Applications of Information Inequalities to Database Theory Problems

Dan Suciu

University of Washington

Motivation

- Information theory has a long history in databases, e.g. [Lee, 1987].
- Influential work by Atserias, Grohe, Marx [Atserias et al., 2013] lead to successful applications to query upper bounds, worst-case optimal join algorithms, query containment under bag semantics.
- This talk: a overview of some of the recent results, intertwined with a brief tutorial on information theory.
- The paper: contains additional details and topics left out of the talk.

Outline

- AGM Bound and Shannon Inequalities
- Max Degree Bounds and Non-Shannon Inequalities
- Query Domination and Max-Inequalities
- Approximate Implication and Conditional Inequalities
- Conclusions

The AGM Bound

and Shannon Inequalities

Full Conjunctive Query

Relational schema: $R_1, \ldots, R_p$.

Definition (Full conjunctive query (CQ))

$$Q(\boldsymbol{X}) = R_{j_1}(\boldsymbol{Y}_1) \wedge \cdots \wedge R_{j_m}(\boldsymbol{Y}_m), \qquad \text{where } \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \subseteq \boldsymbol{X}.$$

E.g. $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$

- Database instance: $\boldsymbol{D} = (R_1^D, \ldots, R_p^D)$.
- Query output: $Q(\boldsymbol{D})$.

Full Conjunctive Query

Relational schema: $R_1, \ldots, R_p$.

Definition (Full conjunctive query (CQ))

$$Q(\boldsymbol{X}) = R_{j_1}(\boldsymbol{Y}_1) \wedge \cdots \wedge R_{j_m}(\boldsymbol{Y}_m), \qquad \text{where } \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \subseteq \boldsymbol{X}.$$

E.g. $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$

- Database instance: $\boldsymbol{D} = (R_1^D, \ldots, R_p^D)$.
- Query output: $Q(\boldsymbol{D})$.

The Output Size Problem

Given statistics on the input $\boldsymbol{D}$, e.g. cardinalities, # distinct values:

- **Estimation Problem.** Compute "estimate" E:

$$|Q(\boldsymbol{D})| \approx E$$

  Adopted in practice, however it is ill defined.

- **Upper Bound Problem.** Compute an upper bound B:

$$|Q(\boldsymbol{D})| \leq B$$

  Challenge: make B tight.

The Output Size Problem

Given statistics on the input $\boldsymbol{D}$, e.g. cardinalities, # distinct values:

- **Estimation Problem.** Compute "estimate" E:

$$|Q(\boldsymbol{D})| \approx E$$

  Adopted in practice, however it is ill defined.

- **Upper Bound Problem.** Compute an upper bound B:

$$|Q(\boldsymbol{D})| \leq B$$

  Challenge: make B tight.

Simple Examples

Assume $|R| \leq N$, $|S| \leq N$, $|T| \leq N$.

- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z)$. $\quad\quad \max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = ?$

Simple Examples

Assume $|R| \leq N$, $|S| \leq N$, $|T| \leq N$.

- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$

Simple Examples

Assume $|R| \leq N$, $|S| \leq N$, $|T| \leq N$.

- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
  If $S.Y$ is a key: $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N$

Simple Examples

Assume $|R| \leq N$, $|S| \leq N$, $|T| \leq N$.

- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
  If $S.Y$ is a key: $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N$
- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = ?$

Simple Examples

Assume $|R| \leq N$, $|S| \leq N$, $|T| \leq N$.

- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
  If $S.Y$ is a key: $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N$
- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$

Simple Examples

Assume $|R| \leq N$, $|S| \leq N$, $|T| \leq N$.

- $Q(X,Y,Z) = R(X,Y) \wedge S(Y,Z)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
  If $S.Y$ is a key: $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N$
- $Q(X,Y,Z,U) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,U)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
- $Q(X,Y,Z) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| =?$

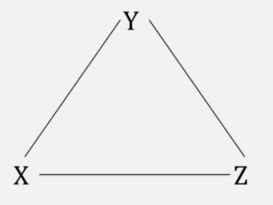
Simple Examples

Assume $|R| \leq N$, $|S| \leq N$, $|T| \leq N$.

- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
  If $S.Y$ is a key: $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N$
- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^{\frac{3}{2}}$

Fractional Edge Covers

Query Q to hypegraph $G = (V, E)$. $R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$

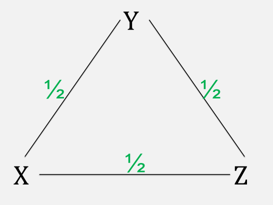


Fractional Edge Covers

Query Q to hypegraph $G = (V, E)$. $R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$

Definition

A *fractional edge cover* is $\boldsymbol{w} = (w_e)_{e \in E}$, $w_e \geq 0$:
$$\forall x \in V, \sum_{e \in E : x \in e} w_e \geq 1.$$



The AGM Bound [Atserias et al., 2013]

$Q(\boldsymbol{X}) = R_1(\boldsymbol{Y}_1) \wedge \cdots \wedge R_m(\boldsymbol{Y}_m)$

Theorem (Upper Bound)

For every fractional edge cover $\boldsymbol{w}$: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

The AGM Bound [Atserias et al., 2013]

$Q(\boldsymbol{X}) = R_1(\boldsymbol{Y}_1) \wedge \cdots \wedge R_m(\boldsymbol{Y}_m)$

Theorem (Upper Bound)

For every fractional edge cover $\boldsymbol{w}$*:* $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Theorem (Lower Bound)

$AGM(Q) \stackrel{def}{=} \min_{\boldsymbol{w}} |R_1|^{w_1} \cdots |R_m|^{w_m}$ *is "tight".*

The AGM Bound [Atserias et al., 2013]

$$Q(\boldsymbol{X}) = R_1(\boldsymbol{Y}_1) \wedge \cdots \wedge R_m(\boldsymbol{Y}_m)$$

Theorem (Upper Bound)

For every fractional edge cover $\boldsymbol{w}$: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Theorem (Lower Bound)

$AGM(Q) \stackrel{def}{=} \min_{\boldsymbol{w}} |R_1|^{w_1} \cdots |R_m|^{w_m}$ *is "tight".*

$$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X) \qquad AGM(Q) = \min \begin{pmatrix} (|R| \cdot |S| \cdot |T|)^{1/2} \\ |R| \cdot |S| \\ |R| \cdot |T| \\ |S| \cdot |T| \end{pmatrix}$$

The AGM Bound [Atserias et al., 2013]

$Q(\boldsymbol{X}) = R_1(\boldsymbol{Y}_1) \wedge \cdots \wedge R_m(\boldsymbol{Y}_m)$

Theorem (Upper Bound)

For every fractional edge cover $\boldsymbol{w}$*:* $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Theorem (Lower Bound)

$AGM(Q) \stackrel{def}{=} \min_{\boldsymbol{w}} |R_1|^{w_1} \cdots |R_m|^{w_m}$ *is "tight".*

$$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X) \qquad AGM(Q) = \min \begin{pmatrix} (|R| \cdot |S| \cdot |T|)^{1/2} \\ |R| \cdot |S| \\ |R| \cdot |T| \\ |S| \cdot |T| \end{pmatrix}$$

Proof. information inequalities.

Entropic Vectors

Definition

Finite probability space $p : D \to [0, 1]$. X = r.v. with outcomes D.

The *entropy* of X is: $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

Entropic Vectors

Definition

Finite probability space $p : D \to [0, 1]$. X = r.v. with outcomes D.

The *entropy* of X is: $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

$N \stackrel{\text{def}}{=} |D|$: $0 \leq h(X) \leq \log N$ $h(X) = \log N$ iff p is uniform.

Entropic Vectors

Definition

Finite probability space $p : D \to [0, 1]$. X = r.v. with outcomes D.

The *entropy* of X is: $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

$N \stackrel{\text{def}}{=} |D|$: $0 \leq h(X) \leq \log N$ $h(X) = \log N$ iff p is uniform.

Definition

R.v. $X_1, \ldots, X_n$. Their *entropic vector* is $\boldsymbol{h} = (h(X_\alpha))_{\alpha \subseteq [n]} \in \mathbb{R}_+^{2^n}$.

Entropic Vectors

Definition

Finite probability space $p : D \to [0, 1]$. X = r.v. with outcomes D.

The *entropy* of X is: $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

$N \stackrel{\text{def}}{=} |D|$: $0 \le h(X) \le \log N$ $h(X) = \log N$ iff p is uniform.

Definition

R.v. $X_1, \ldots, X_n$. Their *entropic vector* is $\boldsymbol{h} = (h(X_\alpha))_{\alpha \subseteq [n]} \in \mathbb{R}_+^{2^n}$.

X	Y
a	p
a	q
b	q
a	m

Entropic Vectors

Definition

Finite probability space $p : D \to [0,1]$. X = r.v. with outcomes D.

The *entropy* of X is: $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

$N \stackrel{\text{def}}{=} |D|$: $0 \leq h(X) \leq \log N$ $h(X) = \log N$ iff p is uniform.

Definition

R.v. $X_1, \ldots, X_n$. Their *entropic vector* is $\boldsymbol{h} = (h(X_\alpha))_{\alpha \subseteq [n]} \in \mathbb{R}_+^{2^n}$.

X	Y	p
a	p	$1/4$
a	q	$1/4$
b	q	$1/4$
a	m	$1/4$

$h(XY) = \log 4$

Entropic Vectors

Definition

Finite probability space $p : D \to [0,1]$. X = r.v. with outcomes D.

The *entropy* of X is: $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

$N \stackrel{\text{def}}{=} |D|$: $0 \leq h(X) \leq \log N$ $h(X) = \log N$ iff p is uniform.

Definition

R.v. $X_1, \ldots, X_n$. Their *entropic vector* is $\boldsymbol{h} = (h(X_\alpha))_{\alpha \subseteq [n]} \in \mathbb{R}_+^{2^n}$.

X	Y	p
a	p	1/4
a	q	1/4
b	q	1/4
a	m	1/4

$h(XY) = \log 4$

X	p
a	3/4
b	1/4

$h(X) \leq \log 2$

Y	p
p	1/4
q	2/4
m	1/4

$h(Y) \leq \log 3$

$\emptyset$	p
	1

$h(\emptyset) = 0$

Information Theory Viewed as Logic

Formulas: $\sum_{\alpha \subseteq [n]} c_\alpha h(X_\alpha) \geq 0$ Information Inequality

Information Theory Viewed as Logic

Formulas: $\sum_{\alpha \subseteq [n]} c_\alpha h(X_\alpha) \geq 0$ Information Inequality

Basic Shannon Inequalities:

$h(\emptyset) = 0$	
$h(\boldsymbol{U} \cup \boldsymbol{V}) \geq h(\boldsymbol{U})$	Monotonicity
$h(\boldsymbol{U}) + h(\boldsymbol{V}) \geq h(\boldsymbol{U} \cup \boldsymbol{V}) + h(\boldsymbol{U} \cap \boldsymbol{V})$	Submodularity

Information Theory Viewed as Logic

Formulas: $\sum_{\alpha\subseteq[n]} c_\alpha h(X_\alpha) \geq 0$ Information Inequality

Basic Shannon Inequalities:

$h(\emptyset) = 0$	
$h(\boldsymbol{U} \cup \boldsymbol{V}) \geq h(\boldsymbol{U})$	Monotonicity
$h(\boldsymbol{U}) + h(\boldsymbol{V}) \geq h(\boldsymbol{U} \cup \boldsymbol{V}) + h(\boldsymbol{U} \cap \boldsymbol{V})$	Submodularity

Model: $\boldsymbol{h} \in \mathbb{R}_+^{2^n}$ $\boldsymbol{h} \models \sum(\cdots) \geq 0$

Information Theory Viewed as Logic

Formulas: $\sum_{\alpha\subseteq[n]} c_\alpha h(X_\alpha) \geq 0$ — Information Inequality

Basic Shannon Inequalities:

$h(\emptyset) = 0$	
$h(\boldsymbol{U} \cup \boldsymbol{V}) \geq h(\boldsymbol{U})$	Monotonicity
$h(\boldsymbol{U}) + h(\boldsymbol{V}) \geq h(\boldsymbol{U} \cup \boldsymbol{V}) + h(\boldsymbol{U} \cap \boldsymbol{V})$	Submodularity

Model: $\boldsymbol{h} \in \mathbb{R}_+^{2^n}$ $\qquad \boldsymbol{h} \models \sum(\cdots) \geq 0$

Classes of Models:

$\Gamma_n \stackrel{\text{def}}{=}$ *polymatroids*: satisfy Shannon inequalities

$\Gamma_n^* \stackrel{\text{def}}{=}$ *entropic vectors*

$M_n \stackrel{\text{def}}{=}$ *modular*: $h(X_1X_2\cdots) = h(X_1) + h(X_2) + \cdots$

$$M_n \subset \Gamma_n^* \subset \Gamma_n (\subset \mathbb{R}_+^{2^n})$$

A Shannon Inequality

Example

$$\Gamma_n \models h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ)$$

A Shannon Inequality

Example

$$\Gamma_n \models h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ)$$

$$h(XY) + h(YZ) + h(XZ)$$

A Shannon Inequality

Example

$$\Gamma_n \models h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ)$$

$$\underline{h(XY) + h(YZ)} + h(XZ)$$

A Shannon Inequality

Example

$$\Gamma_n \models h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ)$$

$$\underline{h(XY) + h(YZ)} + h(XZ)$$
$$\geq h(XYZ) + h(Y) + h(XZ)$$

A Shannon Inequality

Example

$$\Gamma_n \models h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ)$$

$$\underline{h(XY) + h(YZ)} + h(XZ)$$
$$\geq h(XYZ) + \underline{h(Y) + h(XZ)}$$

A Shannon Inequality

Example

$$\Gamma_n \models h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ)$$

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

A Shannon Inequality

Example

$$\Gamma_n \models h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ)$$

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

A Shannon Inequality

Example

$$\Gamma_n \models h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ)$$

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

Note: X is covered 2 times in each expressions. Same for Y, same for Z.

Proof of the AGM Upper Bound: Part 1: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Proof of the AGM Upper Bound: Part 1: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

From Query to Information Inequality:

Example

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X), \qquad |Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2}.$

Proof of the AGM Upper Bound: Part 1: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

From Query to Information Inequality:

Example

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X), \quad |Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2}.$

Instance $\boldsymbol{D} = (R^D, S^D, T^D)$; $p : Q(\boldsymbol{D}) \rightarrow [0, 1]$ uniform; $\boldsymbol{h}$ its entropy.

Proof of the AGM Upper Bound: Part 1: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

From Query to Information Inequality:

Example

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X), \qquad |Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2}.$

Instance $\boldsymbol{D} = (R^D, S^D, T^D)$; $\quad p : Q(\boldsymbol{D}) \to [0, 1]$ uniform; $\quad \boldsymbol{h}$ its entropy.

$$\begin{aligned}\log |R^D| + \log |S^D| + \log |T^D| & \\ \geq & h(XY) + h(YZ) + h(XZ)\end{aligned}$$

Proof of the AGM Upper Bound: Part 1: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

From Query to Information Inequality:

Example

$Q(X,Y,Z) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,X), \qquad |Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2}.$

Instance $\boldsymbol{D} = (R^D, S^D, T^D)$; $\quad p : Q(\boldsymbol{D}) \to [0,1]$ uniform; $\quad \boldsymbol{h}$ its entropy.

$$\log|R^D| + \log|S^D| + \log|T^D| \\ \geq h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ)$$

Proof of the AGM Upper Bound: Part 1: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

From Query to Information Inequality:

Example

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X), \qquad |Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2}.$

Instance $\boldsymbol{D} = (R^D, S^D, T^D)$; $\quad p : Q(\boldsymbol{D}) \to [0, 1]$ uniform; $\quad \boldsymbol{h}$ its entropy.

$$\begin{aligned}
\log|R^D| + \log|S^D| + \log|T^D| & \\
&\geq h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ) \\
&= 2\log|Q(\boldsymbol{D})|
\end{aligned}$$

Proof of the AGM Upper Bound: Part 1: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

From Query to Information Inequality:

Example

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X), \qquad |Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2}.$

Instance $\boldsymbol{D} = (R^D, S^D, T^D)$; $\quad p : Q(\boldsymbol{D}) \rightarrow [0, 1]$ uniform; $\quad \boldsymbol{h}$ its entropy.

$$\begin{aligned}
\log|R^D| + \log|S^D| + \log|T^D| & \\
\geq & h(XY) + h(YZ) + h(XZ) \geq 2h(XYZ) \\
= & 2\log|Q(\boldsymbol{D})|
\end{aligned}$$

Part 2: proof of the inequality $\boxed{\sum_j w_j h(\boldsymbol{Y}_j) \geq h(\boldsymbol{X})}$.

Proof of the AGM Upper Bound: Part 2: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Consider the inequality $\boxed{k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})}$, $k_i \in \mathbb{N}$:

Proof of the AGM Upper Bound: Part 2: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Consider the inequality $\boxed{k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})}$, $k_i \in \mathbb{N}$:

Theorem (Shearer?)

The following are equivalent:
(1) $\Gamma_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(2) $\Gamma_n^* \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(3) $M_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(4) Every variable is "covered" $\geq k_0$ *times.*

[Balister and Bollobás, 2012]

Proof of the AGM Upper Bound: Part 2: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Consider the inequality $\boxed{k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})}$, $k_i \in \mathbb{N}$:

Theorem (Shearer?)

The following are equivalent:
(1) $\Gamma_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(2) $\Gamma_n^* \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(3) $M_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(4) Every variable is "covered" $\geq k_0$ *times.*

[Balister and Bollobás, 2012]

Proof (4) $\Rightarrow$ (1)
Repeatedly replace $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j)$ with $h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

Proof of the AGM Upper Bound: Part 2: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Consider the inequality $k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})$, $k_i \in \mathbb{N}$:

Theorem (Shearer?)

The following are equivalent:
(1) $\Gamma_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(2) $\Gamma_n^* \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(3) $M_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(4) Every variable is "covered" $\geq k_0$ *times.*

[Balister and Bollobás, 2012]

Proof (4) $\Rightarrow$ (1)
Repeatedly replace $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j)$ with $h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

- Every variable remains covered.

Proof of the AGM Upper Bound: Part 2: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Consider the inequality $k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})$, $k_i \in \mathbb{N}$:

Theorem (Shearer?)

The following are equivalent:
(1) $\Gamma_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(2) $\Gamma_n^* \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(3) $M_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(4) Every variable is "covered" $\geq k_0$ *times.*

[Balister and Bollobás, 2012]

Proof (4) $\Rightarrow$ (1)
Repeatedly replace $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j)$ with $h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

- Every variable remains covered.
- $\sum_\ell |\boldsymbol{Y}_\ell|^2$ strictly increases.

Proof of the AGM Upper Bound: Part 2: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Consider the inequality $k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})$, $k_i \in \mathbb{N}$:

Theorem (Shearer?)

The following are equivalent:

(1) $\Gamma_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$

(2) $\Gamma_n^* \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$

(3) $M_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$

(4) Every variable is "covered" $\geq k_0$ *times.*

[Balister and Bollobás, 2012]

Proof (4) $\Rightarrow$ (1)

Repeatedly replace $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j)$ with $h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

- Every variable remains covered.
- $\sum_\ell |\boldsymbol{Y}_\ell|^2$ strictly increases.
- At termination: $k'_1 h(\boldsymbol{Y}'_1) + k'_2 h(\boldsymbol{Y}'_2) + \cdots$ $\boldsymbol{Y}'_1 \supset \boldsymbol{Y}'_2 \supset \cdots$

Proof of the AGM Upper Bound: Part 2: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Consider the inequality $k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})$, $k_i \in \mathbb{N}$:

Theorem (Shearer?)

The following are equivalent:
(1) $\Gamma_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(2) $\Gamma_n^* \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(3) $M_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$
(4) Every variable is "covered" $\geq k_0$ *times.*

[Balister and Bollobás, 2012]

Proof (4) ⇒ (1)
Repeatedly replace $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j)$ with $h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

- Every variable remains covered.
- $\sum_\ell |\boldsymbol{Y}_\ell|^2$ strictly increases.
- At termination: $k_1' h(\boldsymbol{Y}_1') + k_2' h(\boldsymbol{Y}_2') + \cdots$ $\boldsymbol{Y}_1' \supset \boldsymbol{Y}_2' \supset \cdots$
- Thus, $\boldsymbol{Y}_1' = \boldsymbol{X}$ and $k_1' \geq k_0$.

Proof of the AGM Upper Bound: Part 2: $|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}$

Consider the inequality $k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})$, $k_i \in \mathbb{N}$:

Theorem (Shearer?)

The following are equivalent:

(1) $\Gamma_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$

(2) $\Gamma_n^* \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$

(3) $M_n \models \boldsymbol{k} \cdot \boldsymbol{h} \geq k_0 h(\boldsymbol{X})$

(4) Every variable is "covered" $\geq k_0$ *times.*

[Balister and Bollobás, 2012]

Proof (4) $\Rightarrow$ (1)

Repeatedly replace $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j)$ with $h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

- Every variable remains covered.
- $\sum_\ell |\boldsymbol{Y}_\ell|^2$ strictly increases.
- At termination: $k'_1 h(\boldsymbol{Y}'_1) + k'_2 h(\boldsymbol{Y}'_2) + \cdots$ $\boldsymbol{Y}'_1 \supset \boldsymbol{Y}'_2 \supset \cdots$
- Thus, $\boldsymbol{Y}'_1 = \boldsymbol{X}$ and $k'_1 \geq k_0$.

This proves the Upper Bound. Will skip the Lower Bound

Summary of the AGM Bound

- AGM upper bound: apply submodularity in *any* order.
- AGM lower bound: from modular $\boldsymbol{h}^*$ to product relations.

Limitation: AGM uses only cardinality constraints.

Next: add functional dependencies and degree constraints.

The Max-Degree Bound

and Non-Shannon Inequalities

General Statistics

Collect more statistics about the database $\boldsymbol{D}$, such as:

- Relation cardinalities (as in the AGM bound).
- Keys and/or Functional Dependencies.
- Maximum degrees.
- ℓ_p-norms of degree sequences
- ...

Max-Degrees

Fix a relation instance $R(\boldsymbol{X})$, $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$

$$\deg_R(\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u}) \stackrel{\text{def}}{=} |\{\boldsymbol{v} \mid (\boldsymbol{u},\boldsymbol{v}) \in \Pi_{\boldsymbol{UV}}(R)\}|$$

$$\deg_R(\boldsymbol{V}|\boldsymbol{U}) \stackrel{\text{def}}{=} \max_{\boldsymbol{u}} \deg_R(\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u})$$

$R =$

U	V
a	1
a	2
a	3
b	1
b	5

$\deg_R(V|U) = 3.$

Degree constrains generalize:

- Cardinality: $|R| = \deg_R(\boldsymbol{X}|\emptyset)$.
- Functional Dependency $\boldsymbol{U} \to \boldsymbol{V}$: $\deg_R(\boldsymbol{V}|\boldsymbol{U}) = 1$.
- Key $\boldsymbol{U}$: $\deg_R(\boldsymbol{X}|\boldsymbol{U}) = 1$

Max-Degrees

Fix a relation instance $R(\boldsymbol{X})$, $\quad \boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$

$$\mathrm{deg}_R(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}) \stackrel{\text{def}}{=} |\{\boldsymbol{v} \mid (\boldsymbol{u}, \boldsymbol{v}) \in \Pi_{\boldsymbol{UV}}(R)\}|$$

$$\mathrm{deg}_R(\boldsymbol{V}|\boldsymbol{U}) \stackrel{\text{def}}{=} \max_{\boldsymbol{u}} \mathrm{deg}_R(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})$$

$R =$

U	V
a	1
a	2
a	3
b	1
b	5

$\mathrm{deg}_R(V|U) = 3.$

Degree constrains generalize:

- Cardinality: $|R| = \mathrm{deg}_R(\boldsymbol{X}|\emptyset)$.
- Functional Dependency $\boldsymbol{U} \to \boldsymbol{V}$: $\mathrm{deg}_R(\boldsymbol{V}|\boldsymbol{U}) = 1$.
- Key $\boldsymbol{U}$: $\mathrm{deg}_R(\boldsymbol{X}|\boldsymbol{U}) = 1$

Information Measures

The *Conditional Entropy* and *Conditional Mutual Information* are:

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

$$I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) \stackrel{\text{def}}{=} h(\boldsymbol{UV}) + h(\boldsymbol{UW}) - h(\boldsymbol{U}) - h(\boldsymbol{UVW})$$

$\Gamma_n \models h(\boldsymbol{V}|\boldsymbol{U}) \geq 0, \quad \Gamma_n \models I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) \geq 0$

If $\boldsymbol{h} \in \Gamma_n^*$, then:

$$h(\boldsymbol{V}|\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{u}}[h(\boldsymbol{V}|\boldsymbol{U}=\boldsymbol{u})] \leq \log \deg(\boldsymbol{V}|\boldsymbol{U})$$
$$I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) = 0 \text{ iff } \boldsymbol{V} \perp \boldsymbol{W}|\boldsymbol{U}$$

Information Measures

The *Conditional Entropy* and *Conditional Mutual Information* are:

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

$$I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) \stackrel{\text{def}}{=} h(\boldsymbol{UV}) + h(\boldsymbol{UW}) - h(\boldsymbol{U}) - h(\boldsymbol{UVW})$$

$\Gamma_n \models h(\boldsymbol{V}|\boldsymbol{U}) \geq 0, \quad \Gamma_n \models I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) \geq 0$

If $\boldsymbol{h} \in \Gamma_n^*$, then:

$$h(\boldsymbol{V}|\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{u}}[h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})] \leq \log \deg(\boldsymbol{V}|\boldsymbol{U})$$
$$I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) = 0 \ \text{ iff } \boldsymbol{V} \perp \boldsymbol{W}|\boldsymbol{U}$$

Information Measures

The *Conditional Entropy* and *Conditional Mutual Information* are:

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

$$I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) \stackrel{\text{def}}{=} h(\boldsymbol{UV}) + h(\boldsymbol{UW}) - h(\boldsymbol{U}) - h(\boldsymbol{UVW})$$

$\Gamma_n \models h(\boldsymbol{V}|\boldsymbol{U}) \geq 0, \quad \Gamma_n \models I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) \geq 0$

If $\boldsymbol{h} \in \Gamma_n^*$, then:

$$h(\boldsymbol{V}|\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{u}}[h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})] \leq \log \texttt{deg}(\boldsymbol{V}|\boldsymbol{U})$$
$$I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) = 0 \text{ iff } \boldsymbol{V} \perp \boldsymbol{W}|\boldsymbol{U}$$

Max-Degree Bound by Example

Example

$$Q(X,Y,Z,U) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,U) \wedge A(\underline{X,Z},U) \wedge B(X,\underline{Y,U})$$

Max-Degree Bound by Example

Example

$Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge A(\underline{X, Z}, U) \wedge B(X, \underline{Y, U})$

Inequality

$$h(XY) + h(YZ) + h(ZU) + h(U|XZ) + h(X|YU) \geq 2h(XYZU)$$

Implies $\left(|R| \cdot |S| \cdot |T| \cdot \mathtt{deg}_A(U|XZ) \cdot \mathtt{deg}_B(X|YU)\right)^{1/2} \geq |Q|$

Max-Degree Bound by Example

Example

$Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge A(\underline{X, Z}, U) \wedge B(X, \underline{Y, U})$

Inequality

$$\log|R| + \log|S| + \log|T| + \log \mathtt{deg}_A(U|XZ) + \log \mathtt{deg}_B(X|YU) \geq$$
$$h(XY) + h(YZ) + h(ZU) + h(U|XZ) + h(X|YU) \geq 2h(XYZU) = 2\log|Q|$$

Implies $\quad \left(|R| \cdot |S| \cdot |T| \cdot \mathtt{deg}_A(U|XZ) \cdot \mathtt{deg}_B(X|YU)\right)^{1/2} \geq |Q|$

Max-Degree Bound by Example

Example

$$Q(X,Y,Z,U) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,U) \wedge A(\underline{X,Z},U) \wedge B(X,\underline{Y,U})$$

Inequality

$$\underline{h(XY) + h(YZ)} + h(ZU) + h(U|XZ) + h(X|YU)$$

Implies $\left(|R| \cdot |S| \cdot |T| \cdot \texttt{deg}_A(U|XZ) \cdot \texttt{deg}_B(X|YU)\right)^{1/2} \geq |Q|$

Max-Degree Bound by Example

Example

$Q(X,Y,Z,U) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,U) \wedge A(\underline{X,Z},U) \wedge B(X,\underline{Y,U})$

Inequality

$$\underline{h(XY)+h(YZ)}+h(ZU)+h(U|XZ)+h(X|YU)$$
$$\geq h(XYZ)+h(Y)+h(ZU)+h(U|XZ)+h(X|YU)$$

Implies $\left(|R|\cdot|S|\cdot|T|\cdot \texttt{deg}_A(U|XZ)\cdot \texttt{deg}_B(X|YU)\right)^{1/2} \geq |Q|$

Max-Degree Bound by Example

Example

$Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge A(\underline{X, Z}, U) \wedge B(X, \underline{Y, U})$

Inequality

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

Implies $\left(|R| \cdot |S| \cdot |T| \cdot \mathtt{deg}_A(U|XZ) \cdot \mathtt{deg}_B(X|YU)\right)^{1/2} \geq |Q|$

Max-Degree Bound by Example

Example

$$Q(X,Y,Z,U) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,U) \wedge A(\underline{X,Z},U) \wedge B(X,\underline{Y,U})$$

Inequality

$$\underline{h(XY) + h(YZ)} + h(ZU) + h(U|XZ) + h(X|YU)$$
$$\geq h(XYZ) + \underline{h(Y) + h(ZU)} + h(U|XZ) + h(X|YU)$$
$$\geq h(XYZ) + h(YZU) + h(U|XZ) + h(X|YU)$$

Implies $\left(|R| \cdot |S| \cdot |T| \cdot \texttt{deg}_A(U|XZ) \cdot \texttt{deg}_B(X|YU)\right)^{1/2} \geq |Q|$

Max-Degree Bound by Example

Example

$Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge A(\underline{X, Z}, U) \wedge B(X, \underline{Y, U})$

Inequality

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

Implies $\left(|R| \cdot |S| \cdot |T| \cdot \mathtt{deg}_A(U|XZ) \cdot \mathtt{deg}_B(X|YU)\right)^{1/2} \geq |Q|$

Max-Degree Bound by Example

Example

$Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge A(\underline{X, Z}, U) \wedge B(X, \underline{Y, U})$

Inequality

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

Implies $\left(|R| \cdot |S| \cdot |T| \cdot \texttt{deg}_A(U|XZ) \cdot \texttt{deg}_B(X|YU)\right)^{1/2} \geq |Q|$

Max-Degree Bound by Example

Example

$$Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge A(\underline{X, Z}, U) \wedge B(X, \underline{Y, U})$$

Inequality

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

Implies $\left(|R| \cdot |S| \cdot |T| \cdot \texttt{deg}_A(U|XZ) \cdot \texttt{deg}_B(X|YU)\right)^{1/2} \geq |Q|$

The Upper Bound [Khamis et al., 2017]

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

Theorem

If $\Gamma_n^* \models \sum_{i=1,s} w_i h(\boldsymbol{V}_i|\boldsymbol{U}_i) \geq h(\boldsymbol{X})$ *then* $\prod_{i=1,s} deg_{R_{j_i}}^{w_i}(\boldsymbol{V}_i|\boldsymbol{U}_i) \geq |Q|$.

$M_n \models (\cdots) \quad \overset{\Leftarrow}{\not\Rightarrow} \quad \Gamma_n^* \models (\cdots) \quad \overset{\Leftarrow}{\not\Rightarrow} \quad \Gamma_n \models (\cdots)$

The Γ_n^*-bound is tight only "asymptotically".
The Γ_n-bound is not tight, not even asymptotically.

The Upper Bound [Khamis et al., 2017]

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

Theorem

If $\Gamma_n^* \models \sum_{i=1,s} w_i h(\boldsymbol{V}_i|\boldsymbol{U}_i) \geq h(\boldsymbol{X})$ *then* $\prod_{i=1,s} deg_{R_{j_i}}^{w_i}(\boldsymbol{V}_i|\boldsymbol{U}_i) \geq |Q|$.

$$M_n \models (\cdots) \quad \overset{\Leftarrow}{\not\Rightarrow} \quad \Gamma_n^* \models (\cdots) \quad \overset{\Leftarrow}{\not\Rightarrow} \quad \Gamma_n \models (\cdots)$$

The Γ_n^*-bound is tight only "asymptotically".
The Γ_n-bound is not tight, not even asymptotically.

Summary

- Shannon inequalities are sufficient for the AGM bound.
- Shannon inequalities are insufficient for the Max-Degree bound.
- Shannon inequalities are sufficient for the Max-Degree bound for simple degrees.

$$M_n \models (\cdots) \quad \overset{\Leftarrow}{\not\Rightarrow} \quad N_n \models (\cdots) \;\Leftrightarrow\; \Gamma_n^* \models (\cdots) \;\Leftrightarrow \Gamma_n \models (\cdots)$$

Moreover, the bound is tight.

Query Domination and Max-Inequalities

Definition

Fix two full CQs Q, Q'.

Definition

Q' *dominates* Q if $\forall \boldsymbol{D}$, $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$. Write $Q \preceq Q'$.

Query domination problem: decide whether $\boxed{Q \preceq Q'}$

Necessary condition: $\exists \varphi : Q' \rightarrow Q$.
E.g. $R(U,V) \wedge R(V,W) \not\preceq R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$.

Sufficient condition: $\exists \varphi : Q' \rightarrow Q$ surjective.
E.g. $R(X,Y) \wedge R(Y,Z) \wedge R(Z,X) \preceq R(U,V) \wedge R(V,W)$.

Definition

Fix two full CQs Q, Q'.

Definition

Q' *dominates* Q if $\forall \boldsymbol{D}$, $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$. Write $Q \preceq Q'$.

Query domination problem: decide whether $\boxed{Q \preceq Q'}$

Necessary condition: $\exists \varphi : Q' \rightarrow Q$.
E.g. $R(U,V) \wedge R(V,W) \not\preceq R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$.

Sufficient condition: $\exists \varphi : Q' \rightarrow Q$ surjective.
E.g. $R(X,Y) \wedge R(Y,Z) \wedge R(Z,X) \preceq R(U,V) \wedge R(V,W)$.

Definition

Fix two full CQs Q, Q'.

Definition

Q' *dominates* Q if $\forall \boldsymbol{D}$, $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$. Write $Q \preceq Q'$.

Query domination problem: decide whether $\boxed{Q \preceq Q'}$

Necessary condition: $\exists \varphi : Q' \rightarrow Q$.
E.g. $R(U, V) \wedge R(V, W) \;\not\preceq\; R(X, Y) \wedge R(Y, Z) \wedge R(Z, X)$.

Sufficient condition: $\exists \varphi : Q' \rightarrow Q$ surjective.
E.g. $R(X, Y) \wedge R(Y, Z) \wedge R(Z, X) \preceq R(U, V) \wedge R(V, W)$.

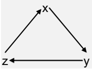
History

Query domination $Q \preceq Q'$ same as *query containment under bag semantics*.

- Introduced in [Chaudhuri and Vardi, 1993].
- Undecidable for Unions of CQ [Ioannidis and Ramakrishnan, 1995].
- Undecidable for CQs with Inequalities [Jayram et al., 2006].
- Sufficient condition [Kopparty and Rossman, 2011].
- Necessary+sufficient condition when Q' is acyclic [Khamis et al., 2021]

Vee's Example [Kopparty and Rossman, 2011]

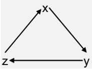
$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$



$Q' = R(U,V) \wedge R(U,W)$

Vee's Example [Kopparty and Rossman, 2011]

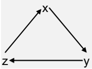
$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$ $\quad$ $Q' = R(U,V) \wedge R(U,W)$



Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$ $Q' = R(U,V) \wedge R(U,W)$

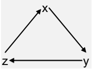


We prove that $Q \preceq Q'$

Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$ $Q' = R(U,V) \wedge R(U,W)$



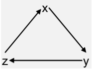
We prove that $Q \preceq Q'$

Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

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

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$ $\qquad Q' = R(U,V) \wedge R(U,W)$



We prove that $Q \preceq Q'$

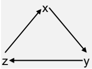
Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \rightarrow Q'$; none surjective.

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

$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X, Y) \wedge R(Y, Z) \wedge R(Z, X)$ $\qquad$ $Q' = R(U, V) \wedge R(U, W)$



We prove that $Q \preceq Q'$

Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

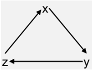
$E \stackrel{\text{def}}{=} h(UV) + h(W|U)$.

$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$

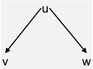
$\max(h(XY) + h(Y|X), h(YZ) + h(Z|Y), h(XZ) + h(X|Z))$

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$ $\qquad$ $Q' = R(U,V) \wedge R(U,W)$



We prove that $Q \preceq Q'$



Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

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

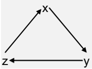
$$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$$

$$\max(h(XY) + h(Y|X), h(YZ) + h(Z|Y), h(XZ) + h(X|Z))$$

$$\geq \frac{1}{3}(h(XY) + h(Y|X) + h(YZ) + h(Z|Y) + h(XZ) + h(X|Z))$$

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$



We prove that $Q \preceq Q'$

$Q' = R(U,V) \wedge R(U,W)$

Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

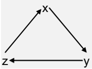
$E \stackrel{\text{def}}{=} h(UV) + h(W|U)$.

$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$

$$\max(h(XY) + h(Y|X), h(YZ) + h(Z|Y), h(XZ) + h(X|Z))$$
$$\geq \frac{1}{3}(h(XY) + h(Y|X) + h(YZ) + h(Z|Y) + h(XZ) + h(X|Z))$$
$$\geq \cdots \geq h(XYZ)$$

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$ $Q' = R(U,V) \wedge R(U,W)$



We prove that $Q \preceq Q'$

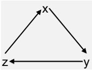
Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

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

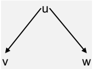
$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X, Y) \wedge R(Y, Z) \wedge R(Z, X)$ $Q' = R(U, V) \wedge R(U, W)$



We prove that $Q \preceq Q'$



Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \rightarrow Q'$; none surjective.

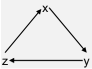
$E \stackrel{\text{def}}{=} h(UV) + h(W|U)$.

$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$

Fix a database instance $\boldsymbol{D}$. We prove $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$:

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$



$Q' = R(U,V) \wedge R(U,W)$

We prove that $Q \preceq Q'$

Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

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

$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$

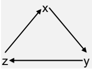
Fix a database instance $\boldsymbol{D}$. We prove $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$:

p on $Q(\boldsymbol{D})$: uniform.

Assume $E \circ \varphi_2 = h(YZ) + h(Z|Y) \geq h(XYZ)$.

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$ $Q' = R(U,V) \wedge R(U,W)$



We prove that $Q \preceq Q'$

Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

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

$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$

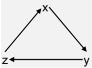
Fix a database instance $\boldsymbol{D}$. We prove $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$:

p on $Q(\boldsymbol{D})$: uniform. Assume $E \circ \varphi_2 = h(YZ) + h(Z|Y) \geq h(XYZ)$.

p' on $Q'(\boldsymbol{D})$, $V \perp W|U$: $p'(U) = p(Y)$, $p'(V|U) = p'(W|U) = p(Z|Y)$:

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$ $\qquad$ $Q' = R(U,V) \wedge R(U,W)$



We prove that $Q \preceq Q'$

Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

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

$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$

Fix a database instance $\boldsymbol{D}$. We prove $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$:

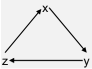
p on $Q(\boldsymbol{D})$: uniform. Assume $E \circ \varphi_2 = h(YZ) + h(Z|Y) \geq h(XYZ)$.

p' on $Q'(\boldsymbol{D})$, $V \perp W|U$: $p'(U) = p(Y)$, $p'(V|U) = p'(W|U) = p(Z|Y)$:

$\log|Q'(\boldsymbol{D})| \geq h'(UVW) = h'(U) + h'(VW|U)$

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$ $\qquad$ $Q' = R(U,V) \wedge R(U,W)$



We prove that $Q \preceq Q'$

Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

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

$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$

Fix a database instance $\boldsymbol{D}$. We prove $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$:

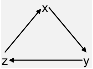
p on $Q(\boldsymbol{D})$: uniform. $\qquad$ Assume $E \circ \varphi_2 = h(YZ) + h(Z|Y) \geq h(XYZ)$.

p' on $Q'(\boldsymbol{D})$, $V \perp W|U$: $\quad p'(U) = p(Y)$, $p'(V|U) = p'(W|U) = p(Z|Y)$:

$$\log|Q'(\boldsymbol{D})| \geq h'(UVW) = h'(U) + h'(VW|U)$$
$$= h'(U) + h'(V|U) + h'(W|U)$$

Vee's Example [Kopparty and Rossman, 2011]

$Q = R(X,Y) \wedge R(Y,Z) \wedge R(Z,X)$ $\quad$ $Q' = R(U,V) \wedge R(U,W)$



We prove that $Q \preceq Q'$

Three homomorphisms $\varphi_1, \varphi_2, \varphi_3 : Q \to Q'$; none surjective.

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

$\max(E \circ \varphi_1, E \circ \varphi_2, E \circ \varphi_3) \geq h(XYZ)$

Fix a database instance $\boldsymbol{D}$. We prove $|Q(\boldsymbol{D})| \leq |Q'(\boldsymbol{D})|$:

p on $Q(\boldsymbol{D})$: uniform. $\quad$ Assume $E \circ \varphi_2 = h(YZ) + h(Z|Y) \geq h(XYZ)$.

p' on $Q'(\boldsymbol{D})$, $V \perp W|U$: $\quad p'(U) = p(Y)$, $p'(V|U) = p'(W|U) = p(Z|Y)$:

$$\begin{aligned}
\log|Q'(\boldsymbol{D})| \geq h'(UVW) &= h'(U) + h'(VW|U) \\
&= h'(U) + h'(V|U) + h'(W|U) \\
&= h(Y) + 2h(Z|Y) \geq h(XYZ) = \log|Q(\boldsymbol{D})|
\end{aligned}$$

Domination and Max-Inequalities

Fix Q, Q' and assume Q' is acyclic.

$$E_{Q'} \stackrel{\text{def}}{=} \sum_{A \in \text{atoms}(Q')} h(\text{vars}(A) | \text{vars}(A) \cap \text{vars}(\text{parent}(A)))$$

Theorem ([Kopparty and Rossman, 2011, Khamis et al., 2021])

$Q \preceq Q'$ *iff* $\max_{\varphi \in \text{hom}(Q', Q)}(E_{Q'} \circ \varphi) \geq h(vars(Q))$ *is valid.*

Domination and Max-Inequalities

Fix Q, Q' and assume Q' is acyclic.

$E_{Q'} \stackrel{\text{def}}{=} \sum_{A \in \text{atoms}(Q')} h(\text{vars}(A) | \text{vars}(A) \cap \text{vars}(\text{parent}(A)))$

Theorem ([Kopparty and Rossman, 2011, Khamis et al., 2021])

$Q \preceq Q'$ *iff* $\max_{\varphi \in \text{hom}(Q', Q)}(E_{Q'} \circ \varphi) \geq h(vars(Q))$ *is valid.*

- Max-inequalities and the domination problem $Q \preceq Q'$ for Q' acyclic are computationally equivalent [Khamis et al., 2021].
- If any two atoms in Q' share at most one variable, then $Q \preceq Q'$ is decidable.

Approximate Implication and

Conditional Inequalities

Constraints (or Dependencies)

Fix a relation $R(\boldsymbol{X})$.

Functional Dependency (FD) $\boxed{\boldsymbol{U} \rightarrow \boldsymbol{V}}$ for $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$.

Multivalued Dependency (MVD) $\boxed{\boldsymbol{U} \twoheadrightarrow \boldsymbol{V} | \boldsymbol{W}}$ for $\boldsymbol{UVW} = \boldsymbol{X}$.

Goal: generalize to Soft Constraints (or Soft Dependencies).

Constraints (or Dependencies)

Fix a relation $R(\boldsymbol{X})$.

Functional Dependency (FD) $\boxed{\boldsymbol{U} \rightarrow \boldsymbol{V}}$ for $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$.

Multivalued Dependency (MVD) $\boxed{\boldsymbol{U} \twoheadrightarrow \boldsymbol{V} | \boldsymbol{W}}$ for $\boldsymbol{U}\boldsymbol{V}\boldsymbol{W} = \boldsymbol{X}$.

Goal: generalize to Soft Constraints (or Soft Dependencies).

Constraints (or Dependencies)

Fix a relation $R(\boldsymbol{X})$.

Functional Dependency (FD) $\boxed{\boldsymbol{U} \to \boldsymbol{V}}$ for $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$.

Multivalued Dependency (MVD) $\boxed{\boldsymbol{U} \twoheadrightarrow \boldsymbol{V} | \boldsymbol{W}}$ for $\boldsymbol{UVW} = \boldsymbol{X}$.

Goal: generalize to Soft Constraints (or Soft Dependencies).

The Constraint Implication Problem

Constraint Implication Problem

Given constraints $\sigma_0, \sigma_1, \ldots, \sigma_p$, check if $\sigma_1 \wedge \cdots \wedge \sigma_p \Rightarrow \sigma_0$.

[Armstrong, 1974] axiomatization for FDs.

[Beeri et al., 1977]: axiomatization for FDs and MVDs.
E.g. $(A \twoheadrightarrow B|CD) \Rightarrow (AC \twoheadrightarrow B|D)$

The Constraint Implication Problem

Constraint Implication Problem

Given constraints $\sigma_0, \sigma_1, \ldots, \sigma_p$, check if $\sigma_1 \wedge \cdots \wedge \sigma_p \Rightarrow \sigma_0$.

[Armstrong, 1974] axiomatization for FDs.

[Beeri et al., 1977]: axiomatization for FDs and MVDs.
E.g. $(A \twoheadrightarrow B|CD) \Rightarrow (AC \twoheadrightarrow B|D)$

Relaxation Problem (informal)

If R satisfies $\sigma_1, \ldots, \sigma_p$ approximatively, does it satisfy σ_0 approximatively?

From Constraints to Information Measure

Fix $R(\boldsymbol{X})$, let $p : R \to [0, 1]$ be uniform, $\boldsymbol{h}$ its entropy.

Theorem ([Lee, 1987])

$R \models \boldsymbol{U} \to \boldsymbol{V}$ *iff* $h(\boldsymbol{V}|\boldsymbol{U}) = 0$

$R \models \boldsymbol{U} \twoheadrightarrow \boldsymbol{V}|\boldsymbol{W}$ *iff* $I(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U}) = 0$

From Constraints to Information Measure

Fix $R(\boldsymbol{X})$, let $p : R \to [0, 1]$ be uniform, $\boldsymbol{h}$ its entropy.

Theorem ([Lee, 1987])

$$R \models \boldsymbol{U} \to \boldsymbol{V} \quad \textit{iff} \quad h(\boldsymbol{V}|\boldsymbol{U}) = 0$$

$$R \models \boldsymbol{U} \twoheadrightarrow \boldsymbol{V}|\boldsymbol{W} \quad \textit{iff} \quad I(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U}) = 0$$

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

From Constraints to Information Measure

Fix $R(\boldsymbol{X})$, let $p : R \to [0, 1]$ be uniform, $\boldsymbol{h}$ its entropy.

Theorem ([Lee, 1987])

$$R \models \boldsymbol{U} \to \boldsymbol{V} \quad \text{iff} \quad h(\boldsymbol{V}|\boldsymbol{U}) = 0$$
$$R \models \boldsymbol{U} \twoheadrightarrow \boldsymbol{V}|\boldsymbol{W} \quad \text{iff} \quad I(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U}) = 0$$

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

$$I(B; CD|A) = 0 \quad \Rightarrow \quad I(B; D|AC) = 0$$ conditional (in)equality!

From Constraints to Information Measure

Fix $R(\boldsymbol{X})$, let $p : R \to [0, 1]$ be uniform, $\boldsymbol{h}$ its entropy.

Theorem ([Lee, 1987])

$$R \models \boldsymbol{U} \to \boldsymbol{V} \quad \textit{iff} \quad h(\boldsymbol{V}|\boldsymbol{U}) = 0$$
$$R \models \boldsymbol{U} \twoheadrightarrow \boldsymbol{V}|\boldsymbol{W} \quad \textit{iff} \quad I(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U}) = 0$$

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

$I(B; CD|A) = 0 \quad \Rightarrow \quad I(B; D|AC) = 0$ conditional (in)equality!

Proof: $I(B; CD|A) = I(B; D|AC) + I(B; C|A) \geq I(B; D|AC)$.

Conditional Information Inequalities

Information inequality: $0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h}$

Conditional information inequality: $\bigwedge_i (0 \geq \boldsymbol{c}_i \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h})$

Conditional Information Inequalities

Information inequality: $0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h}$

Conditional information inequality: $\bigwedge_i (0 \geq \boldsymbol{c}_i \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h})$

Example: $0 \geq I(B; CD|A) \quad \Rightarrow \quad 0 \geq I(B; D|AC)$

Because: $I(B; CD|A) \geq I(B; D|AC)$

The Relaxation Problem

Does the following hold?

If $\boxed{\bigwedge_i (0 \geq \boldsymbol{c}_i \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h})}$ then $\exists \lambda_i \geq 0$, $\boxed{\sum_i \lambda_i \boldsymbol{c}_i \cdot \boldsymbol{h} \geq \boldsymbol{c}_0 \cdot \boldsymbol{h}}$

The Relaxation Problem

Does the following hold?

If $\boxed{\bigwedge_i (0 \geq \boldsymbol{c}_i \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h})}$ then $\exists \lambda_i \geq 0$, $\boxed{\sum_i \lambda_i \boldsymbol{c}_i \cdot \boldsymbol{h} \geq \boldsymbol{c}_0 \cdot \boldsymbol{h}}$

Deduction theorem in logic: if $\boxed{\Sigma, \varphi \models \psi}$ then $\boxed{\Sigma \models \varphi \Rightarrow \psi}$.

The Relaxation Problem

Does the following hold?

If $\boxed{\bigwedge_i (0 \geq \boldsymbol{c_i} \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c_0} \cdot \boldsymbol{h})}$ then $\exists \lambda_i \geq 0$, $\boxed{\sum_i \lambda_i \boldsymbol{c_i} \cdot \boldsymbol{h} \geq \boldsymbol{c_0} \cdot \boldsymbol{h}}$

Deduction theorem in logic: if $\boxed{\Sigma, \varphi \models \psi}$ then $\boxed{\Sigma \models \varphi \Rightarrow \psi}$.

Positive results:

The Relaxation Problem

Does the following hold?

If $\boxed{\bigwedge_i (0 \geq \boldsymbol{c_i} \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h})}$ then $\exists \lambda_i \geq 0$, $\boxed{\sum_i \lambda_i \boldsymbol{c_i} \cdot \boldsymbol{h} \geq \boldsymbol{c}_0 \cdot \boldsymbol{h}}$

Deduction theorem in logic: if $\boxed{\Sigma, \varphi \models \psi}$ then $\boxed{\Sigma \models \varphi \Rightarrow \psi}$.

Positive results:

- Any FD/MVD implication relaxes [Kenig and Suciu, 2022].

The Relaxation Problem

Does the following hold?

If $\boxed{\bigwedge_i (0 \geq \boldsymbol{c}_i \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h})}$ then $\exists \lambda_i \geq 0$, $\boxed{\sum_i \lambda_i \boldsymbol{c}_i \cdot \boldsymbol{h} \geq \boldsymbol{c}_0 \cdot \boldsymbol{h}}$

Deduction theorem in logic: if $\boxed{\Sigma, \varphi \models \psi}$ then $\boxed{\Sigma \models \varphi \Rightarrow \psi}$.

Positive results:

- Any FD/MVD implication relaxes [Kenig and Suciu, 2022].
- Any FD/MVD implication is a Shannon conditional inequality:

$$M_n \models (\cdots) \quad \overset{\Leftarrow}{\not\Rightarrow} \quad N_n \models (\cdots) \quad \Leftrightarrow \quad \Gamma_n^* \models (\cdots) \quad \Leftrightarrow \quad \Gamma_n \models (\cdots)$$

The Relaxation Problem

Does the following hold?

If $\boxed{\bigwedge_i (0 \geq \boldsymbol{c}_i \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h})}$ then $\exists \lambda_i \geq 0$, $\boxed{\sum_i \lambda_i \boldsymbol{c}_i \cdot \boldsymbol{h} \geq \boldsymbol{c}_0 \cdot \boldsymbol{h}}$

Deduction theorem in logic: if $\boxed{\Sigma, \varphi \models \psi}$ then $\boxed{\Sigma \models \varphi \Rightarrow \psi}$.

Positive results:

- Any FD/MVD implication relaxes [Kenig and Suciu, 2022].
- Any FD/MVD implication is a Shannon conditional inequality:

$$M_n \models (\cdots) \quad \overset{\Leftarrow}{\not\Rightarrow} \quad N_n \models (\cdots) \quad \Leftrightarrow \quad \Gamma_n^* \models (\cdots) \quad \Leftrightarrow \quad \Gamma_n \models (\cdots)$$

- If an FD/MVD implication fails, then it fails on some R with 2 tuples.

The Relaxation Problem

Does the following hold?

If $\boxed{\bigwedge_i (0 \geq \boldsymbol{c_i} \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c_0} \cdot \boldsymbol{h})}$ then $\exists \lambda_i \geq 0$, $\boxed{\sum_i \lambda_i \boldsymbol{c_i} \cdot \boldsymbol{h} \geq \boldsymbol{c_0} \cdot \boldsymbol{h}}$

Deduction theorem in logic: if $\boxed{\Sigma, \varphi \models \psi}$ then $\boxed{\Sigma \models \varphi \Rightarrow \psi}$.

Negative Results

The Relaxation Problem

Does the following hold?

If $\boxed{\bigwedge_i (0 \geq \boldsymbol{c}_i \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h})}$ then $\exists \lambda_i \geq 0$, $\boxed{\sum_i \lambda_i \boldsymbol{c}_i \cdot \boldsymbol{h} \geq \boldsymbol{c}_0 \cdot \boldsymbol{h}}$

Deduction theorem in logic: if $\boxed{\Sigma, \varphi \models \psi}$ then $\boxed{\Sigma \models \varphi \Rightarrow \psi}$.

Negative Results

- The following does not relax [Kaced and Romashchenko, 2013]: $I(X;Y|A) = I(X;Y|B) = I(A;B) = I(A;X|Y) = 0 \;\; \Rightarrow \;\; I(X;Y) = 0.$

The Relaxation Problem

Does the following hold?

If $\boxed{\bigwedge_i (0 \geq \boldsymbol{c}_i \cdot \boldsymbol{h}) \Rightarrow (0 \geq \boldsymbol{c}_0 \cdot \boldsymbol{h})}$ then $\exists \lambda_i \geq 0$, $\boxed{\sum_i \lambda_i \boldsymbol{c}_i \cdot \boldsymbol{h} \geq \boldsymbol{c}_0 \cdot \boldsymbol{h}}$

Deduction theorem in logic: if $\boxed{\Sigma, \varphi \models \psi}$ then $\boxed{\Sigma \models \varphi \Rightarrow \psi}$.

Negative Results

- The following does not relax [Kaced and Romashchenko, 2013]: $I(X;Y|A) = I(X;Y|B) = I(A;B) = I(A;X|Y) = 0 \;\; \Rightarrow \;\; I(X;Y) = 0.$
- Every implication relaxes *with error* [Kenig and Suciu, 2022]:

$$\forall \varepsilon > 0, \exists \lambda_i \geq 0, \qquad \sum_i \lambda_i \boldsymbol{c}_i \cdot \boldsymbol{h} + \varepsilon h(\boldsymbol{X}) \geq \boldsymbol{c}_0 \cdot \boldsymbol{h}.$$

Discussion

- Relaxation (Deduction Theorem) fails for Γ_n^* but holds for Γ_n.
- A subtle issue: semantics differs for Γ_n^* and for $\bar{\Gamma}_n^*$.
- Results on FD/MVD extend to *Approximate Acyclic Schemas* [Kenig et al., 2020].
- Open problem: "Soft Logic" based on information measures?

Conclusions

Summary

- AGM Bound.
- Max-Degree Bound.
- Query Domination.
- Approximate Implication.

Information Theory: both a logic, and a tool for logic.

Some Open Problems

- Is $\Gamma_n^* \models \boldsymbol{c} \cdot \boldsymbol{h} \geq 0$ decidable?
- What is the complexity of $\Gamma_n \models \boldsymbol{c} \cdot \boldsymbol{h} \geq 0$ as a function of $||\boldsymbol{c}||_1$?
- Is $Q \preceq Q'$ decidable?
- “Soft logic” based on information theory.

Some Open Problems

- Is $\Gamma_n^* \models \boldsymbol{c} \cdot \boldsymbol{h} \geq 0$ decidable?
- What is the complexity of $\Gamma_n \models \boldsymbol{c} \cdot \boldsymbol{h} \geq 0$ as a function of $||\boldsymbol{c}||_1$?
- Is $Q \preceq Q'$ decidable?
- "Soft logic" based on information theory.

THANK YOU!

Armstrong, W. W. (1974).
Dependency structures of data base relationships.
In Rosenfeld, J. L., editor, *Information Processing, Proceedings of the 6th IFIP Congress 1974, Stockholm, Sweden, August 5-10, 1974*, pages 580–583. North-Holland.

Atserias, A., Grohe, M., and Marx, D. (2013).
Size bounds and query plans for relational joins.
SIAM J. Comput., 42(4):1737–1767.

Balister, P. and Bollobás, B. (2012).
Projections, entropy and sumsets.
Comb., 32(2):125–141.

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

Chan, T. H. and Yeung, R. W. (2002).
On a relation between information inequalities and group theory.
IEEE Transactions on Information Theory, 48(7):1992–1995.

Chaudhuri, S. and Vardi, M. Y. (1993).
Optimization of *Real* conjunctive queries.
In *ACM PODS, 1993*, pages 59–70.

Gogacz, T. and Torunczyk, S. (2017).
Entropy bounds for conjunctive queries with functional dependencies.
In Benedikt, M. and Orsi, G., 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.

Gottlob, G., Lee, S. T., Valiant, G., and Valiant, P. (2012).

Size and treewidth bounds for conjunctive queries.
J. ACM, 59(3):16:1–16:35.

Ioannidis, Y. E. and Ramakrishnan, R. (1995).
Containment of conjunctive queries: Beyond relations as sets.
ACM Trans. Database Syst., 20(3):288–324.

Jayram, T. S., Kolaitis, P. G., and Vee, E. (2006).
The containment problem for REAL conjunctive queries with inequalities.
In *ACM PODS, 2006*, pages 80–89.

Kaced, T. and Romashchenko, A. E. (2013).
Conditional information inequalities for entropic and almost entropic points.
IEEE Trans. Inf. Theory, 59(11):7149–7167.

Kenig, B., Mundra, P., Prasaad, G., Salimi, B., and Suciu, D. (2020).
Mining approximate acyclic schemes from relations.
In Maier, D., Pottinger, R., Doan, A., Tan, W., Alawini, A., and Ngo, H. Q., 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.

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

Khamis, M. A., Kolaitis, P. G., Ngo, H. Q., and Suciu, D. (2021).
Bag query containment and information theory.
ACM Trans. Database Syst., 46(3):12:1–12:39.

Khamis, M. A., Ngo, H. Q., and Suciu, D. (2016).
Computing join queries with functional dependencies.

In Milo, T. and Tan, W., 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.

Khamis, M. A., Ngo, H. Q., and Suciu, D. (2017).
What do shannon-type inequalities, submodular width, and disjunctive datalog have to do with one another?
In Sallinger, E., den Bussche, J. V., and Geerts, F., 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.
Extended version available at http://arxiv.org/abs/1612.02503.

Kopparty, S. and Rossman, B. (2011).
The homomorphism domination exponent.
Eur. J. Comb., 32(7):1097–1114.

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

Matús, F. (2007).
Infinitely many information inequalities.
In *IEEE International Symposium on Information Theory, ISIT 2007, Nice, France, June 24-29, 2007*, pages 41–44. IEEE.

Pippenger, N. (1986).
What are the laws of information theory.
In *1986 Special Problems on Communication and Computation Conference*, pages 3–5.

Zhang, Z. and Yeung, R. W. (1998).
On characterization of entropy function via information inequalities.
IEEE Transactions on Information Theory, 44(4):1440–1452.

Brief History of Upper Bounds on the Query's Output

- The AGM bound [Atserias et al., 2013]
- Add Functional Dependencies [Gottlob et al., 2012, Khamis et al., 2016, Gogacz and Torunczyk, 2017]
- Add Degree Constraints [Khamis et al., 2017].

All results use information inequalities

- Worst Case Optimal Algorithms:
  Generic Join, Leapfrog Tree Join, PANDA – see paper.

Brief History of Upper Bounds on the Query's Output

- The AGM bound [Atserias et al., 2013]
- Add Functional Dependencies [Gottlob et al., 2012, Khamis et al., 2016, Gogacz and Torunczyk, 2017]
- Add Degree Constraints [Khamis et al., 2017].

All results use information inequalities

- Worst Case Optimal Algorithms:
  Generic Join, Leapfrog Tree Join, PANDA – see paper.

Proof of the AGM Lower Bound

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\quad$ $AGM(Q) = |R|^{w_1} \cdot |S|^{w_2} \cdot |T|^{w_3}$

Primal program:
Minimize $w_1 \log|R| + w_2 \log|S| + w_3 \log|T|$
Where $(w_1, w_2, w_3) \in \mathbb{R}_+^3$
$\models w_1 h(XY) + w_2 h(YZ) + w_3 h(XZ) \geq h(XYZ)$

Dual program:
Maximize $h(XYZ)$
Where

$$h(XY) \leq \log|R|$$
$$h(YZ) \leq \log|S|$$
$$h(XZ) \leq \log|T|$$

From the optimal, modular dual $\boldsymbol{h}^*$, we construct a worst-case instance R^*, S^*, T^*:

$$R^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Y)} \rfloor \qquad S^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(Y)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor \qquad T^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$$

$$|Q^*| = \lfloor 2^{h^*(X)} \rfloor \cdot \lfloor 2^{h^*(Y)} \rfloor \cdot \lfloor 2^{h^*(Z)} \rfloor \geq \frac{1}{8} 2^{h^*(XYZ)} = \frac{1}{8} 2^{w_1^* \log|R| + w_2^* \log|S| + w_3^* \log|T|} = \frac{1}{8} 2^{AGM(Q)}$$

Proof of the AGM Lower Bound

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\qquad$ $AGM(Q) = |R|^{w_1} \cdot |S|^{w_2} \cdot |T|^{w_3}$

Primal program:
Minimize $w_1 \log|R| + w_2 \log|S| + w_3 \log|T|$
Where $(w_1, w_2, w_3) \in \mathbb{R}_+^3$
$\Gamma_n^* \models w_1 h(XY) + w_2 h(YZ) + w_3 h(XZ) \geq h(XYZ)$
w_1, w_2, w_3 frac. edge cover

Dual program:
Maximize $h(XYZ)$
Where

$$h(XY) \leq \log|R|$$
$$h(YZ) \leq \log|S|$$
$$h(XZ) \leq \log|T|$$

$h(X), h(Y), h(Z)$
frac. edge packing

From the optimal, modular dual h^*, we construct a worst-case instance R^*, S^*, T^*:

$R^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Y)} \rfloor$ $\qquad$ $S^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(Y)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$ $\qquad$ $T^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$

$$|Q^*| = \lfloor 2^{h^*(X)} \rfloor \cdot \lfloor 2^{h^*(Y)} \rfloor \cdot \lfloor 2^{h^*(Z)} \rfloor \geq \frac{1}{8} 2^{h^*(XYZ)} = \frac{1}{8} 2^{w_1^* \log|R| + w_2^* \log|S| + w_3^* \log|T|} = \frac{1}{8} 2^{AGM(Q)}$$

Proof of the AGM Lower Bound

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\quad\quad$ $AGM(Q) = |R|^{w_1} \cdot |S|^{w_2} \cdot |T|^{w_3}$

Primal program:
Minimize $w_1 \log|R| + w_2 \log|S| + w_3 \log|T|$
Where $(w_1, w_2, w_3) \in \mathbb{R}_+^3$
$\Gamma_n^* \models w_1 h(XY) + w_2 h(YZ) + w_3 h(XZ) \geq h(XYZ)$
w_1, w_2, w_3 frac. edge cover

Dual program:
Maximize $h(XYZ)$
Where $\boldsymbol{h} \in \Gamma_n^*$

$$h(XY) \leq \log|R|$$
$$h(YZ) \leq \log|S|$$
$$h(XZ) \leq \log|T|$$

$h(X), h(Y), h(Z)$
frac. edge packing

From the optimal, modular dual $\boldsymbol{h}^*$, we construct a worst-case instance R^*, S^*, T^*:

$R^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Y)} \rfloor \quad S^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(Y)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor \quad T^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$

$$|Q^*| = \lfloor 2^{h^*(X)} \rfloor \cdot \lfloor 2^{h^*(Y)} \rfloor \cdot \lfloor 2^{h^*(Z)} \rfloor \geq \frac{1}{8} 2^{h^*(XYZ)} = \frac{1}{8} 2^{w_1^* \log|R| + w_2^* \log|S| + w_3^* \log|T|} = \frac{1}{8} 2^{AGM(Q)}$$

Proof of the AGM Lower Bound

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$

$AGM(Q) = |R|^{w_1} \cdot |S|^{w_2} \cdot |T|^{w_3}$

Primal program:
Minimize $w_1 \log|R| + w_2 \log|S| + w_3 \log|T|$
Where $(w_1, w_2, w_3) \in \mathbb{R}_+^3$
$M_n \models w_1 h(XY) + w_2 h(YZ) + w_3 h(XZ) \geq h(XYZ)$
w_1, w_2, w_3 frac. edge cover

Dual program:
Maximize $h(XYZ)$
Where $\boldsymbol{h} \in M_n$

$$h(XY) \leq \log|R|$$
$$h(YZ) \leq \log|S|$$
$$h(XZ) \leq \log|T|$$

$h(X), h(Y), h(Z)$
frac. edge packing

From the optimal, modular dual $\boldsymbol{h}^*$, we construct a worst-case instance R^*, S^*, T^*:

$$R^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Y)} \rfloor \qquad S^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(Y)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor \qquad T^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$$

$$|Q^*| = \lfloor 2^{h^*(X)} \rfloor \cdot \lfloor 2^{h^*(Y)} \rfloor \cdot \lfloor 2^{h^*(Z)} \rfloor \geq \frac{1}{8} 2^{h^*(XYZ)} = \frac{1}{8} 2^{w_1^* \log|R| + w_2^* \log|S| + w_3^* \log|T|} = \frac{1}{8} 2^{AGM(Q)}$$

Proof of the AGM Lower Bound

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\quad AGM(Q) = |R|^{w_1} \cdot |S|^{w_2} \cdot |T|^{w_3}$

Primal program:
Minimize $w_1 \log|R| + w_2 \log|S| + w_3 \log|T|$
Where $(w_1, w_2, w_3) \in \mathbb{R}_+^3$
$M_n \models w_1 h(XY) + w_2 h(YZ) + w_3 h(XZ) \geq h(XYZ)$
w_1, w_2, w_3 frac. edge cover

Dual program:
Maximize $h(XYZ)$
Where $\boldsymbol{h} \in M_n$

$$h(XY) \leq \log|R|$$
$$h(YZ) \leq \log|S|$$
$$h(XZ) \leq \log|T|$$

$h(X), h(Y), h(Z)$
frac. edge packing

From the optimal, modular dual $\boldsymbol{h}^*$, we construct a worst-case instance R^*, S^*, T^*:

$$R^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Y)} \rfloor \qquad S^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(Y)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor \qquad T^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$$

$$|Q^*| = \lfloor 2^{h^*(X)} \rfloor \cdot \lfloor 2^{h^*(Y)} \rfloor \cdot \lfloor 2^{h^*(Z)} \rfloor \geq \frac{1}{8} 2^{h^*(XYZ)} = \frac{1}{8} 2^{w_1^* \log|R| + w_2^* \log|S| + w_3^* \log|T|} = \frac{1}{8} 2^{AGM(Q)}$$

Proof of the AGM Lower Bound

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\qquad$ $AGM(Q) = |R|^{w_1} \cdot |S|^{w_2} \cdot |T|^{w_3}$

Primal program:
Minimize $w_1 \log|R| + w_2 \log|S| + w_3 \log|T|$
Where $(w_1, w_2, w_3) \in \mathbb{R}_+^3$
$M_n \models w_1 h(XY) + w_2 h(YZ) + w_3 h(XZ) \geq h(XYZ)$
w_1, w_2, w_3 frac. edge cover

Dual program:
Maximize $h(XYZ)$
Where $\boldsymbol{h} \in M_n$

$$h(XY) \leq \log|R|$$
$$h(YZ) \leq \log|S|$$
$$h(XZ) \leq \log|T|$$

$h(X), h(Y), h(Z)$
frac. edge packing

From the optimal, modular dual $\boldsymbol{h}^*$, we construct a worst-case instance R^*, S^*, T^*:

$$R^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Y)} \rfloor \qquad S^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(Y)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor \qquad T^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$$

$$|Q^*| = \lfloor 2^{h^*(X)} \rfloor \cdot \lfloor 2^{h^*(Y)} \rfloor \cdot \lfloor 2^{h^*(Z)} \rfloor \geq \frac{1}{8} 2^{h^*(XYZ)} = \frac{1}{8} 2^{w_1^* \log|R| + w_2^* \log|S| + w_3^* \log|T|} = \frac{1}{8} 2^{AGM(Q)}$$

Proof of the AGM Lower Bound

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$

$AGM(Q) = |R|^{w_1} \cdot |S|^{w_2} \cdot |T|^{w_3}$

Primal program:
Minimize $w_1 \log|R| + w_2 \log|S| + w_3 \log|T|$
Where $(w_1, w_2, w_3) \in \mathbb{R}^3_+$
$M_n \models w_1 h(XY) + w_2 h(YZ) + w_3 h(XZ) \geq h(XYZ)$
w_1, w_2, w_3 frac. edge cover

Dual program:
Maximize $h(XYZ)$
Where $\boldsymbol{h} \in M_n$
$h(XY) \leq \log|R|$
$h(YZ) \leq \log|S|$
$h(XZ) \leq \log|T|$
$h(X), h(Y), h(Z)$
frac. edge packing

From the optimal, modular dual $\boldsymbol{h}^*$, we construct a worst-case instance R^*, S^*, T^*:

$R^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Y)} \rfloor \qquad S^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(Y)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor \qquad T^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$

$$|Q^*| = \lfloor 2^{h^*(X)} \rfloor \cdot \lfloor 2^{h^*(Y)} \rfloor \cdot \lfloor 2^{h^*(Z)} \rfloor \geq \frac{1}{8} 2^{h^*(XYZ)} = \frac{1}{8} 2^{w_1^* \log|R| + w_2^* \log|S| + w_3^* \log|T|} = \frac{1}{8} 2^{AGM(Q)}$$

Proof of the AGM Lower Bound

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\quad\quad AGM(Q) = |R|^{w_1} \cdot |S|^{w_2} \cdot |T|^{w_3}$

Primal program:
Minimize $w_1 \log|R| + w_2 \log|S| + w_3 \log|T|$
Where $(w_1, w_2, w_3) \in \mathbb{R}_+^3$
$M_n \models w_1 h(XY) + w_2 h(YZ) + w_3 h(XZ) \geq h(XYZ)$
w_1, w_2, w_3 frac. edge cover

Dual program:
Maximize $h(XYZ)$
Where $\boldsymbol{h} \in M_n$

$$h(XY) \leq \log|R|$$
$$h(YZ) \leq \log|S|$$
$$h(XZ) \leq \log|T|$$

$h(X), h(Y), h(Z)$
frac. edge packing

From the optimal, modular dual $\boldsymbol{h}^*$, we construct a worst-case instance R^*, S^*, T^*:

$$R^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Y)} \rfloor \quad\quad S^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(Y)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor \quad\quad T^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$$

$$|Q^*| = \lfloor 2^{h^*(X)} \rfloor \cdot \lfloor 2^{h^*(Y)} \rfloor \cdot \lfloor 2^{h^*(Z)} \rfloor \geq \frac{1}{8} 2^{h^*(XYZ)} = \frac{1}{8} 2^{w_1^* \log|R| + w_2^* \log|S| + w_3^* \log|T|} = \frac{1}{8} 2^{AGM(Q)}$$

Proof of the AGM Lower Bound

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\qquad$ $AGM(Q) = |R|^{w_1} \cdot |S|^{w_2} \cdot |T|^{w_3}$

Primal program:
Minimize $w_1 \log|R| + w_2 \log|S| + w_3 \log|T|$
Where $(w_1, w_2, w_3) \in \mathbb{R}_+^3$
$M_n \models w_1 h(XY) + w_2 h(YZ) + w_3 h(XZ) \geq h(XYZ)$
w_1, w_2, w_3 frac. edge cover

Dual program:
Maximize $h(XYZ)$
Where $\boldsymbol{h} \in M_n$

$$h(XY) \leq \log|R|$$
$$h(YZ) \leq \log|S|$$
$$h(XZ) \leq \log|T|$$

$h(X), h(Y), h(Z)$
frac. edge packing

From the optimal, modular dual $\boldsymbol{h}^*$, we construct a worst-case instance R^*, S^*, T^*:

$R^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Y)} \rfloor$ $\qquad$ $S^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(Y)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$ $\qquad$ $T^* \stackrel{\text{def}}{=} \lfloor 2^{h^*(X)} \rfloor \times \lfloor 2^{h^*(Z)} \rfloor$

$$|Q^*| = \lfloor 2^{h^*(X)} \rfloor \cdot \lfloor 2^{h^*(Y)} \rfloor \cdot \lfloor 2^{h^*(Z)} \rfloor \geq \frac{1}{8} 2^{h^*(XYZ)} = \frac{1}{8} 2^{w_1^* \log|R| + w_2^* \log|S| + w_3^* \log|T|} = \frac{1}{8} 2^{AGM(Q)}$$

Brief History

- Pippenger [Pippenger, 1986]: inequalities are "*laws of information theory*". Do Shannon inequalities form the complete laws?

- The breakthrough: a non-Shannon inequality with $n = 4$ variables [Zhang and Yeung, 1998].

- There exists infinitely many non-equivalent non-Shannon inequalities with $n = 4$ variable [Matús, 2007]. Hence[1], $\bar{\Gamma}_n^*$ is not a polytope.

- The characterization of $\bar{\Gamma}_n^*$ is open to date.

[1] Γ_n^* is not a cone, nor convex, but its topological closure $\bar{\Gamma}_n^*$ is.

A Non-Shannon Inequality [Zhang and Yeung, 1998]

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

A Non-Shannon Inequality [Zhang and Yeung, 1998]

$$I(X;Y) \le I(X;Y|A) + I(X;Y|B) + I(A;B) + I(X;Y|A) + I(A;Y|X) + I(A;X|Y) \quad (1)$$

Theorem

$\Gamma_n^* \models (1)$, *but* $\Gamma_n \not\models (1)$

A Non-Shannon Inequality [Zhang and Yeung, 1998]

$$I(X;Y) \leq I(X;Y|A) + I(X;Y|B) + I(A;B)$$
$$+I(X;Y|A) + I(A;Y|X) + I(A;X|Y) \qquad (1)$$

Theorem

$\Gamma_n^* \models (1)$, *but* $\Gamma_n \not\models (1)$

Proof:

$$\Gamma_n \models I(X;Y) \leq I(X;Y|A) + I(X;Y|B) + I(A;B)$$
$$+I(X;Y|A') + I(A';Y|X) + I(A';X|Y) + 3I(A';AB|XY)$$

A Non-Shannon Inequality [Zhang and Yeung, 1998]

$$I(X;Y) \leq I(X;Y|A) + I(X;Y|B) + I(A;B) + I(X;Y|A) + I(A;Y|X) + I(A;X|Y) \qquad (1)$$

Theorem

$\Gamma_n^* \models (1)$, *but* $\Gamma_n \not\models (1)$

Proof:

$$\Gamma_n \models I(X;Y) \leq I(X;Y|A) + I(X;Y|B) + I(A;B) + I(X;Y|A') + I(A';Y|X) + I(A';X|Y) + 3I(A';AB|XY)$$

Copy Lemma Define r.v. A': $A' \perp AB|XY$, $p(A'|XY) \stackrel{\text{def}}{=} p(A|XY)$. Then:

A Non-Shannon Inequality [Zhang and Yeung, 1998]

$$I(X;Y) \leq I(X;Y|A) + I(X;Y|B) + I(A;B) + I(X;Y|A) + I(A;Y|X) + I(A;X|Y) \qquad (1)$$

Theorem

$\Gamma_n^* \models (1)$, *but* $\Gamma_n \not\models (1)$

Proof:

$$\Gamma_n \models I(X;Y) \leq I(X;Y|A) + I(X;Y|B) + I(A;B) + I(X;Y|A') + I(A';Y|X) + I(A';X|Y) + 3I(A';AB|XY)$$

Copy Lemma Define r.v. A': $A' \perp AB|XY$, $p(A'|XY) \stackrel{\text{def}}{=} p(A|XY)$. Then:

$$\Gamma_n^* \models I(X;Y|A) + I(A;Y|X) + I(A;X|Y) = I(X;Y|A') + I(A';Y|X) + I(A';X|Y) + 3I(A';AB|XY)$$

A Non-Shannon Inequality [Zhang and Yeung, 1998]

$$I(X;Y) \leq I(X;Y|A) + I(X;Y|B) + I(A;B) + I(X;Y|A) + I(A;Y|X) + I(A;X|Y) \quad (1)$$

Theorem

$\Gamma_n^* \models (1)$, *but* $\Gamma_n \not\models (1)$

Proof:

$$\Gamma_n \models I(X;Y) \leq I(X;Y|A) + I(X;Y|B) + I(A;B) + I(X;Y|A') + I(A';Y|X) + I(A';X|Y) + 3I(A';AB|XY)$$

Copy Lemma Define r.v. A': $A' \perp AB|XY$, $p(A'|XY) \stackrel{\text{def}}{=} p(A|XY)$. Then:

$$\Gamma_n^* \models I(X;Y|A) + I(A;Y|X) + I(A;X|Y) = I(X;Y|A') + I(A';Y|X) + I(A';X|Y) + 3I(A';AB|XY)$$

This implies $\Gamma_n^* \models (1)$.

$\Gamma_n \not\models (1)$: see paper.

Discussion

- Non-Shannon inequalities: $\Gamma_n^* \models (\cdots)$ yet $\Gamma_n \not\models (\cdots)$.
  - Decidability of $\Gamma_n^* \models (\cdots)$ is open.
  - Lower bound holds only asymptotically: $\boldsymbol{h}^* \in \Gamma_n^*$ to a worst-case instance uses the *group characterization* [Chan and Yeung, 2002].
- Shannon inequalities $\Gamma_n \models (\cdots)$ decidable in EXPTIME. But:
  - The order of submodularity steps matters.
  - The bound is not tight in general: $\boldsymbol{h}^* \in \Gamma_n$ does not always correspond to a relation instance.

Next: we can recover elegant properties for "simple" inequalities.

"Simple" Inequalities

The **step function** at $\boldsymbol{V} \subseteq \boldsymbol{X}$ is:

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

"Simple" Inequalities

The **step function** at $\boldsymbol{V} \subseteq \boldsymbol{X}$ is:

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

$N_n \stackrel{\text{def}}{=}$ positive, linear combinations of step functions.

$$\boxed{M_n \subset N_n \subset \Gamma_n^* \subset \Gamma_n (\subset \mathbb{R}_+^{2^n})}$$

"Simple" Inequalities

The **step function** at $\boldsymbol{V} \subseteq \boldsymbol{X}$ is:

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

$N_n \stackrel{\text{def}}{=}$ positive, linear combinations of step functions.

$$\boxed{M_n \subset N_n \subset \Gamma_n^* \subset \Gamma_n (\subset \mathbb{R}_+^{2^n})}$$

An inequality $\boxed{\sum_i w_i h(\boldsymbol{V}_i | \boldsymbol{U}_i) \geq h(\boldsymbol{X})}$ is "simple" if $|\boldsymbol{U}_i| \leq 1$.

Theorem

$\Gamma_n \models (\cdots)$ *iff* $\Gamma_n^* \models (\cdots)$ *iff* $N_n \models (\cdots)$.

"Simple" Inequalities

The **step function** at $\boldsymbol{V} \subseteq \boldsymbol{X}$ is:

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

$N_n \stackrel{\text{def}}{=}$ positive, linear combinations of step functions.

$$\boxed{M_n \subset N_n \subset \Gamma_n^* \subset \Gamma_n (\subset \mathbb{R}_+^{2^n})}$$

An inequality $\boxed{\sum_i w_i h(\boldsymbol{V}_i | \boldsymbol{U}_i) \geq h(\boldsymbol{X})}$ is "simple" if $|\boldsymbol{U}_i| \leq 1$.

Theorem

$\Gamma_n \models (\cdots)$ *iff* $\Gamma_n^* \models (\cdots)$ *iff* $N_n \models (\cdots)$.

If all degrees are "simple", then Max-Degree bound is computable, tight.