

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 $AGM(Q, \mathbf{D}) \stackrel{\text{def}}{=} \max_{\mathbf{D} \models \text{statistics}} |Q(\mathbf{D})|$.
- Thursday: Worst Case Optimal Join, by guest lecturer [Hung Ngo](#). An algorithm that computes $Q(\mathbf{D})$ in time $\tilde{O}(AGM(Q, \mathbf{D}))$.

Background on Cardinality Estimation

Cardinality Estimation 101 (1/3)

Given:

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

“Estimate”:

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

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**:

$$|\sigma_p(R)| \approx \theta_p \cdot |R|$$

$$\theta_{p_1 \wedge p_2} \approx \theta_{p_1} \cdot \theta_{p_2}$$

Histograms, multidimensional histograms.

- Join $J(A, B, C) = R(A, B) \wedge S(B, C)$: assume **preservation of values**

$$\triangleright |J| \approx |R| \cdot \text{avg}(\text{deg}_S(C|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(S.B)|}$$

$$\triangleright |J| \approx |S| \cdot \text{avg}(\text{deg}_R(A|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(R.B)|}$$

- Heuristic: take the minimum:

$$|J| \approx \frac{|R| \cdot |S|}{\max(|\text{Dom}(R.B)|, |\text{Dom}(S.B)|)}$$

Cardinality Estimation 101 (2/3)

Bottom-up on the query plan:

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

$$|\sigma_p(R)| \approx \theta_p \cdot |R|$$

$$\theta_{p_1 \wedge p_2} \approx \theta_{p_1} \cdot \theta_{p_2}$$

Histograms, multidimensional histograms.

- Join $J(A, B, C) = R(A, B) \wedge S(B, C)$: assume **preservation of values**

$$\triangleright |J| \approx |R| \cdot \text{avg}(\text{deg}_S(C|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(S.B)|}$$

$$\triangleright |J| \approx |S| \cdot \text{avg}(\text{deg}_R(A|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(R.B)|}$$

- Heuristic: take the minimum:

$$|J| \approx \frac{|R| \cdot |S|}{\max(|\text{Dom}(R.B)|, |\text{Dom}(S.B)|)}$$

Cardinality Estimation 101 (2/3)

Bottom-up on the query plan:

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

$$|\sigma_p(R)| \approx \theta_p \cdot |R|$$

$$\theta_{p_1 \wedge p_2} \approx \theta_{p_1} \cdot \theta_{p_2}$$

Histograms, multidimensional histograms.

- Join $J(A, B, C) = R(A, B) \wedge S(B, C)$: assume **preservation of values**

$$\triangleright |J| \approx |R| \cdot \text{avg}(\deg_S(C|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(S.B)|}$$

$$\triangleright |J| \approx |S| \cdot \text{avg}(\deg_R(A|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(R.B)|}$$

- Heuristic: take the minimum:

$$|J| \approx \frac{|R| \cdot |S|}{\max(|\text{Dom}(R.B)|, |\text{Dom}(S.B)|)}$$

Cardinality Estimation 101 (2/3)

Bottom-up on the query plan:

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

$$|\sigma_p(R)| \approx \theta_p \cdot |R|$$

$$\theta_{p_1 \wedge p_2} \approx \theta_{p_1} \cdot \theta_{p_2}$$

Histograms, multidimensional histograms.

- Join $J(A, B, C) = R(A, B) \wedge S(B, C)$: assume **preservation of values**

$$\triangleright |J| \approx |R| \cdot \text{avg}(\text{deg}_S(C|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(S.B)|}$$

$$\triangleright |J| \approx |S| \cdot \text{avg}(\text{deg}_R(A|B)) = \frac{|R| \cdot |S|}{|\text{Dom}(R.B)|}$$

- Heuristic: take the minimum:

$$|J| \approx \frac{|R| \cdot |S|}{\max(|\text{Dom}(R.B)|, |\text{Dom}(S.B)|)}$$

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 \mathbf{D} , e.g. cardinalities, # distinct values,

Compute an upper bound B :

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

Challenge: make B tight.

Simple Examples

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

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

$$\max_D |Q(\mathbf{D})| = ?$$

Simple Examples

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

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

$$\max_D |Q(\mathbf{D})| = N^2$$

Simple Examples

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

- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z)$.
If $S.Y$ is a key:

$$\begin{aligned}\max_D |Q(\mathbf{D})| &= N^2 \\ \max_D |Q(\mathbf{D})| &= N\end{aligned}$$

Simple Examples

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

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

If $S.Y$ is a key:

$$\max_D |Q(D)| = N^2$$
$$\max_D |Q(D)| = N$$

- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_D |Q(D)| = ?$

Simple Examples

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

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

If $S.Y$ is a key:

$$\max_D |Q(D)| = N^2$$
$$\max_D |Q(D)| = N$$

- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_D |Q(D)| = N^2$

Simple Examples

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

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

If $S.Y$ is a key:

$$\begin{aligned}\max_D |Q(\mathbf{D})| &= N^2 \\ \max_D |Q(\mathbf{D})| &= N\end{aligned}$$

- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_D |Q(\mathbf{D})| = N^2$

Notice the role of an **edge cover**

Simple Examples

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

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

If $S.Y$ is a key:

$$\max_D |Q(\mathbf{D})| = N^2$$
$$\max_D |Q(\mathbf{D})| = N$$

- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_D |Q(\mathbf{D})| = N^2$

Notice the role of an **edge cover**

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

Simple Examples

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

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

If $S.Y$ is a key:

$$\max_D |Q(\mathbf{D})| = N^2$$
$$\max_D |Q(\mathbf{D})| = N$$

- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_D |Q(\mathbf{D})| = N^2$

Notice the role of an **edge cover**

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

Simple Examples

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

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

If $S.Y$ is a key:

$$\max_D |Q(D)| = N^2$$

$$\max_D |Q(D)| = N$$

- $Q(X, Y, Z, U) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, U)$. $\max_D |Q(D)| = N^2$

Notice the role of an **edge cover**

- $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$. $\max_D |Q(D)| = N^{\frac{3}{2}}$

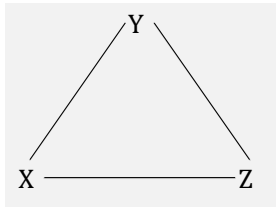
Here we use a **fractional edge cover**

AGM Bound: The Statement

Fractional Edge Covers

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

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



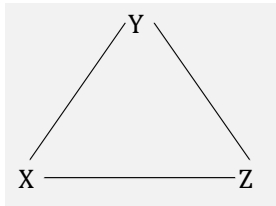
Fractional Edge Covers

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

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

Definition

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



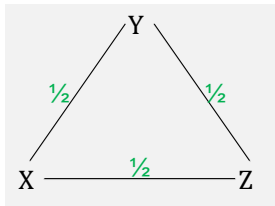
Fractional Edge Covers

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

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

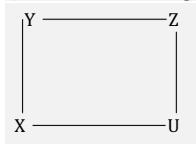
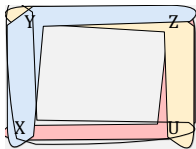
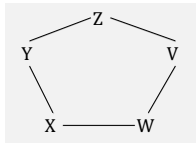
Definition

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



Examples

What are fractional edge covers?



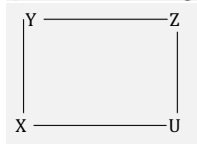
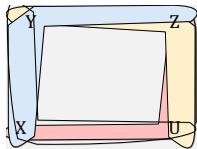
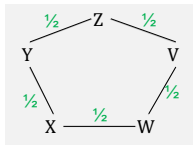
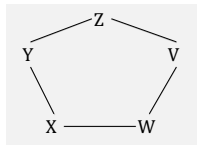
5-cycle

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?



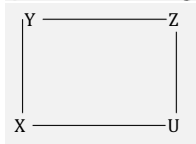
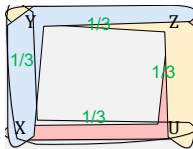
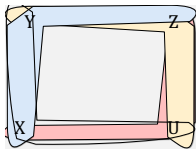
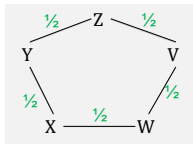
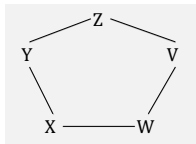
5-cycle

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?



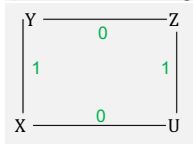
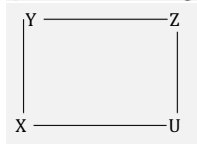
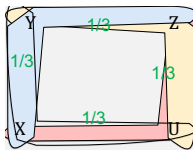
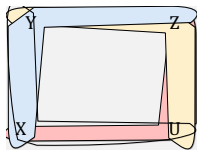
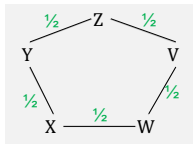
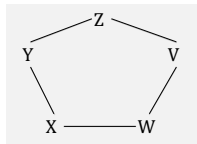
5-cycle

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?



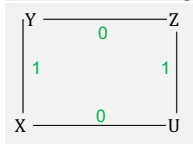
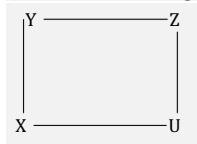
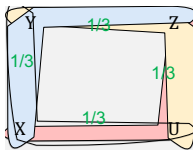
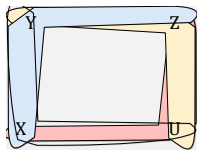
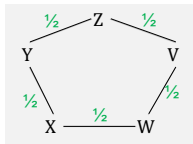
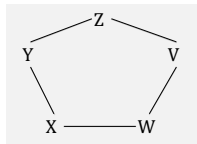
5-cycle

Loomis-Whitney:

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

Examples

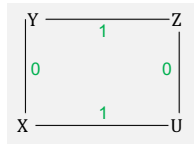
What are fractional edge covers?



5-cycle

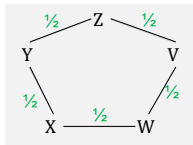
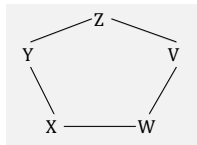
Loomis-Whitney:

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

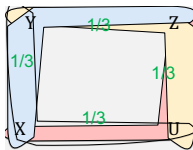
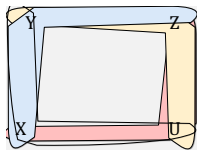


Examples

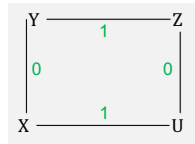
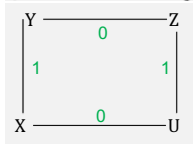
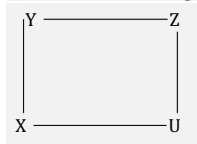
What are fractional edge covers?



5-cycle



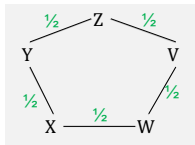
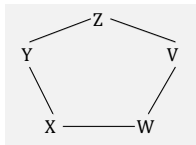
Loomis-Whitney:



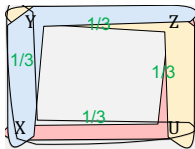
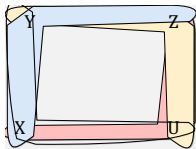
$(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ is a convex combination of $(1, 0, 1, 0)$ and $(0, 1, 0, 1)$.

Examples

What are fractional edge covers?

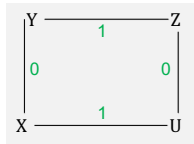
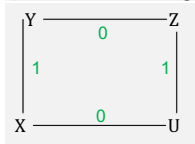
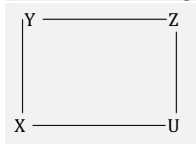


5-cycle



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: not \geq convex combination of others.

The AGM Bound [Atserias et al., 2013]

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

Theorem (Upper Bound)

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

The AGM Bound [Atserias et al., 2013]

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

Theorem (Upper Bound)

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

Theorem (Lower Bound)

$AGM(Q) \stackrel{\text{def}}{=} \min_{\mathbf{w}} |R_1|^{w_1} \cdots |R_m|^{w_m}$ is “tight”.

The AGM Bound [Atserias et al., 2013]

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

Theorem (Upper Bound)

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

Theorem (Lower Bound)

$AGM(Q) \stackrel{\text{def}}{=} \min_{\mathbf{w}} |R_1|^{w_1} \cdots |R_m|^{w_m}$ is “tight”.

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

The AGM Bound [Atserias et al., 2013]

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

Theorem (Upper Bound)

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

Theorem (Lower Bound)

$AGM(Q) \stackrel{\text{def}}{=} \min_{\mathbf{w}} |R_1|^{w_1} \cdots |R_m|^{w_m}$ is “tight”.

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

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 $p : D \rightarrow [0, 1]$. $X = \text{r.v. with outcomes } D$.

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 $p : D \rightarrow [0, 1]$. $X = \text{r.v. with outcomes } D$.

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 $p : D \rightarrow [0, 1]$. $X = \text{r.v. with outcomes } D$.

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

Entropic Vectors

Definition

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

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

X	Y
a	p
a	q
b	q
a	m

Entropic Vectors

Definition

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

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

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 $p : D \rightarrow [0, 1]$. $X = \text{r.v. with outcomes } D$.

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

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

$$h(\emptyset) = 0$$

$$h(\mathbf{U} \cup \mathbf{V}) \geq h(\mathbf{U})$$

Monotonicity

$$h(\mathbf{U}) + h(\mathbf{V}) \geq h(\mathbf{U} \cup \mathbf{V}) + h(\mathbf{U} \cap \mathbf{V})$$

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

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

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

A Shannon Inequality

Example

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

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

A Shannon Inequality

Example

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

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

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:

$$|Q| \leq |R_1|^{w_1} \dots |R_m|^{w_m}$$

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

From Query to Information Inequality:

Example

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

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

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

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

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 $\mathbf{D} = (R^D, S^D, T^D)$; $p : Q(\mathbf{D}) \rightarrow [0, 1]$ uniform; 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: $|Q| \leq |R_1|^{w_1} \dots |R_m|^{w_m}$

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

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

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

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

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

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

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

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

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

$$\text{If } \sum_j w_j h(\mathbf{Y}_j) \geq h(\mathbf{X}) \text{ then } |R_1|^{w_1} \dots |R_m|^{w_m} \geq |Q|$$

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

Theorem (Generalized Shearer's Inequality)

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

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

Theorem (Generalized Shearer's Inequality)

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

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

$$k_1 h(\mathbf{Y}_1) + \dots + k_m h(\mathbf{Y}_m) \geq k_0 h(\mathbf{X})$$

Each variable is “covered $\geq k_0$ times”.

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

Theorem (Generalized Shearer's Inequality)

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

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

$$k_1 h(\mathbf{Y}_1) + \dots + k_m h(\mathbf{Y}_m) \geq k_0 h(\mathbf{X})$$

Each variable is “covered $\geq k_0$ times”.

Repeatedly rewrite $h(\mathbf{Y}_i) + h(\mathbf{Y}_j) \rightarrow h(\mathbf{Y}_i \cup \mathbf{Y}_j) + h(\mathbf{Y}_i \cap \mathbf{Y}_j)$

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

Theorem (Generalized Shearer's Inequality)

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

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

$$k_1 h(\mathbf{Y}_1) + \dots + k_m h(\mathbf{Y}_m) \geq k_0 h(\mathbf{X})$$

Each variable is “covered $\geq k_0$ times”.

Repeatedly rewrite $h(\mathbf{Y}_i) + h(\mathbf{Y}_j) \rightarrow h(\mathbf{Y}_i \cup \mathbf{Y}_j) + h(\mathbf{Y}_i \cap \mathbf{Y}_j)$

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

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

Theorem (Generalized Shearer's Inequality)

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

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

$$k_1 h(\mathbf{Y}_1) + \dots + k_m h(\mathbf{Y}_m) \geq k_0 h(\mathbf{X})$$

Each variable is “covered $\geq k_0$ times”.

Repeatedly rewrite $h(\mathbf{Y}_i) + h(\mathbf{Y}_j) \rightarrow h(\mathbf{Y}_i \cup \mathbf{Y}_j) + h(\mathbf{Y}_i \cap \mathbf{Y}_j)$

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

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

Theorem (Generalized Shearer's Inequality)

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

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

$$k_1 h(\mathbf{Y}_1) + \dots + k_m h(\mathbf{Y}_m) \geq k_0 h(\mathbf{X})$$

Each variable is “covered $\geq k_0$ times”.

Repeatedly rewrite $h(\mathbf{Y}_i) + h(\mathbf{Y}_j) \rightarrow h(\mathbf{Y}_i \cup \mathbf{Y}_j) + h(\mathbf{Y}_i \cap \mathbf{Y}_j)$

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

- $\sum_{\ell} |\mathbf{Y}_{\ell}|^2$ strictly increases (homework!).

When do we stop?

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

Theorem (Generalized Shearer's Inequality)

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

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

$$k_1 h(\mathbf{Y}_1) + \dots + k_m h(\mathbf{Y}_m) \geq k_0 h(\mathbf{X})$$

Each variable is “covered $\geq k_0$ times”.

Repeatedly rewrite $h(\mathbf{Y}_i) + h(\mathbf{Y}_j) \rightarrow h(\mathbf{Y}_i \cup \mathbf{Y}_j) + h(\mathbf{Y}_i \cap \mathbf{Y}_j)$

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

- $\sum_{\ell} |\mathbf{Y}_{\ell}|^2$ strictly increases (homework!).

When do we stop?

- We stop when $\mathbf{Y}_1 \supset \mathbf{Y}_2 \supset \dots$

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

Theorem (Generalized Shearer's Inequality)

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

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

$$k_1 h(\mathbf{Y}_1) + \dots + k_m h(\mathbf{Y}_m) \geq k_0 h(\mathbf{X})$$

Each variable is “covered $\geq k_0$ times”.

Repeatedly rewrite $h(\mathbf{Y}_i) + h(\mathbf{Y}_j) \rightarrow h(\mathbf{Y}_i \cup \mathbf{Y}_j) + h(\mathbf{Y}_i \cap \mathbf{Y}_j)$

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

- $\sum_{\ell} |\mathbf{Y}_{\ell}|^2$ strictly increases (homework!).

When do we stop?

- We stop when $\mathbf{Y}_1 \supset \mathbf{Y}_2 \supset \dots$

- Thus, $\mathbf{Y}_1 = \mathbf{X}$ and $k_1 \geq k_0$.

$$\dots \geq k_1 h(\mathbf{Y}_1) \geq k_0 h(\mathbf{X})$$

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:

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

$$AGM(Q) = \min_{\mathbf{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$$

Proof of the AGM Lower Bound

By example:

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

$$AGM(Q) = \min_{\mathbf{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$$

Primal program:

Minimize

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

where \mathbf{w} is frac. edge cover:

$$X : \quad w_R + \quad \quad w_T \geq 1$$

$$Y : \quad w_R + \quad w_S \geq 1$$

$$Z : \quad \quad w_S + \quad w_T \geq 1$$

Proof of the AGM Lower Bound

By example:

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

$$AGM(Q) = \min_{\mathbf{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$$

Primal program:

Minimize

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

where \mathbf{w} is frac. edge cover:

$$X : \quad w_R + \quad \quad w_T \geq 1$$

$$Y : \quad w_R + \quad w_S \geq 1$$

$$Z : \quad \quad w_S + \quad w_T \geq 1$$

Dual program:

Maximize

$$v_X + v_Y + v_Z$$

where \mathbf{v} is “frac. vertex packing”:

$$R : \quad v_X + \quad v_Y \leq \log |R|$$

$$S : \quad \quad v_Y + \quad v_Z \leq \log |S|$$

$$T : \quad v_X + \quad \quad v_Z \leq \log |T|$$

Proof of the AGM Lower Bound

By example:

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

$$AGM(Q) = \min_{\mathbf{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$$

Primal program:

Minimize

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

where \mathbf{w} is frac. edge cover:

$$X : \quad w_R + \quad \quad w_T \geq 1$$

$$Y : \quad w_R + \quad w_S \geq 1$$

$$Z : \quad \quad w_S + \quad w_T \geq 1$$

Dual program:

Maximize

$$v_X + v_Y + v_Z$$

where \mathbf{v} is “frac. vertex packing”:

$$R : \quad v_X + \quad v_Y \leq \log |R|$$

$$S : \quad \quad v_Y + \quad v_Z \leq \log |S|$$

$$T : \quad v_X + \quad \quad v_Z \leq \log |T|$$

Take optimum \mathbf{v} , define: $\text{Dom}(X) \stackrel{\text{def}}{=} \llbracket 2^{v_X} \rrbracket$, $\text{Dom}(Y) \stackrel{\text{def}}{=} \llbracket 2^{v_Y} \rrbracket$, $\text{Dom}(Z) \stackrel{\text{def}}{=} \llbracket 2^{v_Z} \rrbracket$.

Proof of the AGM Lower Bound

By example:

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

$$AGM(Q) = \min_{\mathbf{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$$

Primal program:

Minimize

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

where \mathbf{w} is frac. edge cover:

$$X : \quad w_R + \quad \quad w_T \geq 1$$

$$Y : \quad w_R + \quad w_S \geq 1$$

$$Z : \quad \quad w_S + \quad w_T \geq 1$$

Dual program:

Maximize

$$v_X + v_Y + v_Z$$

where \mathbf{v} is “frac. vertex packing”:

$$R : \quad v_X + \quad v_Y \leq \log |R|$$

$$S : \quad \quad v_Y + \quad v_Z \leq \log |S|$$

$$T : \quad v_X + \quad \quad v_Z \leq \log |T|$$

Take optimum \mathbf{v} , define: $\text{Dom}(X) \stackrel{\text{def}}{=} \llbracket 2^{v_X} \rrbracket$, $\text{Dom}(Y) \stackrel{\text{def}}{=} \llbracket 2^{v_Y} \rrbracket$, $\text{Dom}(Z) \stackrel{\text{def}}{=} \llbracket 2^{v_Z} \rrbracket$.

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

Proof of the AGM Lower Bound

By example:

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

$$AGM(Q) = \min_{\mathbf{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$$

Primal program:

Minimize

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

where \mathbf{w} is frac. edge cover:

$$X : \quad w_R + \quad \quad w_T \geq 1$$

$$Y : \quad w_R + \quad w_S \geq 1$$

$$Z : \quad \quad w_S + \quad w_T \geq 1$$

Dual program:

Maximize

$$v_X + v_Y + v_Z$$

where \mathbf{v} is “frac. vertex packing”:

$$R : \quad v_X + \quad v_Y \leq \log |R|$$

$$S : \quad \quad v_Y + \quad v_Z \leq \log |S|$$

$$T : \quad v_X + \quad \quad v_Z \leq \log |T|$$

Take optimum \mathbf{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^* , $T^* \stackrel{\text{def}}{=} \dots$

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

Proof of the AGM Lower Bound

By example:

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

$$AGM(Q) = \min_{\mathbf{w}} |R|^{w_R} \cdot |S|^{w_S} \cdot |T|^{w_T}$$

Primal program:

Minimize

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

where \mathbf{w} is frac. edge cover:

$$X : \quad w_R + \quad \quad w_T \geq 1$$

$$Y : \quad w_R + \quad w_S \geq 1$$

$$Z : \quad \quad w_S + \quad w_T \geq 1$$

Dual program:

Maximize

$$v_X + v_Y + v_Z$$

where \mathbf{v} is “frac. vertex packing”:

$$R : \quad v_X + \quad v_Y \leq \log |R|$$

$$S : \quad \quad v_Y + \quad v_Z \leq \log |S|$$

$$T : \quad v_X + \quad \quad v_Z \leq \log |T|$$

Take optimum \mathbf{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): $R^* \stackrel{\text{def}}{=} \text{Dom}(X) \times \text{Dom}(Y)$, $S^*, T^* \stackrel{\text{def}}{=} \dots$

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

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

Definition

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

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

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

Definition

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

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

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

$$v_R + v_S \leq \log N$$

$$v_R + v_T \leq \log N$$

...

with

$$v_R + v_S \leq 1$$

$$v_R + v_T \leq 1$$

...

times $\log N$.

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

Definition

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

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

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

$$v_R + v_S \leq \log N$$

$$v_R + v_T \leq \log N$$

...

with

$$v_R + v_S \leq 1$$

$$v_R + v_T \leq 1$$

...

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

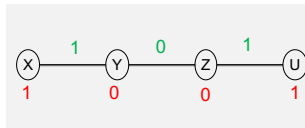
$$|R| = |S| = \dots = N$$

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

Examples

$$|R| = |S| = \dots = N$$

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

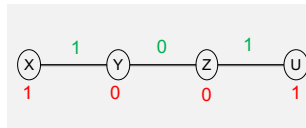


Examples

$$|R| = |S| = \dots = N$$

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

$$R = [N] \times [1], S = [1] \times [1], T = [1] \times [N].$$

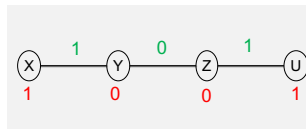


Examples

$$|R| = |S| = \dots = N$$

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

$$R = [N] \times [1], S = [1] \times [1], T = [1] \times [N].$$



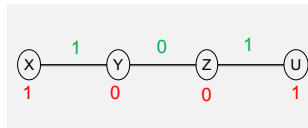
$$R(X, Y) \wedge S(Y, Z) \wedge T(Z, U) \wedge K(U, V)$$

Examples

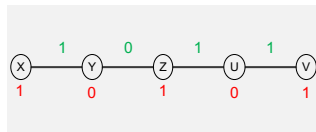
$$|R| = |S| = \dots = N$$

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

$$R = [N] \times [1], S = [1] \times [1], T = [1] \times [N].$$



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

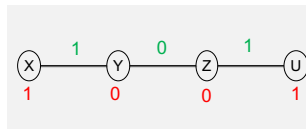


Examples

$$|R| = |S| = \dots = N$$

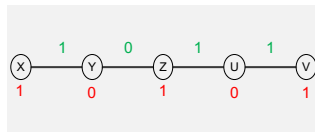
$$R(\textcolor{red}{X}, Y) \wedge S(Y, Z) \wedge T(Z, \textcolor{red}{U})$$

$$R = [\textcolor{red}{N}] \times [1], S = [1] \times [1], T = [1] \times [\textcolor{red}{N}].$$



$$R(\textcolor{red}{X}, Y) \wedge S(Y, \textcolor{red}{Z}) \wedge T(\textcolor{red}{Z}, U) \wedge K(U, \textcolor{red}{V})$$

$$R = T = [\textcolor{red}{N}] \times [1], S = K = [1] \times [\textcolor{red}{N}]$$



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

More Statistics

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

- Cardinality $|R|$.
- Same for any projection: $|R.U|, |R.UV|, \dots$
- Max degree: $\max(\deg_R(VW|U)), \dots$
- Note: an FD $U \rightarrow V$ is $\max(\deg_R(V|U)) = 0$.
- ℓ_p -norm degree sequences: $\|\deg_R(V|U)\|_2, \dots$

Simple Functional Dependencies

Given FDs, $|Q| \ll AGM(Q)$.

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

Simple Functional Dependencies

Given FDs, $|Q| \ll AGM(Q)$.

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

$U \rightarrow V$ is **simple** if $|U| = 1$.

Simple Functional Dependencies

Given FDs, $|Q| \ll AGM(Q)$.

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

$U \rightarrow V$ is **simple** if $|U| = 1$.

Method [Khamis et al., 2016]:

- **Expand** Q to Q^+ by replacing each atom $R(Y)$ with $R'(Y^+)$.
- Return $AGM(Q^+)$.
- 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: $(1, 1, 0), (1, 0, 1), (0, 1, 1), (1/2, 1/2, 1/2)$

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

Example

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

Fractional edge covers: $(1, 1, 0), (1, 0, 1), (0, 1, 1), (1/2, 1/2, 1/2)$

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

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: $(1, 1, 0), (1, 0, 1), (0, 1, 1), (1/2, 1/2, 1/2)$

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

Assume that $S.Y$ is a key:

$$Y \rightarrow Z$$

$$Q^+(X, Y, Z) = R'(X, Y, Z) \wedge S(Y, Z) \wedge T(Z, X)$$

Fractional edge covers: $(1, 0, 0), (0, 1, 1)$

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

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.

Need to use entropic inequalities, beyond Shearer

Conditional Entropy

The *Conditional Entropy*

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

Conditional Entropy

The *Conditional Entropy*

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

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

Conditional Entropy

The *Conditional Entropy*

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

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

The submodularity inequality can be written equivalently as:

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

Example of Statistics:

$R =$

U	V	W
a	1	m
a	1	n
a	2	m
a	3	m
b	1	m
b	5	m

$$|R| = 6$$

$$|R.U| = 2$$

$$|R.V| = 4$$

$$|R.UV| = 5$$

$$h(UVV) \leq \log |R|$$

$$h(U) \leq \log |R.U|$$

...

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

$$h(VW|U) \leq \log \max(\deg_R(VW|U))$$

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

$$h(V|U) \leq \log \max(\deg_R(V|U))$$

Example of Upper Bound

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

Assume $|R| = |S| = |T| = N$:

$$AGM(Q) = N^2.$$

Example of Upper Bound

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

Assume $|R| = |S| = |T| = N$:

If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold:

$$AGM(Q) = N^2.$$

$$|Q| \leq N^{3/2}.$$

Example of Upper Bound

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

Assume $|R| = |S| = |T| = N$:

$$AGM(Q) = N^2.$$

If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold:

$$|Q| \leq N^{3/2}.$$

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

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

Example of Upper Bound

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

Assume $|R| = |S| = |T| = N$:

$$AGM(Q) = N^2.$$

If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold:

$$|Q| \leq N^{3/2}.$$

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

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

Example of Upper Bound

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

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 of Upper Bound

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

Assume $|R| = |S| = |T| = N$:

$$AGM(Q) = N^2.$$

If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold:

$$|Q| \leq N^{3/2}.$$

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

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

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

Example of Upper Bound

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

Assume $|R| = |S| = |T| = N$:

$$AGM(Q) = N^2.$$

If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold:

$$|Q| \leq N^{3/2}.$$

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

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

Example of Upper Bound

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

Assume $|R| = |S| = |T| = N$:

$$AGM(Q) = N^2.$$

If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold:

$$|Q| \leq N^{3/2}.$$

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

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

Example of Upper Bound

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

Assume $|R| = |S| = |T| = N$:

$$AGM(Q) = N^2.$$

If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold:

$$|Q| \leq N^{3/2}.$$

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

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

Example of Upper Bound

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

Assume $|R| = |S| = |T| = N$:

$$AGM(Q) = N^2.$$

If the FDs $XZ \rightarrow U$ and $YU \rightarrow X$ hold:

$$|Q| \leq N^{3/2}.$$

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

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

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

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.