CS294-248 Special Topics in Database Theory
Unit 4: AGM Bound, WCOJ

Dan Suciu

University of Washington

Outline

- Today: the AGM bound. This is a mathematical formula that gives us $AGM(Q, \boldsymbol{D}) \stackrel{\text{def}}{=} \max_{\boldsymbol{D} \models \text{statistics}} |Q(\boldsymbol{D})|$.

- Thursday: Worst Case Optimal Join. This is an algorithm that computes $Q(\boldsymbol{D})$ in time $\tilde{O}(AGM(Q, \boldsymbol{D}))$.

Background on Cardinality Estimation

Cardinality Estimation 101 (1/3)

Given:

- Statistics on the input relations $R_1, R_2, \ldots$
- A full conjunctive query Q

"Estimate":

- The size $|Q(\boldsymbol{D})|$.

Numerous applications: query optimization, memory provisioning, data partitioning.

Cardinality Estimation 101 (2/3)

Bottom-up on the query plan:

- Selection $\sigma_p(R)$: assume independence:

$$\boxed{|\sigma_p(R)| \approx \theta_p \cdot |R|} \qquad \boxed{\theta_{p_1 \wedge p_2} \approx \theta_{p_1} \cdot \theta_{p_2}}$$

Histograms, multidimensional histograms.

- Join $J(A,B,C) = R(A,B) \wedge S(B,C)$: assume preservation of values
  - $|J| \approx |R| \cdot \text{avg}(\text{deg}_S(C|B)) = \frac{|R|\cdot|S|}{|\text{Dom}(S.B)|}$.
  - $|J| \approx |S| \cdot \text{avg}(\text{deg}_R(A|B)) = \frac{|R|\cdot|S|}{|\text{Dom}(R.B)|}$.
  - Heuristic: take the *minimum*: $\boxed{|J| \approx \frac{|R|\cdot|S|}{\max(|\text{Dom}(R.B)|\cdot|\text{Dom}(S.B)|)}}$

Cardinality Estimation 101 (2/3)

Bottom-up on the query plan:

- Selection $\sigma_p(R)$: assume independence:

$$\boxed{|\sigma_p(R)| \approx \theta_p \cdot |R|} \qquad \boxed{\theta_{p_1 \wedge p_2} \approx \theta_{p_1} \cdot \theta_{p_2}}$$

  Histograms, multidimensional histograms.

- Join $J(A, B, C) = R(A, B) \wedge S(B, C)$: assume preservation of values
  - $|J| \approx |R| \cdot \text{avg}(\text{deg}_S(C|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(S.B)|}$.
  - $|J| \approx |S| \cdot \text{avg}(\text{deg}_R(A|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(R.B)|}$.
  - Heuristic: take the *minimum*: $\boxed{|J| \approx \frac{|R| \cdot |S|}{\max(|\text{Dom}(R.B)| \cdot |\text{Dom}(S.B)|)}}$

Cardinality Estimation 101 (2/3)

Bottom-up on the query plan:

- Selection $\sigma_p(R)$: assume independence:

$$\boxed{|\sigma_p(R)| \approx \theta_p \cdot |R|} \qquad \boxed{\theta_{p_1 \wedge p_2} \approx \theta_{p_1} \cdot \theta_{p_2}}$$

  Histograms, multidimensional histograms.

- Join $J(A, B, C) = R(A, B) \wedge S(B, C)$: assume preservation of values
  - $|J| \approx |R| \cdot \text{avg}(\text{deg}_S(C|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(S.B)|}$.
  - $|J| \approx |S| \cdot \text{avg}(\text{deg}_R(A|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(R.B)|}$.
  - Heuristic: take the *minimum*: $\boxed{|J| \approx \frac{|R| \cdot |S|}{\max(|\text{Dom}(R.B)| \cdot |\text{Dom}(S.B)|)}}$

Cardinality Estimation 101 (2/3)

Bottom-up on the query plan:

- Selection $\sigma_p(R)$: assume independence:

$$\boxed{|\sigma_p(R)| \approx \theta_p \cdot |R|} \qquad \boxed{\theta_{p_1 \wedge p_2} \approx \theta_{p_1} \cdot \theta_{p_2}}$$

  Histograms, multidimensional histograms.

- Join $J(A, B, C) = R(A, B) \wedge S(B, C)$: assume preservation of values
  - $|J| \approx |R| \cdot \text{avg}(\text{deg}_S(C|B)) = \frac{|R|\cdot|S|}{|\text{Dom}(S.B)|}$.
  - $|J| \approx |S| \cdot \text{avg}(\text{deg}_R(A|B)) = \frac{|R|\cdot|S|}{|\text{Dom}(R.B)|}$.
  - Heuristic: take the *minimum*: $\boxed{|J| \approx \frac{|R|\cdot|S|}{\max(|\text{Dom}(R.B)|\cdot|\text{Dom}(S.B)|)}}$

Cardinality Estimation 101 (3/3)

- Notoriously hard to estimate cardinality of complex queries.
- No rigorous definition of the *estimate*: there is no probability space.
- How do we combine multiple sources of information?
  - We had two formulas for the join, why choose min?
  - Given $R(A, B, C)$ and histograms on A, B, C, AB, AC, how do we estimate $|\sigma_{A=2,B=4,C=6}(R)|$?

Upper Bound on the Output of a Query

The Output Bound Problem

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

Compute an upper bound B: $$|Q(\boldsymbol{D})| \leq B$$

Challenge: make B tight.

Simple Examples

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

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

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: $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N$
- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
  Notice the role of an edge cover

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: $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N$
- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
  Notice the role of an edge cover
- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = ?$

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: $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N$
- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
  Notice the role of an edge cover
- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$

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: $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N$

- $Q(X,Y,Z,U) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,U)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^2$
  Notice the role of an edge cover

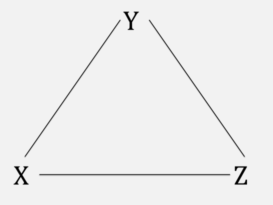
- $Q(X,Y,Z) = R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$. $\max_{\boldsymbol{D}} |Q(\boldsymbol{D})| = N^{\frac{3}{2}}$
  Here we use a fractional edge cover

AGM Bound: The Statement

Fractional Edge Covers

Query Q to hypegraph $G = (V, E)$.

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

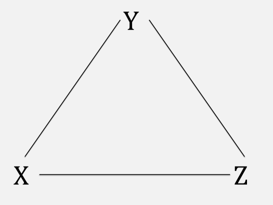


Fractional Edge Covers

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

Definition

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



Fractional Edge Covers

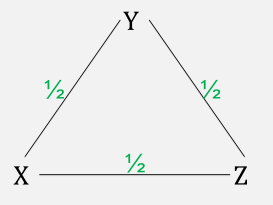
Query Q to hypegraph $G = (V, E)$.

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

Definition

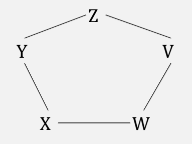
A *fractional edge cover* is $\boldsymbol{w} = (w_e)_{e \in E}$, $w_e \geq 0$:

$$\forall x \in V, \sum_{e \in E : x \in e} w_e \geq 1.$$

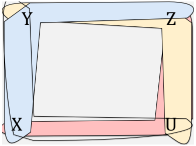


Examples

What are fractional edge covers?

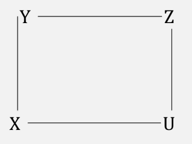


5-cycle



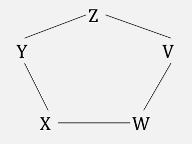
Loomis-Whitney:

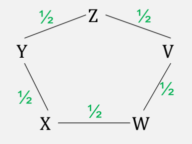
$$R(X, Y, Z) \wedge S(Y, Z, U)$$
$$\wedge T(Z, U, X) \wedge K(U, X, Y)$$



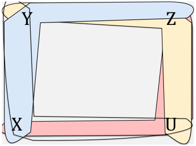
Examples

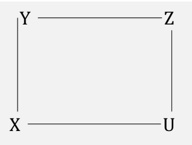
What are fractional edge covers?





5-cycle



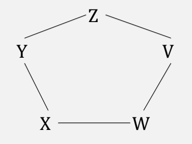


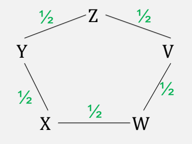
Loomis-Whitney:

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

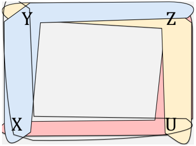
Examples

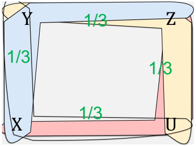
What are fractional edge covers?





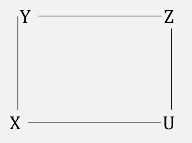
5-cycle





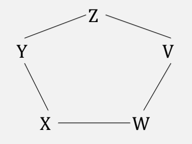
Loomis-Whitney:

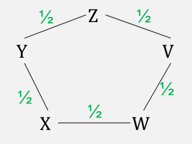
$$R(X, Y, Z) \wedge S(Y, Z, U)$$
$$\wedge T(Z, U, X) \wedge K(U, X, Y)$$



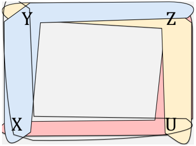
Examples

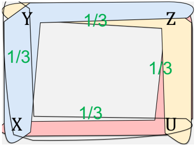
What are fractional edge covers?





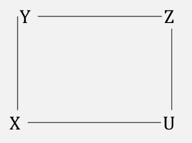
5-cycle

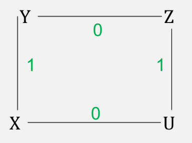




Loomis-Whitney:

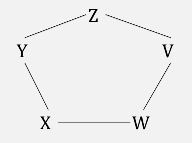
$$R(X,Y,Z) \wedge S(Y,Z,U)$$
$$\wedge T(Z,U,X) \wedge K(U,X,Y)$$

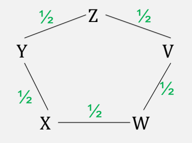




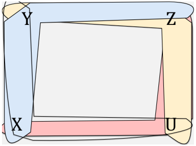
Examples

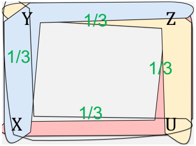
What are fractional edge covers?





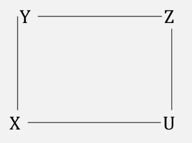
5-cycle

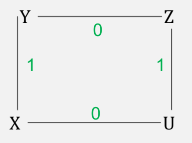


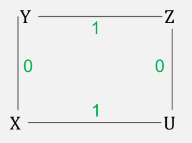


Loomis-Whitney:

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

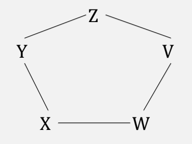


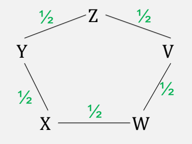




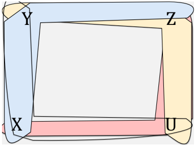
Examples

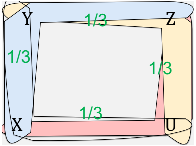
What are fractional edge covers?





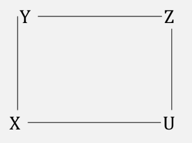
5-cycle

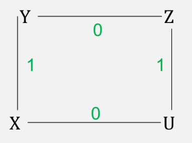


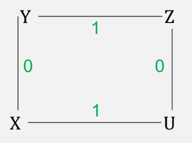


Loomis-Whitney:

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



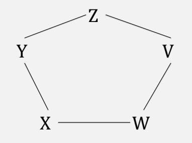


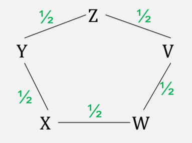


$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ is a convex combination of $(1, 0, 1, 0)$ and $(0, 1, 0, 1)$.

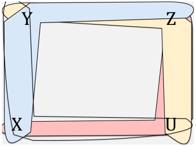
Examples

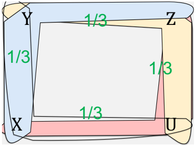
What are fractional edge covers?





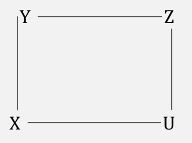
5-cycle

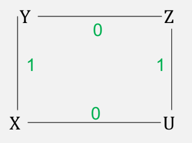


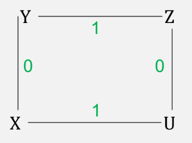


Loomis-Whitney:

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







$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ is a convex combination of $(1, 0, 1, 0)$ and $(0, 1, 0, 1)$.

Vertex of the edge covering polytope: no convex combination of others.

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".*

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

$$AGM(Q) = \min \begin{pmatrix} (|R| \cdot |S| \cdot |T|)^{1/2} \\ |R| \cdot |S| \\ |R| \cdot |T| \\ |S| \cdot |T| \end{pmatrix}$$

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".*

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

$$AGM(Q) = \min \begin{pmatrix} (|R| \cdot |S| \cdot |T|)^{1/2} \\ |R| \cdot |S| \\ |R| \cdot |T| \\ |S| \cdot |T| \end{pmatrix}$$

Minimum over vertices of the edge-covering polytope. WHY?

Proof Outline

- Proof of the upper bound: information inequalities (a.k.a. entropic inequalities).

- Proof of the lower bound: construct a worst-case database instance by using strong duality of linear optimization.

Proof of the Upper Bound

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 $p : D \rightarrow [0, 1]$. 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: $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

$N \stackrel{\text{def}}{=} |D|$: $0 \leq h(X) \leq \log N$ $h(X) = \log N$ 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
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: $h(X) \stackrel{\text{def}}{=} -\sum_{x \in D} p(x) \log p(x)$

$N \stackrel{\text{def}}{=} |D|$: $0 \le h(X) \le \log N$ $h(X) = \log N$ 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	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

$h(\emptyset) = 0$

Shannon Inequalities

Basic Shannon Inequalities

$$h(\emptyset) = 0$$

$$h(\boldsymbol{U} \cup \boldsymbol{V}) \geq h(\boldsymbol{U}) \qquad \text{Monotonicity}$$

$$h(\boldsymbol{U}) + h(\boldsymbol{V}) \geq h(\boldsymbol{U} \cup \boldsymbol{V}) + h(\boldsymbol{U} \cap \boldsymbol{V}) \qquad \text{Submodularity}$$

A Shannon inequality is a consequence of these inequalities.

A Shannon Inequality

Example

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

A Shannon Inequality

Example

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

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

A Shannon Inequality

Example

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

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

A Shannon Inequality

Example

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

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

A Shannon Inequality

Example

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

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

A Shannon Inequality

Example

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

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

A Shannon Inequality

Example

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

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

A Shannon Inequality

Example

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

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

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

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

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

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: $\boxed{|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}}$

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.

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

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.

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

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

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.

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

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

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: $\boxed{|Q| \leq |R_1|^{w_1} \cdots |R_m|^{w_m}}$

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.

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

For a general query $Q(\boldsymbol{X}) = R_1(\boldsymbol{Y}_1) \wedge \cdots \wedge R_m(\boldsymbol{Y}_m)$:

If $\boxed{\sum_j w_j h(\boldsymbol{Y}_j) \geq h(\boldsymbol{X})}$ then $\boxed{|R_1|^{w_1} \cdots |R_m|^{w_m} \geq |Q|}$

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

Theorem (Generalized Shearer's Inequality)

If $\boldsymbol{w}$ is a frac. edge cover, then $\boxed{w_1 h(\boldsymbol{Y}_1) + \cdots + w_m h(\boldsymbol{Y}_m) \geq h(\boldsymbol{X})}$.

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

Theorem (Generalized Shearer's Inequality)

If $\boldsymbol{w}$ is a frac. edge cover, then $\boxed{w_1 h(\boldsymbol{Y}_1) + \cdots + w_m h(\boldsymbol{Y}_m) \geq h(\boldsymbol{X})}$.

Proof following [Balister and Bollobás, 2012]. Convert to integers:

$$\boxed{k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})}$$

Each variable is "covered $\geq k_0$ times".

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

Theorem (Generalized Shearer's Inequality)

If $\boldsymbol{w}$ is a frac. edge cover, then $\boxed{w_1 h(\boldsymbol{Y}_1) + \cdots + w_m h(\boldsymbol{Y}_m) \geq h(\boldsymbol{X})}$.

Proof following [Balister and Bollobás, 2012]. Convert to integers:

$$\boxed{k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})}$$

Each variable is "covered $\geq k_0$ times".

Repeatedly rewrite $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j) \to h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

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

Theorem (Generalized Shearer's Inequality)

If $\boldsymbol{w}$ *is a frac. edge cover, then* $\boxed{w_1 h(\boldsymbol{Y}_1) + \cdots + w_m h(\boldsymbol{Y}_m) \geq h(\boldsymbol{X})}$.

Proof following [Balister and Bollobás, 2012]. Convert to integers:

$$\boxed{k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})}$$

Each variable is "covered $\geq k_0$ times".

Repeatedly rewrite $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j) \to h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

- Every variable remains covered $\geq k_0$ times.

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

Theorem (Generalized Shearer's Inequality)

If $\boldsymbol{w}$ *is a frac. edge cover, then* $\boxed{w_1 h(\boldsymbol{Y}_1) + \cdots + w_m h(\boldsymbol{Y}_m) \geq h(\boldsymbol{X})}$.

Proof following [Balister and Bollobás, 2012]. Convert to integers:

$$\boxed{k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})}$$

Each variable is "covered $\geq k_0$ times".

Repeatedly rewrite $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j) \to h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

- Every variable remains covered $\geq k_0$ times.
- $\sum_\ell |\boldsymbol{Y}_\ell|^2$ strictly increases (homework!).

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

Theorem (Generalized Shearer's Inequality)

If $\boldsymbol{w}$ is a frac. edge cover, then $\boxed{w_1 h(\boldsymbol{Y}_1) + \cdots + w_m h(\boldsymbol{Y}_m) \geq h(\boldsymbol{X})}$.

Proof following [Balister and Bollobás, 2012]. Convert to integers:

$$\boxed{k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})}$$

Each variable is "covered $\geq k_0$ times".

Repeatedly rewrite $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j) \to h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

- Every variable remains covered $\geq k_0$ times.
- $\sum_\ell |\boldsymbol{Y}_\ell|^2$ strictly increases (homework!).

When do we stop?

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

Theorem (Generalized Shearer's Inequality)

If $\boldsymbol{w}$ *is a frac. edge cover, then* $\boxed{w_1 h(\boldsymbol{Y}_1) + \cdots + w_m h(\boldsymbol{Y}_m) \geq h(\boldsymbol{X})}$.

Proof following [Balister and Bollobás, 2012]. Convert to integers:

$$\boxed{k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})}$$

Each variable is "covered $\geq k_0$ times".

Repeatedly rewrite $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j) \to h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

- Every variable remains covered $\geq k_0$ times.
- $\sum_\ell |\boldsymbol{Y}_\ell|^2$ strictly increases (homework!). When do we stop?
- We stop when $\boldsymbol{Y}_1 \supset \boldsymbol{Y}_2 \supset \cdots$

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

Theorem (Generalized Shearer's Inequality)

If $\boldsymbol{w}$ is a frac. edge cover, then $\boxed{w_1 h(\boldsymbol{Y}_1) + \cdots + w_m h(\boldsymbol{Y}_m) \geq h(\boldsymbol{X})}$.

Proof following [Balister and Bollobás, 2012]. Convert to integers:

$$\boxed{k_1 h(\boldsymbol{Y}_1) + \cdots + k_m h(\boldsymbol{Y}_m) \geq k_0 h(\boldsymbol{X})}$$

Each variable is "covered $\geq k_0$ times".

Repeatedly rewrite $h(\boldsymbol{Y}_i) + h(\boldsymbol{Y}_j) \to h(\boldsymbol{Y}_i \cup \boldsymbol{Y}_j) + h(\boldsymbol{Y}_i \cap \boldsymbol{Y}_j)$

- Every variable remains covered $\geq k_0$ times.
- $\sum_\ell |\boldsymbol{Y}_\ell|^2$ strictly increases (homework!). When do we stop?
- We stop when $\boldsymbol{Y}_1 \supset \boldsymbol{Y}_2 \supset \cdots$
- Thus, $\boldsymbol{Y}_1 = \boldsymbol{X}$ and $k_1 \geq k_0$. $\boxed{\cdots \geq k_1 h(\boldsymbol{Y}_1) \geq k_0 h(\boldsymbol{X})}$

This completes the proof of the Upper AGM Bound.

Discussion

- Shearer's inequality: apply submodularity repeatedly, in any order!
- Shearer inequalities correspond 1-1 to fractional edge covers.
- Any inequality is an upper bound on $|Q|$: $AGM(Q)$ is the smallest.
- How tight is $AGM(Q)$ upper bound? Next

Proof of the Lower Bound

Proof of the AGM Lower Bound

By example:

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X) \qquad AGM(Q) = \min_{\boldsymbol{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$

Proof of the AGM Lower Bound

By example:

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\quad AGM(Q) = \min_{\boldsymbol{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$

Primal program:

Minimize

$w_R \log |R| + w_S \log |S| + w_T \log |T|$

where $\boldsymbol{w}$ is frac. edge cover:

$$\begin{array}{lllll} X: & w_R + & & w_T & \geq 1 \\ Y: & w_R + & w_S & & \geq 1 \\ Z: & & w_S + & w_T & \geq 1 \end{array}$$

Proof of the AGM Lower Bound

By example:

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\quad AGM(Q) = \min_{\boldsymbol{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$

Primal program:

Minimize

$w_R \log|R| + w_S \log|S| + w_T \log|T|$

where $\boldsymbol{w}$ is frac. edge cover:

$$\begin{array}{llllll} X: & w_R+ & & w_T & \geq 1 \\ Y: & w_R+ & w_S & & \geq 1 \\ Z: & & w_S+ & w_T & \geq 1 \end{array}$$

Dual program:

Maximize

$$v_X + v_Y + v_Z$$

where $\boldsymbol{v}$ is "frac. vertex packing":

$$\begin{array}{llllll} R: & v_X+ & v_Y & & \leq \log|R| \\ S: & & v_Y+ & v_Z & \leq \log|S| \\ T: & v_X+ & & v_Z & \leq \log|T| \end{array}$$

Proof of the AGM Lower Bound

By example:

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\quad$ $AGM(Q) = \min_{\boldsymbol{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$

Primal program:

Minimize

$w_R \log|R| + w_S \log|S| + w_T \log|T|$

where $\boldsymbol{w}$ is frac. edge cover:

$$\begin{array}{llllll} X: & w_R+ & & w_T & \geq 1 \\ Y: & w_R+ & w_S & & \geq 1 \\ Z: & & w_S+ & w_T & \geq 1 \end{array}$$

Dual program:

Maximize

$$v_X + v_Y + v_Z$$

where $\boldsymbol{v}$ is "frac. vertex packing":

$$\begin{array}{lllll} R: & v_X+ & v_Y & & \leq \log|R| \\ S: & & v_Y+ & v_Z & \leq \log|S| \\ T: & v_X+ & & v_Z & \leq \log|T| \end{array}$$

Take optimum $\boldsymbol{v}$, define: $\text{Dom}(X) \stackrel{\text{def}}{=} [\lfloor 2^{v_X} \rfloor]$, $\text{Dom}(Y) \stackrel{\text{def}}{=} [\lfloor 2^{v_Y} \rfloor]$, $\text{Dom}(Z) \stackrel{\text{def}}{=} [\lfloor 2^{v_Z} \rfloor]$.

Proof of the AGM Lower Bound

By example:

$R(X,Y) \wedge S(Y,Z) \wedge T(Z,X)$ $\qquad AGM(Q) = \min_{\boldsymbol{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$

Primal program:
Minimize
$w_R \log|R| + w_S \log|S| + w_T \log|T|$
where $\boldsymbol{w}$ is frac. edge cover:

$$\begin{array}{llllll} X: & w_R+ & & w_T & \geq 1 \\ Y: & w_R+ & w_S & & \geq 1 \\ Z: & & w_S+ & w_T & \geq 1 \end{array}$$

Dual program:
Maximize

$$v_X + v_Y + v_Z$$

where $\boldsymbol{v}$ is "frac. vertex packing":

$$\begin{array}{lllll} R: & v_X+ & v_Y & & \leq \log|R| \\ S: & & v_Y+ & v_Z & \leq \log|S| \\ T: & v_X+ & & v_Z & \leq \log|T| \end{array}$$

Take optimum $\boldsymbol{v}$, define: $\text{Dom}(X) \stackrel{\text{def}}{=} [\lfloor 2^{v_X} \rfloor]$, $\text{Dom}(Y) \stackrel{\text{def}}{=} [\lfloor 2^{v_Y} \rfloor]$, $\text{Dom}(Z) \stackrel{\text{def}}{=} [\lfloor 2^{v_Z} \rfloor]$.

Worst-case instance (cartesian products): $R^* \stackrel{\text{def}}{=} \text{Dom}(X) \times \text{Dom}(Y)$, S^*, $T^* \stackrel{\text{def}}{=} \cdots$

Proof of the AGM Lower Bound

By example:

$R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$ $\qquad AGM(Q) = \min_{\boldsymbol{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$

Primal program:
Minimize
$w_R \log |R| + w_S \log |S| + w_T \log |T|$
where $\boldsymbol{w}$ is frac. edge cover:

$$\begin{array}{llllll} X: & w_R+ & & w_T & \geq 1 \\ Y: & w_R+ & w_S & & \geq 1 \\ Z: & & w_S+ & w_T & \geq 1 \end{array}$$

Dual program:
Maximize
$$v_X + v_Y + v_Z$$
where $\boldsymbol{v}$ is "frac. vertex packing":

$$\begin{array}{llllll} R: & v_X+ & v_Y & & \leq \log |R| \\ S: & & v_Y+ & v_Z & \leq \log |S| \\ T: & v_X+ & & v_Z & \leq \log |T| \end{array}$$

Take optimum $\boldsymbol{v}$, define: $\text{Dom}(X) \stackrel{\text{def}}{=} [\lfloor 2^{v_X} \rfloor]$, $\text{Dom}(Y) \stackrel{\text{def}}{=} [\lfloor 2^{v_Y} \rfloor]$, $\text{Dom}(Z) \stackrel{\text{def}}{=} [\lfloor 2^{v_Z} \rfloor]$.

Worst-case instance (cartesian products): $R^* \stackrel{\text{def}}{=} \text{Dom}(X) \times \text{Dom}(Y)$, $S^*, T^* \stackrel{\text{def}}{=} \cdots$

$$|Q^*| = \lfloor 2^{v_X} \rfloor \cdot \lfloor 2^{v_Y} \rfloor \cdot \lfloor 2^{v_Z} \rfloor \geq \frac{1}{8} 2^{v_X + v_Y + v_Z}$$

Proof of the AGM Lower Bound

By example:

$R(X, Y) \wedge S(Y, Z) \wedge T(Z, X) \qquad AGM(Q) = \min_{\boldsymbol{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$

Primal program:
Minimize
$w_R \log |R| + w_S \log |S| + w_T \log |T|$
where $\boldsymbol{w}$ is frac. edge cover:

$$\begin{array}{llllll} X: & w_R + & & w_T & \geq 1 \\ Y: & w_R + & w_S & & \geq 1 \\ Z: & & w_S + & w_T & \geq 1 \end{array}$$

Dual program:
Maximize
$$v_X + v_Y + v_Z$$
where $\boldsymbol{v}$ is "frac. vertex packing":

$$\begin{array}{lllll} R: & v_X + & v_Y & & \leq \log |R| \\ S: & & v_Y + & v_Z & \leq \log |S| \\ T: & v_X + & & v_Z & \leq \log |T| \end{array}$$

Take optimum $\boldsymbol{v}$, define: $\text{Dom}(X) \stackrel{\text{def}}{=} [\lfloor 2^{v_X} \rfloor]$, $\text{Dom}(Y) \stackrel{\text{def}}{=} [\lfloor 2^{v_Y} \rfloor]$, $\text{Dom}(Z) \stackrel{\text{def}}{=} [\lfloor 2^{v_Z} \rfloor]$.

Worst-case instance (cartesian products): $R^* \stackrel{\text{def}}{=} \text{Dom}(X) \times \text{Dom}(Y)$, $S^*, T^* \stackrel{\text{def}}{=} \cdots$

$$|Q^*| = \lfloor 2^{v_X} \rfloor \cdot \lfloor 2^{v_Y} \rfloor \cdot \lfloor 2^{v_Z} \rfloor \geq \frac{1}{8} 2^{v_X + v_Y + v_Z} = \frac{1}{8} 2^{w_1^* \log |R| + w_2^* \log |S| + w_3^* \log |T|} = \frac{1}{8} 2^{AGM(Q)}$$

Special Case: $|R| = |S| = \cdots = N$

Definition

Fix a hypergraph (V, E); $(v_X)_{X \in V} \in \mathbb{R}_+^{|V|}$ is a fractional vertex packing if:

$$\forall \boldsymbol{Y} \in E :, \boxed{\sum_{X \in \boldsymbol{Y}} v_X \leq 1}$$

Special Case: $|R| = |S| = \cdots = N$

Definition

Fix a hypergraph (V, E); $(v_X)_{X \in V} \in \mathbb{R}_+^{|V|}$ is a fractional vertex packing if:

$$\forall \boldsymbol{Y} \in E :, \boxed{\sum_{X \in \boldsymbol{Y}} v_X \leq 1}$$

When $|R| = |S| = \cdots = N$, then replace

$$\boxed{\begin{aligned} v_R + v_S &\leq \log N \\ v_R + v_T &\leq \log N \\ &\ldots \end{aligned}} \quad \text{with} \quad \boxed{\begin{aligned} v_R + v_S &\leq 1 \\ v_R + v_T &\leq 1 \\ &\ldots \end{aligned}} \quad \text{times } \log N.$$

Special Case: $|R| = |S| = \cdots = N$

Definition

Fix a hypergraph (V, E); $(v_X)_{X \in V} \in \mathbb{R}_+^{|V|}$ is a fractional vertex packing if:

$$\forall \boldsymbol{Y} \in E :, \boxed{\sum_{X \in \boldsymbol{Y}} v_X \leq 1}$$

When $|R| = |S| = \cdots = N$, then replace

$$\boxed{\begin{aligned} v_R + v_S &\leq \log N \\ v_R + v_T &\leq \log N \\ &\ldots \end{aligned}} \quad \text{with} \quad \boxed{\begin{aligned} v_R + v_S &\leq 1 \\ v_R + v_T &\leq 1 \\ &\ldots \end{aligned}} \quad \text{times } \log N.$$

Then: $R = [N^{v_X}] \times [N^{v_Y}]$, $S = [N^{v_Y}] \times [N^{v_Z}]$, $T = [N^{v_X}] \times [N^{v_Z}]$.

$$Q = [N^{v_X}] \times [N^{v_Y}] \times [N^{v_Z}]$$

Examples

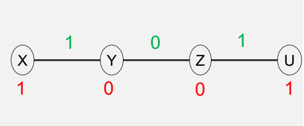
$$|R| = |S| = \cdots = N$$

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

Examples

$$|R| = |S| = \cdots = N$$

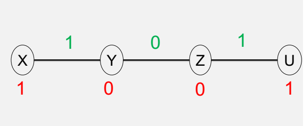
$$R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$$



Examples

$|R| = |S| = \cdots = N$

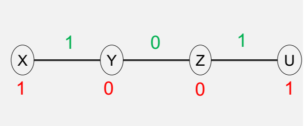
$R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$
$R = [N] \times [1],\ S = [1] \times [1],\ T = [1] \times [N].$



Examples

$|R| = |S| = \cdots = N$

$R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$
$R = [N] \times [1],\ S = [1] \times [1],\ T = [1] \times [N].$

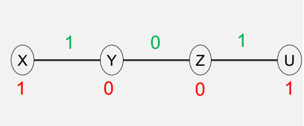


$R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge K(U, V)$

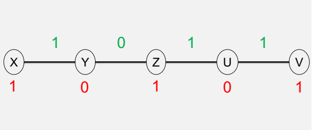
Examples

$|R| = |S| = \cdots = N$

$R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$
$R = [N] \times [1],\ S = [1] \times [1],\ T = [1] \times [N].$



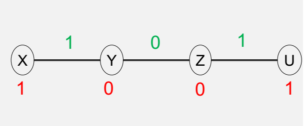
$R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge K(U, V)$



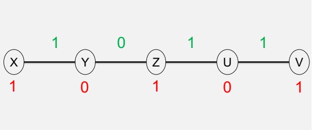
Examples

$$|R| = |S| = \cdots = N$$

$$R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$$
$$R = [N] \times [1],\ S = [1] \times [1],\ T = [1] \times [N].$$



$$R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge K(U, V)$$
$$R = T = [N] \times [1],\ S = K = [1] \times [N]$$



Summary of the AGM Bound

- Upper / lower bound: fractional edge cover / vertex packing.
- Their equality follows from strong duality.
- The worst-case instance of the AGM bound is a Product Database.
- Full CQs only. Otherwise, ignore non-head variables.

Limitation of AGM: only cardinalities. Next: extensions to other stats.

Extensions of the AGM Bound

Simple Functional Dependencies

Given functional dependencies, query output is $\ll$ AGM bound.

Example: $R(X, Y) \wedge S(Y, Z)$: N^2 becomes N when $Y \rightarrow Z$.

An FD $\boldsymbol{U} \rightarrow \boldsymbol{V}$ is simple if $\boldsymbol{U}$ is a single variable.

Method [Khamis et al., 2016]:

- Expand Q to Q^+ by replacing each atom $R(\boldsymbol{Y})$ with $R'(\boldsymbol{Y}^+)$.
- Compute the AGM bound of Q^+.
- This bound is tight. Proof: very useful exercise.

Example

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$
Fractional edge covers: $(1, 1, 0), (1, 0, 1), (0, 1, 1), (1/2, 1/2, 1/2)$

$$\boxed{|Q| \leq \min(|R| \cdot |S|, |R| \cdot |T|, |S| \cdot |T|, \sqrt{|R| \cdot |S| \cdot |T|})}$$

Example

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$
Fractional edge covers: $(1, 1, 0), (1, 0, 1), (0, 1, 1), (1/2, 1/2, 1/2)$

$$\boxed{|Q| \leq \min(|R| \cdot |S|, |R| \cdot |T|, |S| \cdot |T|, \sqrt{|R| \cdot |S| \cdot |T|})}$$

Assume that $S.Y$ is a key: $Y \rightarrow Z$

Example

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$
Fractional edge covers: $(1, 1, 0), (1, 0, 1), (0, 1, 1), (1/2, 1/2, 1/2)$

$$\boxed{|Q| \leq \min(|R| \cdot |S|, |R| \cdot |T|, |S| \cdot |T|, \sqrt{|R| \cdot |S| \cdot |T|})}$$

Assume that $S.Y$ is a key: $Y \rightarrow Z$
$Q^+(X, Y, Z) = R'(X, Y, {\color{red}Z}) \wedge S(Y, Z) \wedge T(Z, X)$
Fractional edge covers: ${\color{red}(1, 0, 0)}, (0, 1, 1)$

$$\boxed{|Q| \leq \min(|R|, |S| \cdot |T|)}$$

Discussion

The expansion procedure is very easy, but limited only to simple FDs:

$AGM(Q^+)$ is always an upper bound on Q's output, but may not be tight.

Example

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

$$A\text{: } XZ \rightarrow U; \qquad\qquad B\text{: } YU \rightarrow X$$

Expansion is useless ($Q^+ = Q$).

More Statistics

Statistics for a relation $R(U, V, W, \ldots)$:

- Its cardinality $|R|$.
- Number distinct values of an attribute / set of attributes, e.g. $|R.X|$.
- Max degree of an attribute / set of attributes, e.g. $\max(\deg_R(VW|U))$
- The max degree of a projection, e.g. $\max(\deg_R(V|U))$.
- The ℓ_p-norm of some degree sequence, e.g. $||\deg_R(V|U)||_2$.

Will use entropic inequalities, beyond Shearer

Example

$$R = \begin{array}{|ccc|} \hline U & V & W \\ \hline a & 1 & m \\ a & 1 & n \\ a & 2 & m \\ a & 3 & m \\ b & 1 & m \\ b & 5 & m \\ \hline \end{array}$$

$$\begin{aligned} |R| &= 6 \\ |R.U| &= 2 \\ |R.V| &= 4 \\ |R.UV| &= 5 \end{aligned}$$

$\max(\deg_R(VW|U)) = 4$

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

Conditional Entropy

The *Conditional Entropy*

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

Conditional Entropy

The *Conditional Entropy*

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

What it means: $h(\boldsymbol{V}|\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{u}}[h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})]$

Conditional Entropy

The *Conditional Entropy*

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

What it means: $h(\boldsymbol{V}|\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{u}}[h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})]$

The submodularity inequality can be written equivalently as:

$$h(\boldsymbol{V}|\boldsymbol{U}) \geq h(\boldsymbol{V}|\boldsymbol{UW})$$

From Entropy to Statistics

Fix a joint probability distribution of the variables $\boldsymbol{X}$, with support $R(\boldsymbol{X})$:

$$h(\boldsymbol{X}) \leq \log |R|$$

$$h(\boldsymbol{V}|\boldsymbol{U}) \leq \log\left(\max \deg_R(\boldsymbol{V}|\boldsymbol{U})\right)$$

$$h(\boldsymbol{UV}) + (p-1)h(\boldsymbol{V}|\boldsymbol{V}) \leq \log ||\deg_R(\boldsymbol{V}|\boldsymbol{U})||_p^p$$

Example

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

Assume $|R| = |S| = |T| = N$: $AGM(Q) = N^2$.

Example

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

Assume $|R| = |S| = |T| = N$: $AGM(Q) = N^2$.
If the FDs $XZ \to U$ and $YU \to X$ hold: $|Q| \le N^{3/2}$.

$$\log|R| + \log|S| + \log|T| + \log\max\deg_A(U|XZ) + \log\max\deg_B(X|YU) \ge$$
$$\ge h(XY) + h(YZ) + h(ZU) + h(U|XZ) + h(X|YU)$$

Example

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

Assume $|R| = |S| = |T| = N$: $AGM(Q) = N^2$.

If the FDs $XZ \to U$ and $YU \to X$ hold: $|Q| \leq N^{3/2}$.

$$\log|R| + \log|S| + \log|T| + \log\max\deg_A(U|XZ) + \log\max\deg_B(X|YU) \geq$$
$$\geq \underline{h(XY) + h(YZ)} + h(ZU) + h(U|XZ) + h(X|YU)$$

Example

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

Assume $|R| = |S| = |T| = N$: $AGM(Q) = N^2$.

If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold: $|Q| \leq N^{3/2}$.

$$\log|R| + \log|S| + \log|T| + \log\max\deg_A(U|XZ) + \log\max\deg_B(X|YU) \geq$$

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

$$\geq h(XYZ) + h(Y) + h(ZU) + h(U|XZ) + h(X|YU)$$

Example

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

Assume $|R| = |S| = |T| = N$: $AGM(Q) = N^2$.
If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold: $|Q| \leq N^{3/2}$.

$$\begin{aligned} &\log|R| + \log|S| + \log|T| + \log\max\deg_A(U|XZ) + \log\max\deg_B(X|YU) \geq \\ &\geq h(XY) + h(YZ) + h(ZU) + h(U|XZ) + h(X|YU) \\ &\qquad \geq h(XYZ) + \underline{h(Y) + h(ZU)} + h(U|XZ) + h(X|YU) \end{aligned}$$

Example

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

Assume $|R| = |S| = |T| = N$: $AGM(Q) = N^2$.
If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold: $|Q| \leq N^{3/2}$.

$$\begin{aligned}
&\log|R| + \log|S| + \log|T| + \log\max\deg_A(U|XZ) + \log\max\deg_B(X|YU) \geq \\
&\geq h(XY) + h(YZ) + h(ZU) + h(U|XZ) + h(X|YU) \\
&\qquad \geq h(XYZ) + \underline{h(Y) + h(ZU)} + h(U|XZ) + h(X|YU) \\
&\qquad \geq h(XYZ) + h(YZU) + h(U|XZ) + h(X|YU)
\end{aligned}$$

Example

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

Assume $|R| = |S| = |T| = N$: $AGM(Q) = N^2$.

If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold: $|Q| \leq N^{3/2}$.

$$\begin{aligned}
&\log|R| + \log|S| + \log|T| + \log\max\deg_A(U|XZ) + \log\max\deg_B(X|YU) \geq \\
&\geq h(XY) + h(YZ) + h(ZU) + h(U|XZ) + h(X|YU) \\
&\qquad \geq h(XYZ) + h(Y) + h(ZU) + h(U|XZ) + h(X|YU) \\
&\qquad \geq h(XYZ) + h(YZU) + \underline{h(U|XZ)} + \underline{h(X|YU)}
\end{aligned}$$

Example

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

Assume $|R| = |S| = |T| = N$: $AGM(Q) = N^2$.
If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold: $|Q| \leq N^{3/2}$.

$$\begin{aligned}
&\log|R| + \log|S| + \log|T| + \log\max\deg_A(U|XZ) + \log\max\deg_B(X|YU) \geq \\
&\geq h(XY) + h(YZ) + h(ZU) + h(U|XZ) + h(X|YU) \\
&\qquad \geq h(XYZ) + h(Y) + h(ZU) + h(U|XZ) + h(X|YU) \\
&\qquad \geq h(XYZ) + h(YZU) + \underline{h(U|XZ)} + \underline{h(X|YU)} \\
&\qquad \geq h(XYZ) + h(YZU) + h(U|XYZ) + h(X|YZU) \\
&\qquad = 2h(XYZU) = 2\log|Q|
\end{aligned}$$

Example

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

Assume $|R| = |S| = |T| = N$: $AGM(Q) = N^2$.

If the FDs $XZ \to U$ and $YU \to X$ hold: $|Q| \leq N^{3/2}$.

$$\begin{aligned}
&\log|R| + \log|S| + \log|T| + \log\max\deg_A(U|XZ) + \log\max\deg_B(X|YU) \geq \\
&\geq h(XY) + h(YZ) + h(ZU) + h(U|XZ) + h(X|YU) \\
&\quad \geq h(XYZ) + h(Y) + h(ZU) + h(U|XZ) + h(X|YU) \\
&\quad \geq h(XYZ) + h(YZU) + h(U|XZ) + h(X|YU) \\
&\quad \geq h(XYZ) + h(YZU) + h(U|XYZ) + h(X|YZU) \\
&\quad = 2h(XYZU) = 2\log|Q|
\end{aligned}$$

$$\boxed{|Q| \leq \sqrt{|R| \cdot |S| \cdot |T| \cdot \max(\deg(U|XZ)) \cdot \max(\deg(X|YU))}}$$

Discussion

- AGM/Shearer limited to cardinality statistics.
- More general statistics require general entropic inequalities.
- Everything gets harder: fractional edge cover no longer sufficient, order of the submodularity matters.
- Can we compute the upper bound? Is it tight? Yes and no, it's complicated [Suciu, 2023].
- Do they work in practice? Yes, but you need to do the engineering work [Deeds et al., 2023].

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.

Deeds, K. B., Suciu, D., and Balazinska, M. (2023).
Safebound: A practical system for generating cardinality bounds.
Proc. ACM Manag. Data, 1(1):53:1–53:26.

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.

Suciu, D. (2023).
Applications of information inequalities to database theory problems.
In *LICS*, pages 1–30.