log-L-Bound}_{\Gamma^*_n}(Q, \Sigma, k\boldsymbol{b})}{\textit{Log-L-Bound}_{\bar{\Gamma}^*_n}(Q, \Sigma, k\boldsymbol{b})} = 1 \tag{66}$$

Proof. The proof is similar to that of Theorem 5.5 item (1) in Sec. 5.2. The LHS is obviously ≤ 1. To prove that it is ≥ 1, denote by[7] $L \stackrel{\text{def}}{=} \text{Log-L-Bound}_{\bar{\Gamma}^*_n}(Q, \Sigma, \boldsymbol{b})$, and observe that $\text{Log-L-Bound}_{\bar{\Gamma}^*_n}$ is linear in $\boldsymbol{b}$, $\text{Log-L-Bound}_{\bar{\Gamma}^*_n}(Q, \Sigma, k\boldsymbol{b}) = kL$, because $\bar{\Gamma}^*_n$ is a convex cone. It suffices to show that, for all $\varepsilon > 0$, there exists $\boldsymbol{h} \in \Gamma^*_n$ such that $\boldsymbol{h} \models (\Sigma, k\boldsymbol{b})$ and $h(\boldsymbol{X}) \geq (1-\varepsilon)^2 kL$.

Start with some $\boldsymbol{h} \in \bar{\Gamma}^*_n$ satisfying:

$$\forall \sigma \in \Sigma: \ h(\sigma) \leq b_\sigma, \qquad\qquad h(\boldsymbol{X}) = L$$

which exists by the definition of $\text{Log-L-Bound}_{\bar{\Gamma}^*_n}$ (Def. 5.1) and the fact that the set $\{\boldsymbol{h} \in \mathbb{R}^{2^{[n]}} \mid ||\boldsymbol{h}||_\infty \leq L\}$ is compact. Chan and Yeung's Theorem 5.9 proves that Υ_n is dense in Γ^*_n, and therefore it is also dense in $\bar{\Gamma}^*_n$. Assume w.l.o.g. that $h(\boldsymbol{X}) > 0$, then $g \stackrel{\text{def}}{=} \min_{\boldsymbol{U},\boldsymbol{V}: h(\boldsymbol{V}|\boldsymbol{U})>0} h(\boldsymbol{V}|\boldsymbol{U}) > 0$ (the

[7] If $\text{Log-L-Bound}_{\bar{\Gamma}^*_n}(Q, \Sigma, \boldsymbol{b}) = \infty$ then we choose L an arbitrarily large number, and make minor adjustments to the proof; we omit the details.

smallest non-zero value of any expression $h(\boldsymbol{V}|\boldsymbol{U})$). Let $\delta \stackrel{\text{def}}{=} \varepsilon g/4$. Since Υ_n is dense, there exists $r \in \mathbb{N}$ and $\boldsymbol{h}^{(r)} \in \Upsilon_n$ such that $||\frac{1}{r}\boldsymbol{h}^{(r)} - \boldsymbol{h}||_\infty \leq \delta$. We have $h(\boldsymbol{X}) > 0$ hence $h(\boldsymbol{X}) \geq g$, therefore:

$$\begin{aligned}\frac{1}{r}h^{(r)}(\boldsymbol{X}) \geq & h(\boldsymbol{X}) - \delta = h(\boldsymbol{X}) - \varepsilon g/4 \\ \geq & h(\boldsymbol{X}) - (\varepsilon/4)h(\boldsymbol{X}) \geq (1-\varepsilon)h(\boldsymbol{X}) \\ = & (1-\varepsilon)L\end{aligned}$$

We claim that $\frac{1}{r}h^{(r)}(\boldsymbol{V}|\boldsymbol{U}) \leq (1+\varepsilon/2)h(\boldsymbol{V}|\boldsymbol{U})$, for all $\boldsymbol{U}, \boldsymbol{V}$. If $h(\boldsymbol{V}|\boldsymbol{U}) = 0$, then $\boldsymbol{h}$ satisfies the FD $\boldsymbol{U} \to \boldsymbol{V}$, and therefore $\boldsymbol{h}^{(r)}$ also satisfies this FD (see the note after Theorem 5.9), implying $h^{(r)}(\boldsymbol{V}|\boldsymbol{U}) = 0$. Otherwise, $h(\boldsymbol{V}|\boldsymbol{U}) \geq g$ and,

$$\begin{aligned}\frac{1}{r}h^{(r)}(\boldsymbol{V}|\boldsymbol{U}) \leq & h(\boldsymbol{V}|\boldsymbol{U}) + 2\delta = h(\boldsymbol{V}|\boldsymbol{U}) + \varepsilon g/2 \\ \leq & (1+\varepsilon/2)h(\boldsymbol{V}|\boldsymbol{U})\end{aligned}$$

Therefore, we have:

$$\forall \sigma \in \Sigma: \; h^{(r)}(\sigma) \leq (1+\varepsilon/2)rb_\sigma \qquad\qquad h^{(r)}(\boldsymbol{X}) \geq (1-\varepsilon)rL$$

Finally, by the Slack lemma, $\exists k \in \mathbb{N}, \boldsymbol{h}' \in \Gamma_n^*$ such that:

$$\begin{aligned}\forall \sigma \in \Sigma: \; h'(\sigma) \leq & (1-\varepsilon/2)kh^{(r)}(\sigma) \leq (1-(\varepsilon/2)^2)krb_\sigma \leq krb_\sigma \\ h'(\boldsymbol{X}) \geq & (1-\varepsilon)kh^{(r)}(\boldsymbol{X}) \geq (1-\varepsilon)^2 krL\end{aligned}$$

This completes the proof. □

In the remainder of this section we prove Theorem 9.7.

Proof. (of Theorem 9.7) We will use the following definition from [BV04, Example 5.12]:

Definition 9.9. *Let K be a proper cone (meaning: closed, convex, with a non-empty interior, and pointed i.e. $\boldsymbol{x}, -\boldsymbol{x} \in K$ implies $\boldsymbol{x} = 0$). A primal/dual cone program in standard form*[8] *is the following:*

Primal		*Dual*	
Maximize	$\boldsymbol{c}^T \cdot \boldsymbol{x}$	*Minimize*	$\boldsymbol{y}^T \cdot b$
where	$\boldsymbol{A} \cdot \boldsymbol{x} = \boldsymbol{b}$	*where*	$(\boldsymbol{y}^T \cdot \boldsymbol{A} - \boldsymbol{c}^T)^T \in K^*$
	$\boldsymbol{x} \in K$		

Denote by P^*, D^* the optimal value of the primal and dual respectively. *Weak duality* states that $P^* \leq D^*$, and is easy prove. When *Slater's condition* holds, which says that there exists $\boldsymbol{x}$ in the interior of K such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, then *strong duality* holds too: $P^* = D^*$.

Log-L-Bound$_K(Q, \Sigma, \boldsymbol{b})$ and Log-U-Bound$_K(Q, \Sigma, \boldsymbol{b})$ can be expressed as a cone program, by letting $\boldsymbol{A}$ and $\boldsymbol{c}$ be the matrix and vector defined in the proof of Theorem 5.2 (thus, $\boldsymbol{A} \cdot \boldsymbol{h} = (h(\sigma))_{\sigma \in \Sigma}$ and $c_{\boldsymbol{X}} = 1$, $c_{\boldsymbol{U}} = 0$ for $\boldsymbol{U} \neq \boldsymbol{X}$):

Log-L-Bound$_K$		Log-U-Bound$_K$	
Maximize	$\boldsymbol{c}^T \cdot h$	Minimize	$\boldsymbol{w}^T \cdot \boldsymbol{b}$
where	$\boldsymbol{A} \cdot \boldsymbol{h} + \boldsymbol{\beta} = \boldsymbol{b}$	where	$(\boldsymbol{w}^T \cdot \boldsymbol{A} - \boldsymbol{c}^T)^T \in K^*$
	$(\boldsymbol{h}, \boldsymbol{\beta}) \in K \times \mathbb{R}_+^s$		$\boldsymbol{w} \geq 0$

(67)

[8] We changed to the original formulation [BV04] by replacing $\boldsymbol{c}$ with $-\boldsymbol{c}$, replacing $\boldsymbol{y}$ with $-\boldsymbol{y}$.

Here $\boldsymbol{\beta}$ are slack variables that convert an inequality $h(\sigma) \leq b_\sigma$ into an equality $h(\sigma) + \beta_\sigma = b_\sigma$. We leave it to the reader to check that these two programs are indeed primal/dual as in Def. 9.9.

However, in general, Slater's condition need not hold for (67). For example, $K = \bar{\Gamma}^*_n$ lies in the hyperplane $h(\emptyset) = 0$, and thus has an empty interior. This could be addressed by removing the $\emptyset$ dimension, but we have a bigger problem. Some of the constraints may be tight: when $b_\sigma = 0$, then $h(\sigma) = 0$, meaning that no feasible solution $\boldsymbol{h}$ exists in the interior of K. Instead, we will define a different cone, K_0, by using polymatroids on a lattice, as in Sec. 5.1.

Partition Σ into $\Sigma_0 = \{\sigma \mid b_\sigma = 0\}$ and $\Sigma_1 = \{\sigma \mid b_\sigma > 0\}$, and denote by $\boldsymbol{b}_1$ the restriction of $\boldsymbol{b}$ to Σ_1. In other words, Σ_0 defines a set of functional dependencies, while $(\Sigma_1, \boldsymbol{b}_1)$ defines non-tight statistics. Let $(L_{\Sigma_0}, \subseteq)$ be the lattice of the closed sets of Σ_0 (defined in Sec. 2); we will drop the subscript and write simply $(L, \subseteq)$ to reduce clutter. Define $F \subseteq \mathbb{R}^{2^{[n]}}$ the following cone[9]:

$$F \stackrel{\text{def}}{=} \{\boldsymbol{h} \in \mathbb{R}^{2^{[n]}} \mid \forall \sigma \in \Sigma_0 : h(\sigma) = 0\}$$

Let $L_0 \stackrel{\text{def}}{=} L - \{\hat{0}\}$, $N \stackrel{\text{def}}{=} |L_0|$. Our cone $K_0 \subseteq \mathbb{R}^N_+$ is:

$$K_0 \stackrel{\text{def}}{=} \Pi_{L_0}(K \cap F) \tag{68}$$

The function Π_{L_0} projects a $2^{[n]}$-dimensional vector $(h_{\boldsymbol{U}})_{\boldsymbol{U} \in 2^{\boldsymbol{X}}}$ to the N-dimensional vector $(h_{\boldsymbol{U}})_{\boldsymbol{U} \in L_0}$. Thus, K_0 not only removes the $\emptyset$ dimension, but also removes all dimensions subject to a tight constraint. We prove the following:

(1) K_0 is proper

(2) Log-L-Bound$_K$ = Log-L-Bound$_{K_0}$ and
Log-U-Bound$_K$ = Log-U-Bound$_{K_0}$

(3) Log-L-Bound$_{K_0}$ = Log-U-Bound$_{K_0}$

Theorem 9.7 follows from these three claims.

We start with item (1), and observe that K_0 is a closed, convex cone, because $K \cap F$ is a closed, convex cone, and Π_{L_0} is a linear isomorphism $K \cap F \to K_0$: indeed, Π_{L_0} is surjective by the definition of K_0, and it is injective because, if $\boldsymbol{h}, \boldsymbol{h}' \in K \cap F$, then $\Pi_{L_0}(\boldsymbol{h}) = \Pi_{L_0}(\boldsymbol{h}')$ implies that, for every set $\boldsymbol{U}$, $h(\boldsymbol{U}) = h(\boldsymbol{U}^+) = h'(\boldsymbol{U}^+) = h'(\boldsymbol{U})$. It is immediate to check that K_0 is pointed, and we will show below that K_0 has a non-empty interior: this implies that it is proper.

Next, we prove item (2), and for that we will write $\boldsymbol{h}$ for a vector in $K \cap F$ and write $\boldsymbol{h}^{(0)}$ for a vector in K_0. For any statistics $\sigma = (\boldsymbol{V}|\boldsymbol{U}) \in \Sigma_1$, denote by $\sigma^+ = ((\boldsymbol{U}\boldsymbol{V})^+|\boldsymbol{U}^+)$. We say that a vector $\boldsymbol{h}^{(0)}$ *satisfies* the statistics $(\Sigma_1, \boldsymbol{b}_1)$, in notation $\boldsymbol{h}^{(0)} \models (\Sigma_1, \boldsymbol{b}_1)$, if $\boldsymbol{h}^{(0)}(\sigma^+) \leq \boldsymbol{b}_\sigma$ for all $\sigma \in \Sigma_1$. By definition, a vector $\boldsymbol{h}^{(0)} \in K_0$ satisfies the FDs Σ_0.

Lemma 9.10. *The following holds:*

$$\textit{Log-L-Bound}_K(Q, \Sigma, \boldsymbol{b}) = \textit{Log-L-Bound}_{K_0}(Q, \Sigma_1, \boldsymbol{b}_1)$$

Equivalently, the lemma states:

$$\begin{aligned}
&\sup_{\boldsymbol{h}} \{h(\boldsymbol{X}) \mid \boldsymbol{h} \in K, \boldsymbol{h} \models (\Sigma, \boldsymbol{b})\} \\
&\quad = \sup_{\boldsymbol{h}^{(0)}} \left\{h^{(0)}(\hat{1}) \mid \boldsymbol{h}^{(0)} \in K_0, \boldsymbol{h}^{(0)} \models (\Sigma_1, \boldsymbol{b}_1)\right\}
\end{aligned}$$

[9] In fact F is even a vector space.

and the proof is immediate, because the projection of a vector $\boldsymbol{h} \in K$ satisfying $(\Sigma, \boldsymbol{b})$ is a vector $\boldsymbol{h}^{(0)} \in K_0$ satisfying $(\Sigma_1, \boldsymbol{b}_1)$, and, conversely, every such $\boldsymbol{h}^{(0)}$ is the projection of a vector $\boldsymbol{h}$.

Recall that we have assumed $N_n \subseteq K \subseteq \Gamma_n$:

Lemma 9.11. *The following holds:*

$$\textit{Log-U-Bound}_K(Q, \Sigma, \boldsymbol{b}) = \textit{Log-U-Bound}_{K_0}(Q, \Sigma_1, \boldsymbol{b}_1)$$

Proof. We need to prove:

$$\begin{aligned}
&\inf_{\boldsymbol{w}} \left\{ \sum_{\sigma \in \Sigma} w_\sigma b_\sigma \mid K \models \sum_{\sigma \in \Sigma} w_\sigma h(\sigma) \geq h(\boldsymbol{X}) \right\} \\
&= \inf_{\boldsymbol{w}^{(0)}} \left\{ \sum_{\sigma \in \Sigma_1} w^{(0)}_{\sigma^+} b_\sigma \mid K_0 \models \sum_{\sigma \in \Sigma_1} w^{(0)}_{\sigma^+} h^{(0)}(\sigma^+) \geq h^{(0)}(\boldsymbol{X}) \right\}
\end{aligned} \tag{69}$$

A vector $\boldsymbol{w}$ on the LHS defines an unconstrained inequality, while a vector $\boldsymbol{w}^{(0)}$ on the RHS defines a constrained inequality, because $\boldsymbol{w}^{(0)}$ satisfies

$$\forall \boldsymbol{h}^{(0)} \in K_0, \sum_{\sigma \in \Sigma_1} w^{(0)}_{\sigma^+} h^{(0)}(\sigma^+) \geq h^{(0)}(\boldsymbol{X})$$

iff it satisfies

$$\forall \boldsymbol{h} \in K, \bigwedge_{\sigma \in \Sigma_0} h(\sigma) = 0 \Rightarrow \sum_{\sigma \in \Sigma_1} w^{(0)}_{\sigma^+} h(\sigma^+) \geq h(\boldsymbol{X}) \tag{70}$$

which is a constrained inequality.

We start by showing that LHS$\geq$RHS in Eq. (69). For that it suffices observe that, if $\boldsymbol{w} = (\boldsymbol{w}_\sigma)_{\sigma \in \Sigma}$ defines a valid inequality $\sum_{\sigma \in \Sigma} w_\sigma h(\sigma) \geq h(\boldsymbol{X})$, then its projection $\boldsymbol{w}^{(0)} = (w_{\sigma^+})_{\sigma \in \Sigma_1}$ defines a valid constrained inequality $\sum_{\sigma \in \Sigma_1} w^{(0)}_{\sigma^+} h^{(0)}(\sigma^+) \geq h^{(0)}(\boldsymbol{X})$, because all the missing terms $h(\sigma)$ for $\sigma \in \Sigma_0$ are $= 0$. Therefore $\inf_{\boldsymbol{w}}(\cdots) \geq \inf_{\boldsymbol{w}^{(0)}}(\cdots)$

We prove now that LHS$\leq$RHS. Let $\boldsymbol{w}^{(0)}$ be a vector defining a valid constrained inequality (70). The objective value of the RHS of (69) is $\sum_{\sigma \in \Sigma_1} w^{(0)}_{\sigma^+} b_\sigma$. By the relaxation theorem 9.5, for every $\varepsilon > 0$ there exists $\lambda_\sigma \geq 0$, for $\sigma \in \Sigma_0$, such that the following is a valid, unconstrained inequality:

$$K \models \sum_{\sigma \in \Sigma_0} \lambda_\sigma h(\sigma) + \sum_{\sigma \in \Sigma_1} w^{(0)}_{\sigma^+} h(\sigma^+) + \varepsilon h(\boldsymbol{X}) \geq h(\boldsymbol{X})$$

This inequality is not yet of the form on the LHS of (69), because we have terms $h(\sigma^+)$ instead of $h(\sigma)$. For each such term, $h(\sigma^+) = h((\boldsymbol{U}\boldsymbol{V})^+ | \boldsymbol{U}^+)$, where $\sigma = (\boldsymbol{V}|\boldsymbol{U})$, we use the following Shannon inequality:

$$\begin{aligned}
h(\sigma^+) &= h((\boldsymbol{U}\boldsymbol{V})^+ | \boldsymbol{U}^+) \leq h((\boldsymbol{U}\boldsymbol{V})^+ | \boldsymbol{U}) \\
&= h(\boldsymbol{U}\boldsymbol{V} | \boldsymbol{U}) + h((\boldsymbol{U}\boldsymbol{V})^+ | \boldsymbol{U}\boldsymbol{V}) \\
&= h(\sigma) + h((\boldsymbol{U}\boldsymbol{V})^+ | \boldsymbol{U}\boldsymbol{V}) \leq h(\sigma) + \sum_{\sigma \in \Sigma_0} h(\sigma)
\end{aligned}$$

The last inequality, $h((\boldsymbol{U}\boldsymbol{V})^+ | \boldsymbol{U}\boldsymbol{V}) \leq \sum_{\sigma \in \Sigma_0} h(\sigma)$, can be checked by induction on the number of steps needed to compute the closure $(\boldsymbol{U}\boldsymbol{V})^+$ using the FDs in Σ_0. (It also follows from Theorem 9.2.)

This implies that there exists coefficients $\lambda'_\sigma \geq 0$ such that the following is a valid, unconstrained inequality:

$$K \models \sum_{\sigma \in \Sigma_0} \lambda'_\sigma h(\sigma) + \sum_{\sigma \in \Sigma_1} w^{(0)}_{\sigma^+} h(\sigma) + \varepsilon h(\boldsymbol{X}) \geq h(\boldsymbol{X})$$

or, equivalently,

$$K \models \sum_{\sigma \in \Sigma_0} \frac{\lambda'_\sigma}{1-\varepsilon} h(\sigma) + \sum_{\sigma \in \Sigma_1} \frac{w^{(0)}_{\sigma^+}}{1-\varepsilon} h(\sigma) \geq h(\boldsymbol{X})$$

This is a valid inequality for the LHS of (69), and its objective value is $(\sum_{\sigma \in \Sigma_1} w^{(0)}_{\sigma^+} b_\sigma)/(1-\varepsilon)$, because $b_\sigma = 0$ for all $\sigma \in \Sigma_0$. Since ε can be chosen arbitrarily small, it follows that LHS$\leq$RHS in (69). $\square$

This completes the proof of item (2). It remains to prove item (3). For that we represent both Log-L-Bound$_{K_0}(Q, \Sigma_1, \boldsymbol{b}_1)$ and Log-U-Bound$_{K_0}(Q, \Sigma_1, \boldsymbol{b}_1)$ as the solutions to the primal/dual cone program (67) over the cone K_0. Notice that the vector $\boldsymbol{b}$ is restricted to $\boldsymbol{b}_1$ and, therefore, $b_\sigma > 0$ for all $\sigma \in \Sigma_1$: we no longer have tight constraints. Similarly, the matrix $\boldsymbol{A}$ will be restricted to a matrix $\boldsymbol{A}_1$ whose rows correspond to the closed sets $\boldsymbol{U} \subseteq \boldsymbol{X}$. It remains to check Slater's condition, and, in particular, prove that K_0 has a non-empty interior. For that purpose we extended the definition of step functions from Sec. 6.1 to our lattice L. For each closed set $\boldsymbol{W} \in L$, s.t. $\boldsymbol{W} \neq \hat{1}$, we define the step function at $\boldsymbol{W}$ as follows.

$$\forall \boldsymbol{U} \in L: \quad h^{(0)}_{\boldsymbol{W}}(\boldsymbol{U}) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } \boldsymbol{U} \subseteq \boldsymbol{W} \\ 1 & \text{otherwise} \end{cases}$$

Let $\boldsymbol{h_W} \in K$ be the standard step function in Eq. (40) (we assumed $N_n \subseteq K$). $\boldsymbol{h_W}$ satisfies all FDs Σ_0, because the only FDs that it does not satisfy are of the form $\boldsymbol{U} \rightarrow \boldsymbol{V}$ where $\boldsymbol{U} \subseteq \boldsymbol{W}$, $\boldsymbol{V} \not\subseteq \boldsymbol{W}$, and none of the FDs in Σ_0 have this form because $\boldsymbol{W}^+ = \boldsymbol{W}$. It follows $\boldsymbol{h_W} \in K \cap F$, and this proves $\boldsymbol{h}^{(0)}_{\boldsymbol{W}} = \Pi_{L_0}(\boldsymbol{h_W}) \in K_0$. There are N step functions $\boldsymbol{h}^{(0)}_{\boldsymbol{W}} \in K_0$, and it is straightforward to check that they are independent vectors in $\mathbb{R}^N$. Let $\varepsilon > 0$ be small enough such that $2\varepsilon N < \min_{\sigma \in \Sigma_1} b_\sigma$. Define $\boldsymbol{h} \stackrel{\text{def}}{=} \sum_{\boldsymbol{W} \in L_0} \varepsilon \boldsymbol{h_W}$, and $\boldsymbol{h}^{(0)} \stackrel{\text{def}}{=} \Pi_{L_0}(\boldsymbol{h})$. Since $K \cap F$ is a convex cone, $\boldsymbol{h} \in K \cap F$ and therefore $\boldsymbol{h}^{(0)} \in K_0$. We claim that there exists slack variables $\boldsymbol{\beta}$ such that $(\boldsymbol{h}^{(0)}, \beta)$ is a feasible solution to the cone program in (67) and, furthermore, $(\boldsymbol{h}^{(0)}, \boldsymbol{\beta})$ belongs to the interior of $K_0 \times \mathbb{R}^s_+$. Indeed, for all $\sigma \in \Sigma_1$, $h^{(0)}(\sigma) \leq \varepsilon N < b_\sigma$, hence, if we define $\beta_\sigma \stackrel{\text{def}}{=} b_\sigma - h^{(0)}(\sigma)$, the pair $(\boldsymbol{h}^{(0)}, \boldsymbol{\beta}) \in K_0 \times \mathbb{R}^s_+$ is a feasible solution to the primal (67). Next, we prove that $(\boldsymbol{h}^{(0)}, \boldsymbol{\beta})$ is in the interior of $K_0 \times \mathbb{R}^s_+$. Since $\beta_\sigma > 0$ for all $\sigma \in \Sigma_1$, it follows that $\boldsymbol{\beta}$ is in the interior of $\mathbb{R}^s_+$. Set $\boldsymbol{h}' \stackrel{\text{def}}{=} \sum_{\boldsymbol{W} \in L_0} (\varepsilon + \delta_{\boldsymbol{W}}) \boldsymbol{h_W}$, where $\delta_{\boldsymbol{W}} \in (-\varepsilon, \varepsilon)$ are N arbitrary numbers, we have $\boldsymbol{h}' \in K \cap F$, hence $\Pi_{L_0}(\boldsymbol{h}') \in K_0$. This proves that $\boldsymbol{h}^{(0)}$ is in the interior of K_0, thus, verifying Slater's condition. It follows that Log-L-Bound$_{K_0}(Q, \Sigma_1, \boldsymbol{b}_1)$ = Log-U-Bound$_{K_0}(Q, \Sigma_1, \boldsymbol{b}_1)$, completing the proof of item (3). $\square$

10 Conclusions

Data is ultimately information, and therefore the connection between databases and information theory is no surprise. We have discussed applications of information inequalities to several database theory problems: query upper bounds, query evaluation, query domination, and reasoning about approximate constraints. There are major open problems in information theory, for example the decidability of entropic information inequalities, the complexity of deciding Shannon inequalities, a characterization of the cone $\bar{\Gamma}_n^*$, and each such open problem has a corresponding open problem in database theory. In some cases the converse holds too, for example the query domination problem is computationally equivalent to checking validity of max-information inequalities, hence any proof of (un)-decidability of one problem carries over to the other.

A broader question is whether information theory can find wider applications in finite model theory. For example, functional dependencies and multivalued dependencies can be specified either using first order logic sentences, or using entropic terms. Are there other properties in finite model theory that can be captured using information theory? Such a connection would enable logical implications to be relaxed to approximate reasoning, with lots of potential in modern, data-driven applications that rely heavily on statistical reasoning.

References

[ACK+11] Michael Armbrust, Kristal Curtis, Tim Kraska, Armando Fox, Michael J. Franklin, and David A. Patterson. PIQL: success-tolerant query processing in the cloud. *PVLDB*, 5(3):181–192, 2011.

[AD80] William Ward Armstrong and Claude Delobel. Decomposition and functional dependencies in relations. *ACM Trans. Database Syst.*, 5(4):404–430, 1980.

[AFP+09] Michael Armbrust, Armando Fox, David A. Patterson, Nick Lanham, Beth Trushkowsky, Jesse Trutna, and Haruki Oh. SCADS: scale-independent storage for social computing applications. In *CIDR 2009, Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2009, Online Proceedings*, 2009.

[AGM13] Albert Atserias, Martin Grohe, and Dániel Marx. Size bounds and query plans for relational joins. *SIAM J. Comput.*, 42(4):1737–1767, 2013.

[AL02] Marcelo Arenas and Leonid Libkin. A normal form for XML documents. In Lucian Popa, Serge Abiteboul, and Phokion G. Kolaitis, editors, *Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA*, pages 85–96. ACM, 2002.

[AL05] Marcelo Arenas and Leonid Libkin. An information-theoretic approach to normal forms for relational and XML data. *J. ACM*, 52(2):246–283, 2005.

[ALK+13] Michael Armbrust, Eric Liang, Tim Kraska, Armando Fox, Michael J. Franklin, and David A. Patterson. Generalized scale independence through incremental precomputation. In *SIGMOD 2013*, pages 625–636, 2013.

[AtCG+15] Molham Aref, Balder ten Cate, Todd J. Green, Benny Kimelfeld, Dan Olteanu, Emir Pasalic, Todd L. Veldhuizen, and Geoffrey Washburn. Design and implementation of the logicblox system. In Timos K. Sellis, Susan B. Davidson, and Zachary G. Ives, editors, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 1371–1382. ACM, 2015.

[BB12] Paul Balister and Béla Bollobás. Projections, entropy and sumsets. *Comb.*, 32(2):125–141, 2012.

[BBF+16] Tobias Bleifuß, Susanne Bülow, Johannes Frohnhofen, Julian Risch, Georg Wiese, Sebastian Kruse, Thorsten Papenbrock, and Felix Naumann. Approximate discovery of functional dependencies for large datasets. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 1803–1812, 2016.

[BFH77] Catriel Beeri, Ronald Fagin, and John 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.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CGFS86] Fan R. K. Chung, Ronald L. Graham, Peter Frankl, and James B. Shearer. Some intersection theorems for ordered sets and graphs. *J. Comb. Theory, Ser. A*, 43(1):23–37, 1986.

[CIPY14] Xu Chu, Ihab F. Ilyas, Paolo Papotti, and Yin Ye. Ruleminer: Data quality rules discovery. In *IEEE 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, March 31 - April 4, 2014*, pages 1222–1225, 2014.

[Csi22] Laszlo Csirmaz. Around entropy inequalities. `https://www.lirmm.fr/~romashchen/seminar-aait.html`, 2022. Slides and video accessed March 2023.

[CV93] Surajit Chaudhuri and Moshe Y. Vardi. Optimization of *Real* conjunctive queries. In *ACM PODS, 1993*, pages 59–70, 1993.

[CY02] Terence H. Chan and Raymond W. Yeung. On a relation between information inequalities and group theory. *IEEE Transactions on Information Theory*, 48(7):1992–1995, 2002.

[Fag82] Ronald Fagin. Horn clauses and database dependencies. *J. ACM*, 29(4):952–985, 1982.

[Fag83] Ronald Fagin. Degrees of acyclicity for hypergraphs and relational database schemes. *J. ACM*, 30(3):514–550, 1983.

[FBS+20] Michael J. Freitag, Maximilian Bandle, Tobias Schmidt, Alfons Kemper, and Thomas Neumann. Adopting worst-case optimal joins in relational database systems. *Proc. VLDB Endow.*, 13(11):1891–1904, 2020.

[FD16] Abram Friesen and Pedro Domingos. The sum-product theorem: A foundation for learning tractable models. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1909–1918, New York, New York, USA, 20–22 Jun 2016. PMLR.

[FK98] Ehud Friedgut and Jeff Kahn. On the number of copies of one hypergraph in another. *Israel Journal of Mathematics*, 105(1):251–256, 1998.

[Fri04] Ehud Friedgut. Hypergraphs, entropy, and inequalities. *Am. Math. Mon.*, 111(9):749–760, 2004.

[GLVV12] Georg Gottlob, Stephanie Tien Lee, Gregory Valiant, and Paul Valiant. Size and treewidth bounds for conjunctive queries. *J. ACM*, 59(3):16:1–16:35, 2012.

[GM06] Martin Grohe and Dániel Marx. Constraint solving via fractional edge covers. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006, Miami, Florida, USA, January 22-26, 2006*, pages 289–298. ACM Press, 2006.

[GM14] Martin Grohe and Dániel Marx. Constraint solving via fractional edge covers. *ACM Trans. Algorithms*, 11(1):4:1–4:20, 2014.

[GP93] Dan Geiger and Judea Pearl. Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics*, 21(4):2001–2021, 1993.

[GR04] Chris Giannella and Edward L. Robertson. On approximation measures for functional dependencies. *Inf. Syst.*, 29(6):483–507, 2004.

[GT17] Tomasz Gogacz and Szymon Torunczyk. Entropy bounds for conjunctive queries with functional dependencies. In Michael Benedikt and Giorgio Orsi, editors, *20th International Conference on Database Theory, ICDT 2017, March 21-24, 2017, Venice, Italy*, volume 68 of *LIPIcs*, pages 15:1–15:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[Her06] Christian Herrmann. Corrigendum to "on the undecidability of implications between embedded multivalued database dependencies" [inform. and comput. 122(1995) 221-235]. *Inf. Comput.*, 204(12):1847–1851, 2006.

[IR95] Yannis E. Ioannidis and Raghu Ramakrishnan. Containment of conjunctive queries: Beyond relations as sets. *ACM Trans. Database Syst.*, 20(3):288–324, 1995.

[JKV06] T. S. Jayram, Phokion G. Kolaitis, and Erik Vee. The containment problem for REAL conjunctive queries with inequalities. In *ACM PODS, 2006*, pages 80–89, 2006.

[KKNS20] Mahmoud Abo Khamis, Phokion G. Kolaitis, Hung Q. Ngo, and Dan Suciu. Decision problems in information theory. In Artur Czumaj, Anuj Dawar, and Emanuela Merelli, editors, *47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 168 of *LIPIcs*, pages 106:1–106:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[KKNS21] Mahmoud Abo Khamis, Phokion G. Kolaitis, Hung Q. Ngo, and Dan Suciu. Bag query containment and information theory. *ACM Trans. Database Syst.*, 46(3):12:1–12:39, 2021.

[KMP$^+$20] Batya Kenig, Pranay Mundra, Guna Prasaad, Babak Salimi, and Dan Suciu. Mining approximate acyclic schemes from relations. In David Maier, Rachel Pottinger, AnHai Doan, Wang-Chiew Tan, Abdussalam Alawini, and Hung 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.

[KN18] Sebastian Kruse and Felix Naumann. Efficient discovery of approximate dependencies. *PVLDB*, 11(7):759–772, 2018.

[KNS16] Mahmoud Abo Khamis, Hung Q. Ngo, and Dan Suciu. Computing join queries with functional dependencies. In Tova Milo and Wang-Chiew Tan, editors, *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 327–342. ACM, 2016.

[KNS17] Mahmoud Abo Khamis, Hung Q. Ngo, and Dan Suciu. What do shannon-type inequalities, submodular width, and disjunctive datalog have to do with one another? In Emanuel Sallinger, Jan Van den Bussche, and Floris Geerts, editors, *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017*, pages 429–444. ACM, 2017. Extended version available at `http://arxiv.org/abs/1612.02503`.

[KR11] Swastik Kopparty and Benjamin Rossman. The homomorphism domination exponent. *Eur. J. Comb.*, 32(7):1097–1114, 2011.

[KR13] Tarik Kaced and Andrei E. Romashchenko. Conditional information inequalities for entropic and almost entropic points. *IEEE Trans. Inf. Theory*, 59(11):7149–7167, 2013.

[KS22] Batya Kenig and Dan Suciu. Integrity constraints revisited: From exact to approximate implication. *Log. Methods Comput. Sci.*, 18(1), 2022.

[Lee87a] Tony 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.

[Lee87b] Tony T. Lee. An information-theoretic analysis of relational databases - part II: information structures of database schemas. *IEEE Trans. Software Eng.*, 13(10):1061–1072, 1987.

[LW49] L. H. Loomis and H. Whitney. An inequality related to the isoperimetric inequality. *Bull. Am. Math. Soc.*, 55:961–962, 1949.

[Mar13] Dániel Marx. Tractable hypergraph properties for constraint satisfaction and conjunctive queries. *J. ACM*, 60(6):42:1–42:51, 2013.

[Mat07] Frantisek Matús. Infinitely many information inequalities. In *IEEE International Symposium on Information Theory, ISIT 2007, Nice, France, June 24-29, 2007*, pages 41–44. IEEE, 2007.

[MKS21] Amine Mhedhbi, Chathura Kankanamge, and Semih Salihoglu. Optimizing one-time and continuous subgraph queries using worst-case optimal joins. *ACM Trans. Database Syst.*, 46(2):6:1–6:45, 2021.

[NPRR12] Hung Q. Ngo, Ely Porat, Christopher Ré, and Atri Rudra. Worst-case optimal join algorithms: [extended abstract]. In Michael Benedikt, Markus Krötzsch, and Maurizio Lenzerini, editors, *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 37–48. ACM, 2012.

[NPRR18] Hung Q. Ngo, Ely Porat, Christopher Ré, and Atri Rudra. Worst-case optimal join algorithms. *J. ACM*, 65(3):16:1–16:40, 2018.

[NRR13] Hung Q. Ngo, Christopher Ré, and Atri Rudra. Skew strikes back: new developments in the theory of join algorithms. *SIGMOD Rec.*, 42(4):5–16, 2013.

[PD11] Hoifung Poon and Pedro M. Domingos. Sum-product networks: A new deep architecture. In *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pages 337–346, 2011.

[Pip86] Nicholas Pippenger. What are the laws of information theory. In *1986 Special Problems on Communication and Computation Conference*, pages 3–5, 1986.

[Rom22] Andrei E. Romashchenko. The zhang-yeung inequality: an attempt to explain. `https://www.lirmm.fr/~romashchen/seminar-aait.html`, 2022. Slides and video accessed March 2023.

[SBHR06] Yannis Sismanis, Paul Brown, Peter J. Haas, and Berthold Reinwald. GORDIAN: efficient and scalable discovery of composite keys. In *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006*, pages 691–702, 2006.

[Sch03] A. Schrijver. *Combinatorial Optimization - Polyhedra and Efficiency.* Springer, 2003.

[SGS18] Babak Salimi, Johannes Gehrke, and Dan Suciu. Bias in OLAP queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 1021–1035, 2018.

[SOC16] Maximilian Schleich, Dan Olteanu, and Radu Ciucanu. Learning linear regression models over factorized joins. In Fatma Özcan, Georgia Koutrika, and Sam Madden, editors, *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 3–18. ACM, 2016.

[Stu90] Milan 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.

[Vel14] Todd L. Veldhuizen. Triejoin: A simple, worst-case optimal join algorithm. In Nicole Schweikardt, Vassilis Christophides, and Vincent Leroy, editors, *Proc. 17th International Conference on Database Theory (ICDT), Athens, Greece, March 24-28, 2014*, pages 96–106. OpenProceedings.org, 2014.

[Wan22] Yilei Wang. Personal communication, 2022.

[WWS23] Yisu Remy Wang, Max Willsey, and Dan Suciu. Free join: Unifying worst-case optimal and traditional joins. *CoRR*, abs/2301.10841, 2023.

[Yeu08] Raymond W. Yeung. *Information Theory and Network Coding.* Springer Publishing Company, Incorporated, 1 edition, 2008.

[ZY97] Zhen Zhang and Raymond W. Yeung. A non-shannon-type conditional inequality of information quantities. *IEEE Trans. Information Theory*, 43(6):1982–1986, 1997.

[ZY98] Zhen Zhang and Raymond W Yeung. On characterization of entropy function via information inequalities. *IEEE Transactions on Information Theory*, 44(4):1440–1452, 1998.

.1 Proof of the claim in Example 4.9

We prove that any valid Σ-information inequality is a positive linear combination of the four inequalities shown in Example 4.9. Such an inequality has the following form:

$$\begin{aligned} w_R h(XY) + w_S h(YZ) + w_T h(ZU) \\ + w_B h(U|XZ) + w_A h(X|YU) \geq & h(XYZU) \end{aligned} \tag{71}$$

where $w_R, \ldots, w_A \geq 0$ are non-negative real numbers. Since the inequality holds for all polymatroids, it also holds for every step function $\boldsymbol{h}^{\boldsymbol{V}}$ (see Eq. (42)), for all $\boldsymbol{V} \subseteq \{X, Y, Z, U\}$. There are $2^4 - 1 = 15$ step functions, but we only use 5 of them:

$$\begin{aligned} \boldsymbol{h}^X : \quad & w_R + w_B \geq 1 \\ \boldsymbol{h}^Y : \quad & w_R + w_S \geq 1 \\ \boldsymbol{h}^Z : \quad & w_S + w_T \geq 1 \\ \boldsymbol{h}^U : \quad & w_T + w_A \geq 1 \\ \boldsymbol{h}^{XU} : \quad & w_R + w_T \geq 1 \end{aligned}$$

Consider the three constraints for $\boldsymbol{h}^Y, \boldsymbol{h}^Z, \boldsymbol{h}^{XU}$, which mention only the variables w_R, w_S, w_T. Any solution to these three constraints can be immediately extended to a solution to all 5 constraints, by setting $w_B \geq \max(0, 1 - w_R)$ and $w_A \geq \max(0, 1 - w_T)$. On the other hand, the three constraints on w_R, w_S, w_T assert that they form a fractional edge cover of a triangle. The fractional edge covering polytope of a triangle has four extreme vertices, $(0, 1, 1), (1, 0, 1), (1, 1, 0), (1/2, 1/2, 1/2)$. It follows that the extreme vertices of our polytope over all 5 variables are:

w_R	w_S	w_T	w_A	w_B
0	1	1	0	1
1	0	1	0	0
1	1	0	1	0
1/2	1/2	1/2	1/2	1/2

Each of these vectors corresponds precisely to one of the four inequalities that we listed in Example 4.9, and, conversely, any Σ-inequality of the form (71) is dominated by some convex combination of one of these four.

.2 Proof of (33), and Failure of the Copy Lemma

We start by proving Inequality (33). The proof follows immediately from the following identity:

$$\begin{aligned}
&-I(X;Y)+I(X;Y|A)+I(X;Y|B)+I(A:B)+\\
&\quad+I(X;Y|A')+I(A';Y|X)+I(A';X|Y)+3I(A';AB|XY)\\
=&I(A;B|A')+I(A;A'|Y)+I(A;A'|X)\\
&\quad+I(A;A'|BXY)+I(B;A'|Y)+I(B;A'|X)\\
&\quad+I(B;A'|AXY)+I(X;Y|BA')+I(X;Y|AA')\\
&\quad+I(X;A'|ABY)+I(Y;A'|AB)
\end{aligned}$$

Next, we prove that the polymatroid in Fig. 5 does not satisfy the Copy Lemma. Assuming otherwise, let h' denote the polymatroid over variables X, Y, A, B, A', B'. Using only basic Shannon inequalities, we derive a contradiction. We will drop the index h' from $I_{h'}(\cdots)$ and write simply $I(\cdots)$. By assumption, the values $h'(\boldsymbol{U})$ for all sets $\boldsymbol{U}$ that do not contain both A and A', or both B and A' are known, for example $h'(A'|XY) = h(A|XY) = 1$. We also known the value $h'(\boldsymbol{U})$ when $\boldsymbol{U}$ contain XY, example $h'(AA'XY) = h'(AA'|XY)+h'(XY) = h'(A|XY)+h'(A'|XY)+h(XY) = 1+1+3 = 5$. We proceed by examining the other sets where A, A' or B, B' co-occur, and start by showing $h'(AA') \leq 3$:

$$\begin{aligned}
I(A;A'|X) =&h'(AX)+h'(A'X)-h'(X)-h'(AA'X)\\
=&3+3-2-h'(AA'X) \geq 0
\end{aligned}$$

and we derive $h'(AA'X) \leq 4$. Similarly (replacing X with Y) we derive $h'(AA'Y) \leq 4$. Finally, we have:

$$\begin{aligned}
&I(X;Y|AA') =\\
&\quad=h'(AA'X)+h'(AA'Y)-h'(AA'XY)-h'(AA') \geq 0
\end{aligned}$$

and we derive that $h'(AA') \leq 3$. We repeat the argument above by replacing A with B, and derive similarly that $h'(BA') \leq 3$. Next, we show that $h'(ABA') \geq 5$, which follows from:

$$\begin{aligned}
&I(XY;A'|AB) =\\
&\qquad=h'(ABXY)+h'(ABA')-h'(AB)-h'(ABA'XY)\\
&\qquad=4+h'(ABA')-4-5 \geq 0
\end{aligned}$$

thus $h'(ABA') \geq 5$. (We also have $h'(ABA') \leq h'(ABA'XY) = 5$, hence $h'(ABA') = 5$, but the inequality suffices for us.) Finally, we derive a contradiction:

$$\begin{aligned}
I(A;B|A') =&h'(AA')+h'(BA')-h'(A')-h'(ABA')\\
\leq&3+3-2-5=-1
\end{aligned}$$

.3 Addendum to Theorem 5.9

We briefly sketch here the proof that, if an entropic function $\boldsymbol{h}$ satisfies a set of functional dependencies, then so do all $\boldsymbol{h}^{(r)}$, for all $r \geq 0$. For that we need to review the main argument of the proof in [CY02].

Let $\boldsymbol{h}$ be an entropic function, realized by a probability distribution (R, p). The first step is to ensure that the probabilities $p(t)$ can be assumed to be rational numbers. Assume w.l.o.g. that R is the support of p, then, by Lee's result [Lee87a], $h \models \boldsymbol{U} \rightarrow \boldsymbol{V}$, iff $R \models \boldsymbol{U} \rightarrow \boldsymbol{V}$. Consider now any sequence of probability distributions on R, $p^{(k)} : R \rightarrow [0, 1]$, of rational numbers, such that $\lim_k p^{(k)} = p$. Then $p^{(k)}$, and its entropic vector $\boldsymbol{h}^{(k)}$, continue to satisfy the same FDs as R and, thus, the same FDs as $\boldsymbol{h}$. Since $\boldsymbol{h}^{(k)}$ can be arbitrarily close to $\boldsymbol{h}$, it suffices to prove that the theorem holds for an entropic vector $\boldsymbol{h}$ realized by a probability distribution (R, p) where the probabilities are rational numbers. Assume they have a common denominator $q > 0$, and let $N = |R|$.

From here on, we follow Chan and Yeung's proof [CY02]. For each $r = q, 2q, 3q, \ldots$ define the following $r \times n$ matrix $\boldsymbol{M}_r = (m_{\rho i})_{\rho=1,r; i=1,n}$. Its rows are copies of the tuples in R, where each tuple $\boldsymbol{x} \in R$ occurs $r \cdot p(\boldsymbol{x})$ times in the matrix $\boldsymbol{M}_r$. Intuitively, $\boldsymbol{M}_r$ can be viewed as a relation with n attributes and r tuples, including duplicates, whose uniform probability distribution has the same entropic vector h as (R, p). Let G be the symmetric group S_r, i.e. the group of permutation on the set $\{1, 2, \ldots, r\}$; one should think of G as the group of permutations on the rows of $\boldsymbol{M}_r$. For each $i = 1, \ldots, n$, let G_i the subgroup that leaves the column i invariant, in other words:

$$G_i = \{\sigma \in G \mid m_{\sigma(\rho),i} = m_{\rho,i}, \forall \rho = 1, r\}$$

Denoting similarly G_α the subgroup of permutations that leave the set of columns $\alpha \subseteq [n]$ invariant, one can check that $G_\alpha = \bigcap_{i \in \alpha} G_i$. Let $\boldsymbol{h}^{(r)}$ be the entropy of the uniform probability distribution on the relational instance $\{(aG_1, \ldots, aG_n) \mid a \in G\}$. Using a combinatorial argument, Chan and Yeung [CY02] prove that $\lim_{r\rightarrow\infty} \frac{1}{r}\boldsymbol{h}^{(r)} = \boldsymbol{h}$. We will not repeat that argument here, but make the additional observation that, if R satisfies the FD $\boldsymbol{U} \rightarrow \boldsymbol{V}$, then $G_{\boldsymbol{U}} \subseteq G_{\boldsymbol{V}}$, which implies that $\boldsymbol{h}^{(r)}$ also satisfies the same FD.

.4 Proof of Equation (44)

Möbius inversion formula states that, if $f, g : 2^{\boldsymbol{X}} \rightarrow \mathbb{R}$ are two set functions, and one of the identities below holds, then so does the other:

$$f(\boldsymbol{U}) = \sum_{\boldsymbol{V} \subseteq \boldsymbol{U}} g(\boldsymbol{V}) \qquad\qquad g(\boldsymbol{U}) = \sum_{\boldsymbol{V} \subseteq \boldsymbol{U}} (-1)^{|\boldsymbol{U}-\boldsymbol{V}|} f(\boldsymbol{V}) \tag{72}$$

It is immediate to derive that the following identities are also equivalent:[10]

$$f(\boldsymbol{U}) = \sum_{\boldsymbol{V} \subseteq \boldsymbol{X}-\boldsymbol{U}} g(\boldsymbol{V}) \qquad\qquad g(\boldsymbol{U}) = \sum_{\boldsymbol{V} : \boldsymbol{X}-\boldsymbol{U} \subseteq \boldsymbol{V}} (-1)^{|\boldsymbol{V} \cap \boldsymbol{U}|} f(\boldsymbol{V})$$

To prove Equation (44), we use Equation (45), and the fact that $h(\boldsymbol{X}) = \sum_{\boldsymbol{V}:\boldsymbol{V}\subseteq\boldsymbol{X}} a_{\boldsymbol{V}}$:

$$\begin{aligned}
h(\boldsymbol{U}) &= \sum_{\boldsymbol{V}:\boldsymbol{V}\cap\boldsymbol{U}\neq\emptyset} a_{\boldsymbol{V}} = h(\boldsymbol{X}) - \sum_{\boldsymbol{V}:\boldsymbol{V}\subseteq\boldsymbol{X}-\boldsymbol{U}} a_{\boldsymbol{V}} \\
h(\boldsymbol{X}|\boldsymbol{U}) &= \sum_{\boldsymbol{V}:\boldsymbol{V}\subseteq\boldsymbol{X}-\boldsymbol{U}} a_{\boldsymbol{V}}
\end{aligned}$$

[10] Define $h(\boldsymbol{U}) \stackrel{\text{def}}{=} f(\boldsymbol{X} - \boldsymbol{U})$ then (72) becomes:

$$h(\boldsymbol{U}) = \sum_{\boldsymbol{V} \subseteq \boldsymbol{X}-\boldsymbol{U}} g(\boldsymbol{V}) \qquad\qquad g(\boldsymbol{U}) = \sum_{\boldsymbol{V} \subseteq \boldsymbol{U}} (-1)^{|\boldsymbol{U}-\boldsymbol{V}|} h(\boldsymbol{X} - \boldsymbol{V})$$

The claim follows by replacing $\boldsymbol{V}$ with $\boldsymbol{X} - \boldsymbol{V}$ in the second equation, then renaming h to f.

Möbius' inversion formula implies:

$$a_{\boldsymbol{U}} = \sum_{\boldsymbol{V}:\boldsymbol{X}-\boldsymbol{U}\subseteq \boldsymbol{V}} (-1)^{|\boldsymbol{V}\cap \boldsymbol{U}|} h(\boldsymbol{X}|\boldsymbol{V})$$

If $\boldsymbol{X} - \boldsymbol{U} \subseteq \boldsymbol{V}$, then we can write $\boldsymbol{V}$ uniquely as $(\boldsymbol{X} - \boldsymbol{U}) \cup \boldsymbol{V}_0$, where $\boldsymbol{V}_0 \subseteq \boldsymbol{U}$. After renaming $\boldsymbol{V}_0$ to $\boldsymbol{V}$ we derive:

$$\begin{aligned}
a_{\boldsymbol{U}} &= \sum_{\boldsymbol{V}\subseteq \boldsymbol{U}} (-1)^{|\boldsymbol{V}|} h(\boldsymbol{X}|(\boldsymbol{X}-\boldsymbol{U})\cup \boldsymbol{V}) \\
&= \sum_{\boldsymbol{V}\subseteq \boldsymbol{U}} (-1)^{|\boldsymbol{V}|} h(\boldsymbol{X}) - \sum_{\boldsymbol{V}\subseteq \boldsymbol{U}} (-1)^{|\boldsymbol{V}|} h((\boldsymbol{X}-\boldsymbol{U})\cup \boldsymbol{V}) \\
&= -\sum_{\boldsymbol{V}\subseteq \boldsymbol{U}} (-1)^{|\boldsymbol{V}|} h((\boldsymbol{X}-\boldsymbol{U})\cup \boldsymbol{V}) \\
&= -\sum_{\boldsymbol{V}\subseteq \boldsymbol{U}} (-1)^{|\boldsymbol{V}|} \left(h((\boldsymbol{X}-\boldsymbol{U})\cup \boldsymbol{V}) - h(\boldsymbol{X}-\boldsymbol{U})\right) \\
&= -\sum_{\boldsymbol{V}\subseteq \boldsymbol{U}} (-1)^{|\boldsymbol{V}|} h(\boldsymbol{V}|\boldsymbol{X}-\boldsymbol{U})
\end{aligned}$$

We used twice the fact that $\sum_{\boldsymbol{V}\subseteq \boldsymbol{U}} (-1)^{|\boldsymbol{V}|} = 0$ when $\boldsymbol{U} \neq \emptyset$. This completes the proof.

.5 Proof of Lemma 6.8

To prove the lemma we need to establish a simple fact:

Proposition .1. *Let $\alpha_1, \ldots, \alpha_n \geq 0$ be non-negative real numbers. Define the following set function $\boldsymbol{h} : 2^{\boldsymbol{X}} \to \mathbb{R}$:*

$$h(\boldsymbol{U}) \stackrel{def}{=} \max_{i:X_i\in \boldsymbol{U}} \alpha_i$$

Then $\boldsymbol{h}$ is a normal polymatroid.

Proof. Set $\alpha_0 \stackrel{\text{def}}{=} 0$ and assume w.l.o.g. that $\alpha_0 \leq \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$. Define $\delta_i \stackrel{\text{def}}{=} \alpha_i - \alpha_{i-1}$ for $i = 1, n$. We prove:

$$\boldsymbol{h} = \sum_{i=1,n} \delta_i \boldsymbol{h}^{\boldsymbol{X}_{[i+1:n]}} \tag{73}$$

If k is the largest index s.t. $X_k \in \boldsymbol{U}$, then:

$$\sum_{i=1,n} \delta_i \boldsymbol{h}^{\boldsymbol{X}_{[i+1:n]}}(\boldsymbol{U}) = \sum_{i=1,k} \delta_i = \alpha_k = h(\boldsymbol{U})$$

which proves (73). Since $\delta_i \geq 0, \forall i$, $\boldsymbol{h}$ is a normal polymatroid. □

Proof. (of Lemma 6.8) We prove the claim by induction on the number of variables $n = |\boldsymbol{X}|$. When $n = 0$ then the claim holds vacuously, so assume $n \geq 1$. We will use the following identity:

$$\forall \boldsymbol{U} \subseteq \boldsymbol{X} : \quad h(\boldsymbol{U}) = h(\boldsymbol{U}|X_n) + I(\boldsymbol{U}; X_n)$$

Consider the following two set functions $2^{\boldsymbol{X}-\{X_n\}} \to \mathbb{R}_+$:

$$h_1(\boldsymbol{U}) \stackrel{\text{def}}{=} h(\boldsymbol{U}|X_n) \qquad\qquad h_2(\boldsymbol{U}) \stackrel{\text{def}}{=} \max_{X_i \in \boldsymbol{U}} I(X_i; X_n)$$

Since $\boldsymbol{h}_1$, is a polymatroid in $n-1$ variables, by induction hypothesis we obtain a normal polymatroid $\boldsymbol{h}'_1$ satisfying properties (a),(b),(c). The second function is already a normal polymatroid, by Prop. .1. Observe that $h_2(\boldsymbol{U}) \leq I(\boldsymbol{U}; X_n)$ for any set $\boldsymbol{U}$ (see Eq. 31 in Prop. 5.6). Define:

$$a \stackrel{\text{def}}{=} \min_i (h(X_n) - I(X_i; X_n)) = h(X_n) - h_2(\boldsymbol{X})$$

and observe that:

$$0 \leq a \leq h(X_n) - I(\boldsymbol{U}; X_n), \quad \forall \boldsymbol{U} \subseteq \boldsymbol{X} - \{X_n\}$$

Since both $\boldsymbol{h}'_1$ and $\boldsymbol{h}_2$ are normal polymatroids, we can write them as:

$$\boldsymbol{h}'_1 = \sum_{\boldsymbol{V} \subseteq \boldsymbol{X}-\{X_n\}} b_{\boldsymbol{V}} \boldsymbol{h}^{\boldsymbol{V}} \qquad\qquad \boldsymbol{h}_2 = \sum_{\boldsymbol{V} \subseteq \boldsymbol{X}-\{X_n\}} c_{\boldsymbol{V}} \boldsymbol{h}^{\boldsymbol{V}}$$

where $b_{\boldsymbol{V}}, c_{\boldsymbol{V}}$ are non-negative coefficients. Define $\boldsymbol{h}'$ as:

$$\boldsymbol{h}' \stackrel{\text{def}}{=} \sum_{\boldsymbol{V} \subseteq \boldsymbol{X}-\{X_n\}} b_{\boldsymbol{V}} \boldsymbol{h}^{\boldsymbol{V}} + \sum_{\boldsymbol{V} \subseteq \boldsymbol{X}-\{X_n\}} c_{\boldsymbol{V}} \boldsymbol{h}^{\boldsymbol{V} \cup \{X_n\}} + a\boldsymbol{h}^{\boldsymbol{X}}$$

We claim that $\boldsymbol{h}'$ satisfies the conditions of the lemma. Obviously, $\boldsymbol{h}'$ is a normal polymatroid, it remains to check conditions (a),(b),(c). Observe that the following identities hold, for all $\boldsymbol{U} \subseteq \boldsymbol{X} - \{X_n\}$:

$$\begin{aligned} h'(\boldsymbol{U}) =& h'_1(\boldsymbol{U}) + h_2(\boldsymbol{U}) \\ h'(\boldsymbol{U} \cup \{X_n\}) =& h'_1(\boldsymbol{U}) + h_2(\boldsymbol{X} - \{X_n\}) + a \\ =& h'_1(\boldsymbol{U}) + h(X_n) \end{aligned}$$

We check condition (a). If $\boldsymbol{U}$ does not contain X_n, then

$$h'(\boldsymbol{U}) = h'_1(\boldsymbol{U}) + h_2(\boldsymbol{U}) \leq h(\boldsymbol{U}|X_n) + I(\boldsymbol{U}; X_n) = h(\boldsymbol{U})$$

If $X_n \in \boldsymbol{U}$ then:

$$\begin{aligned} h'(\boldsymbol{U}) =& h'_1(\boldsymbol{U} - \{X_n\}) + h(X_n) \\ \leq& h(\boldsymbol{U} - \{X_n\}|X_n) + h(X_n) = h(\boldsymbol{U}) \end{aligned}$$

Next, we check condition (b):

$$\begin{aligned} h'(\boldsymbol{X}) =& h'_1(\boldsymbol{X} - \{X_n\}) + h(X_n) \\ =& h(\boldsymbol{X} - \{X_n\}|X_n) + h(X_n) = h(\boldsymbol{X}) \end{aligned}$$

Finally, we check condition (c) for X_i, $i < n$:

$$\begin{aligned} h'(X_i) =& h'_1(X_i) + h_2(X_i) = h_1(X_i) + h_2(X_i) \\ =& h(X_i|X_n) + I(X_i; X_n) = h(X_i) \end{aligned}$$

and finally for X_n: $h'(X_n) = h'_1(\emptyset) + h(X_n)$. □