

CS294-248 Special Topics in Database Theory

Unit 5: Entropies, Database Constraints

Dan Suciu

University of Washington

Outline

- Today: recap the AGM and its generalization.
- Thursday: Databas Constraints

AGM Bound

Fractional Edge Cover / Vertex Packing

Hypergraph $G = (V, E)$

Fractional Edge Cover \mathbf{w}

Minimize $\sum_e w_e$, where:

$$\forall x \in V : \sum_{e \in E: x \in e} w_e \geq 1$$
$$w_e \geq 0$$

Fractional Vertex Packing \mathbf{v}

Maximize $\sum_x v_x$, where:

$$\forall e \in E : \sum_{x \in V: x \in e} v_x \leq 1$$
$$v_x \geq 0$$

Weak duality: $\sum_e w_e$

Fractional Edge Cover / Vertex Packing

Hypergraph $G = (V, E)$

Fractional Edge Cover \mathbf{w}

Minimize $\sum_e w_e$, where:

$$\forall x \in V : \sum_{e \in E: x \in e} w_e \geq 1$$
$$w_e \geq 0$$

Fractional Vertex Packing \mathbf{v}

Maximize $\sum_x v_x$, where:

$$\forall e \in E : \sum_{x \in V: x \in e} v_x \leq 1$$
$$v_x \geq 0$$

Weak duality: $\sum_e w_e \geq \sum_e w_e (\sum_{x \in e} v_x)$

Fractional Edge Cover / Vertex Packing

Hypergraph $G = (V, E)$

Fractional Edge Cover \mathbf{w}

Minimize $\sum_e w_e$, where:

$$\forall x \in V : \sum_{e \in E: x \in e} w_e \geq 1$$
$$w_e \geq 0$$

Fractional Vertex Packing \mathbf{v}

Maximize $\sum_x v_x$, where:

$$\forall e \in E : \sum_{x \in V: x \in e} v_x \leq 1$$
$$v_x \geq 0$$

Weak duality: $\sum_e w_e \geq \sum_e w_e (\sum_{x \in e} v_x) = \sum_x v_x (\sum_{e: x \in e} w_e)$

Fractional Edge Cover / Vertex Packing

Hypergraph $G = (V, E)$

Fractional Edge Cover \mathbf{w}

Minimize $\sum_e w_e$, where:

$$\forall x \in V : \sum_{e \in E: x \in e} w_e \geq 1$$
$$w_e \geq 0$$

Fractional Vertex Packing \mathbf{v}

Maximize $\sum_x v_x$, where:

$$\forall e \in E : \sum_{x \in V: x \in e} v_x \leq 1$$
$$v_x \geq 0$$

Weak duality: $\sum_e w_e \geq \sum_e w_e (\sum_{x \in e} v_x) = \sum_x v_x (\sum_{e: x \in e} w_e) \geq \sum_x v_x$

Fractional Edge Cover / Vertex Packing

Hypergraph $G = (V, E)$

Fractional Edge Cover \mathbf{w}

Minimize $\sum_e w_e$, where:

$$\forall x \in V : \sum_{e \in E: x \in e} w_e \geq 1$$
$$w_e \geq 0$$

Fractional Vertex Packing \mathbf{v}

Maximize $\sum_x v_x$, where:

$$\forall e \in E : \sum_{x \in V: x \in e} v_x \leq 1$$
$$v_x \geq 0$$

Weak duality: $\sum_e w_e \geq \sum_e w_e (\sum_{x \in e} v_x) = \sum_x v_x (\sum_{e: x \in e} w_e) \geq \sum_x v_x$

Strong duality: $\min_{\mathbf{w}} \sum_e w_e = \max_{\mathbf{v}} \sum_x v_x \stackrel{\text{def}}{=} \rho^*$

Fractional edge covering number

The AGM Bound

[Atserias et al., 2013]

$$Q(\mathbf{x}) = \bigwedge_j R_j(\mathbf{x}_j)$$

Full CQ with m relations, n variables

Assume $|R_j| = N$ for all j .

Upper bound: $|Q| \leq N^{\rho^*}$

Proof: we used entropic inequalities. Elementary proof in [Suciu, 2023]

Lower bound: $|Q| \geq \frac{1}{2^n} N^{\rho^*}$ on product database $R_j \stackrel{\text{def}}{=} \prod_{x_i \in \text{Vars}(R_j)} [N^{v_i^*}]$,

where \mathbf{v}^* = optimal vertex packing.

The AGM Bound

[Atserias et al., 2013]

$$Q(\mathbf{x}) = \bigwedge_j R_j(\mathbf{x}_j)$$

Full CQ with m relations, n variables

Assume $|R_j| = N$ for all j .

Upper bound: $|Q| \leq N^{\rho^*}$

Proof: we used entropic inequalities. Elementary proof in [Suciu, 2023]

Lower bound: $|Q| \geq \frac{1}{2^n} N^{\rho^*}$ on product database $R_j \stackrel{\text{def}}{=} \prod_{x_i \in \text{Vars}(R_j)} [N^{v_i^*}]$,

where \mathbf{v}^* = optimal vertex packing.

The AGM Bound

[Atserias et al., 2013]

$$Q(\mathbf{x}) = \bigwedge_j R_j(\mathbf{x}_j)$$

Full CQ with m relations, n variables

Assume $|R_j| = N$ for all j .

Upper bound: $|Q| \leq N^{\rho^*}$

Proof: we used entropic inequalities. Elementary proof in [Suciu, 2023]

Lower bound: $|Q| \geq \frac{1}{2^n} N^{\rho^*}$ on product database $R_j \stackrel{\text{def}}{=} \prod_{x_i \in \text{Vars}(R_j)} [N^{v_i^*}]$,

where \mathbf{v}^* = optimal vertex packing.

Examples

$$L_5: \boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$$

Examples

$$L_5: \boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$$

$$\mathbf{w}^* = (1, 1, 0, 1), \mathbf{v}^* = (1, 0, 1, 0, 1).$$

$$AGM = N^3, \quad A_1, \dots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$$

Examples

$$L_5: \boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$$

$$\mathbf{w}^* = (1, 1, 0, 1), \mathbf{v}^* = (1, 0, 1, 0, 1).$$

$$AGM = N^3, \quad A_1, \dots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$$

$$C_5: \boxed{A_{12}(x_1, x_2) \wedge A_{23}(x_2, x_3) \wedge A_{34}(x_3, x_4) \wedge A_{45}(x_4, x_5) \wedge A_{51}(x_5, x_1)}$$

Examples

$$L_5: \boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$$

$$\mathbf{w}^* = (1, 1, 0, 1), \mathbf{v}^* = (1, 0, 1, 0, 1).$$

$$AGM = N^3, \quad A_1, \dots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$$

$$C_5: \boxed{A_{12}(x_1, x_2) \wedge A_{23}(x_2, x_3) \wedge A_{34}(x_3, x_4) \wedge A_{45}(x_4, x_5) \wedge A_{51}(x_5, x_1)}$$

$$\mathbf{w}^* = (1/2, \dots, 1/2), \mathbf{v}^* = (1/2, \dots, 1/2).$$

$$AGM = N^{5/2}; \quad A_{12} = A_{23} = \dots = [N^{1/2}] \times [N^{1/2}]$$

Examples

$$L_5: \boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$$

$$\mathbf{w}^* = (1, 1, 0, 1), \mathbf{v}^* = (1, 0, 1, 0, 1).$$

$$AGM = N^3, \quad A_1, \dots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$$

$$C_5: \boxed{A_{12}(x_1, x_2) \wedge A_{23}(x_2, x_3) \wedge A_{34}(x_3, x_4) \wedge A_{45}(x_4, x_5) \wedge A_{51}(x_5, x_1)}$$

$$\mathbf{w}^* = (1/2, \dots, 1/2), \mathbf{v}^* = (1/2, \dots, 1/2).$$

$$AGM = N^{5/2}; \quad A_{12} = A_{23} = \dots = [N^{1/2}] \times [N^{1/2}]$$

$$K_5: \boxed{\bigwedge_{1 \leq i < j \leq 5} A_{ij}(x_i, x_j)}$$

Examples

$$L_5: \boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$$

$$\mathbf{w}^* = (1, 1, 0, 1), \mathbf{v}^* = (1, 0, 1, 0, 1).$$

$$AGM = N^3, \quad A_1, \dots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$$

$$C_5: \boxed{A_{12}(x_1, x_2) \wedge A_{23}(x_2, x_3) \wedge A_{34}(x_3, x_4) \wedge A_{45}(x_4, x_5) \wedge A_{51}(x_5, x_1)}$$

$$\mathbf{w}^* = (1/2, \dots, 1/2), \mathbf{v}^* = (1/2, \dots, 1/2).$$

$$AGM = N^{5/2}; \quad A_{12} = A_{23} = \dots = [N^{1/2}] \times [N^{1/2}]$$

$$K_5: \boxed{\bigwedge_{1 \leq i < j \leq 5} A_{ij}(x_i, x_j)}$$

$$\mathbf{w}^* = (1/4, \dots, 1/4), \mathbf{v}^* = (1/2, 1/2, 1/2, 1/2, 1/2)$$

$$AGM = N^{5/2}; \quad A_{12} = A_{23} = \dots = [N^{1/2}] \times [N^{1/2}]$$

Examples

$$L_5: \boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$$

$$\mathbf{w}^* = (1, 1, 0, 1), \mathbf{v}^* = (1, 0, 1, 0, 1).$$

$$AGM = N^3, \quad A_1, \dots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$$

$$C_5: \boxed{A_{12}(x_1, x_2) \wedge A_{23}(x_2, x_3) \wedge A_{34}(x_3, x_4) \wedge A_{45}(x_4, x_5) \wedge A_{51}(x_5, x_1)}$$

$$\mathbf{w}^* = (1/2, \dots, 1/2), \mathbf{v}^* = (1/2, \dots, 1/2).$$

$$AGM = N^{5/2}; \quad A_{12} = A_{23} = \dots = [N^{1/2}] \times [N^{1/2}]$$

$$K_5: \boxed{\bigwedge_{1 \leq i < j \leq 5} A_{ij}(x_i, x_j)}$$

$$\mathbf{w}^* = (1/4, \dots, 1/4), \mathbf{v}^* = (1/2, 1/2, 1/2, 1/2, 1/2)$$

$$AGM = N^{5/2}; \quad A_{12} = A_{23} = \dots = [N^{1/2}] \times [N^{1/2}]$$

Loomis-Whitney:

$$\boxed{A_1(x_2, x_3, x_4, x_5) \wedge A_2(x_1, x_3, x_4, x_5) \wedge \dots \wedge A_5(x_1, x_2, x_3, x_4)}$$

Examples

$$L_5: \boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$$

$$\mathbf{w}^* = (1, 1, 0, 1), \mathbf{v}^* = (1, 0, 1, 0, 1).$$

$$AGM = N^3, \quad A_1, \dots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$$

$$C_5: \boxed{A_{12}(x_1, x_2) \wedge A_{23}(x_2, x_3) \wedge A_{34}(x_3, x_4) \wedge A_{45}(x_4, x_5) \wedge A_{51}(x_5, x_1)}$$

$$\mathbf{w}^* = (1/2, \dots, 1/2), \mathbf{v}^* = (1/2, \dots, 1/2).$$

$$AGM = N^{5/2}; \quad A_{12} = A_{23} = \dots = [N^{1/2}] \times [N^{1/2}]$$

$$K_5: \boxed{\bigwedge_{1 \leq i < j \leq 5} A_{ij}(x_i, x_j)}$$

$$\mathbf{w}^* = (1/4, \dots, 1/4), \mathbf{v}^* = (1/2, 1/2, 1/2, 1/2, 1/2)$$

$$AGM = N^{5/2}; \quad A_{12} = A_{23} = \dots = [N^{1/2}] \times [N^{1/2}]$$

Loomis-Whitney:

$$\boxed{A_1(x_2, x_3, x_4, x_5) \wedge A_2(x_1, x_3, x_4, x_5) \wedge \dots \wedge A_5(x_1, x_2, x_3, x_4)}$$

$$AGM = N^{5/4}, \quad A_1 = A_2 = \dots = [N^{1/4}] \times [N^{1/4}] \times [N^{1/4}] \times [N^{1/4}]$$

Arbitrary Cardinalities

- Each relation has a different cardinality $|R|, |S|, \dots$
- AGM is no longer N^{ρ^*} , but some function of $|R|, |S|, \dots$
- Need to consider multiple fractional vertex cover: AGM is a $\min(\dots)$.
- In practice: the AGM is given by a linear optimization problem, which generalizes the fractional edge cover/vertex packing.

Arbitrary Cardinalities: the Primal/Dual LPs

$$Q(\mathbf{x}) = \bigwedge_j R_j(\mathbf{x}_j)$$

Full CQ with m relations, n variables

Upper bound:

Minimize $\sum_j w_j \log |R_j|$ where:

$$\forall i = 1, n : \quad \sum_{j: x_i \in \text{Vars}(R_j)} w_j \geq 1$$

$$w_j \geq 0$$

Forall \mathbf{w} : $|Q| \leq \prod_j |R_j|^{w_j}$.

Lower bound:

Maximize $\sum_i v_i$ where:

$$\forall j = 1, m : \quad \sum_{i: x_i \in \text{Vars}(R_j)} v_i \leq \log |R_j|$$

$$v_i \geq 0$$

Forall \mathbf{v} , $\exists \text{DB s.t. } |Q| \geq \frac{1}{2^n} 2^{\sum_i v_i}$.

Weak duality: $\sum_j w_j \log |R_j| \geq \sum_i v_i$.

Strong duality: $\min_{\mathbf{w}} \sum_j w_j \log |R_j| = \max_{\mathbf{v}} \sum_i v_i \stackrel{\text{def}}{=} \log(\text{AGM})$

Discussion

- AGM bound is “tight”: factor $\frac{1}{2^{|\text{Vars}(Q)|}}$, often much better.
- Uses only cardinalities: extension only to **simple** FDs.
- No need for entropies yet.
- AGM bound is computable in PTIME in the size of Q .

Entropic Vectors

Motivation

- Extend the AGM bound to more statistics.
- Use in reasoning about constraints (next lecture).

Entropy, Entropic Vector

Entropy of a finite random variable:
$$h(X) \stackrel{\text{def}}{=} - \sum_i p_i \log p_i$$

Entropic vector defined by n random variables: $(h(\mathbf{X}_S))_{S \subseteq [n]} \in \mathbb{R}_+^{2^n}$

Derived quantities:

Conditional Entropy:

Chain rule:

$$\begin{aligned} h(\mathbf{V}|\mathbf{U}) &\stackrel{\text{def}}{=} h(\mathbf{UV}) - h(\mathbf{U}) \\ h(\mathbf{U}) + h(\mathbf{V}|\mathbf{V}) &= h(\mathbf{UV}) \end{aligned}$$

Conditional Mutual Information:

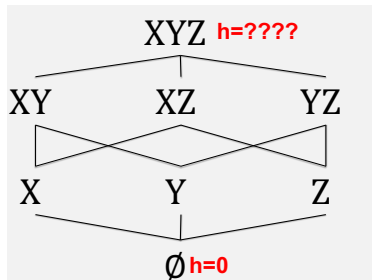
$$I_h(\mathbf{V}; \mathbf{W}|\mathbf{U}) \stackrel{\text{def}}{=} h(\mathbf{UV}) + h(\mathbf{UW}) - h(\mathbf{UVW}) - h(\mathbf{U})$$

Example: The Parity Function

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

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

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

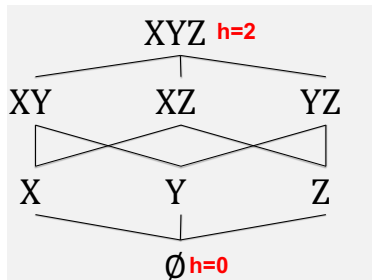


Example: The Parity Function

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

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

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

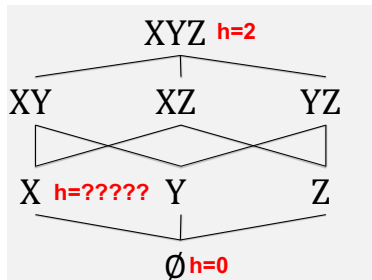


Example: The Parity Function

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

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

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

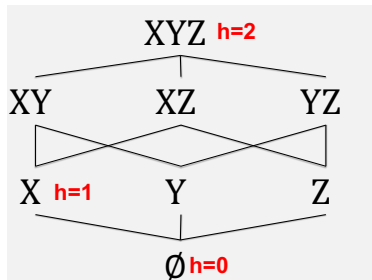


Example: The Parity Function

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

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

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

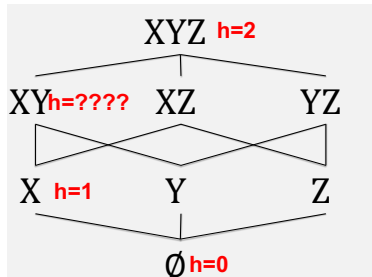


Example: The Parity Function

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

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

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

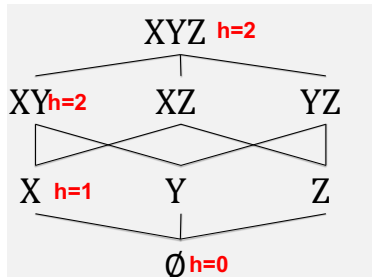


Example: The Parity Function

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

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

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

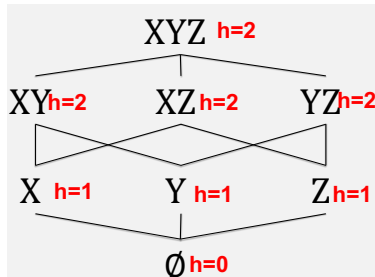


Example: The Parity Function

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

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

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

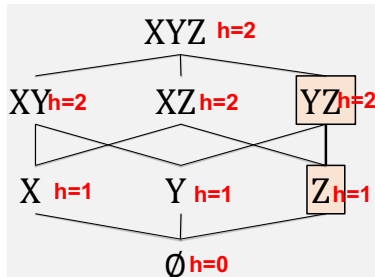


Example: The Parity Function

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

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

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



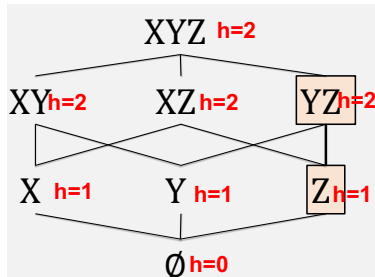
$$h(Y|Z) =$$

Example: The Parity Function

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

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

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



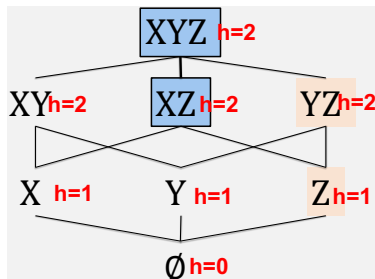
$$h(Y|Z) = 1$$

Example: The Parity Function

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

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

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



$$h(Y|Z) = 1$$

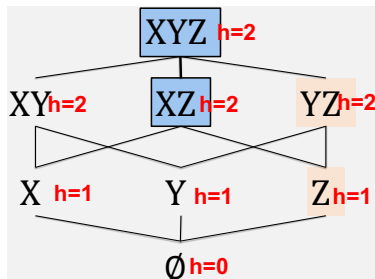
$$h(Y|XZ) =$$

Example: The Parity Function

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

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

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



$$h(Y|Z) = 1$$

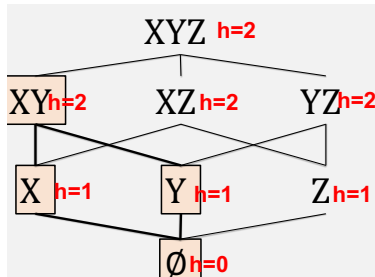
$$h(Y|XZ) = 0 \text{ Always decreases}$$

Example: The Parity Function

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

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

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



$$h(Y|Z) = 1$$

$$h(Y|XZ) = 0 \text{ Always decreases}$$

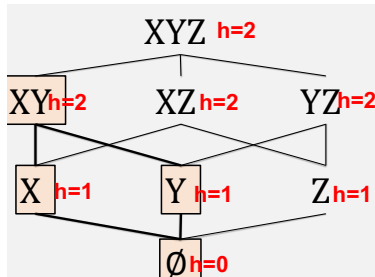
$$I_h(X; Y|\emptyset) =$$

Example: The Parity Function

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

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

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



$$h(Y|Z) = 1$$

$$h(Y|XZ) = 0 \text{ Always decreases}$$

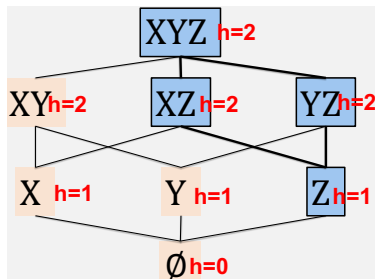
$$I_h(X; Y|\emptyset) = 0$$

Example: The Parity Function

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

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

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



$$h(Y|Z) = 1$$

$$h(Y|XZ) = 0 \text{ Always decreases}$$

$$I_h(X; Y|\emptyset) = 0$$

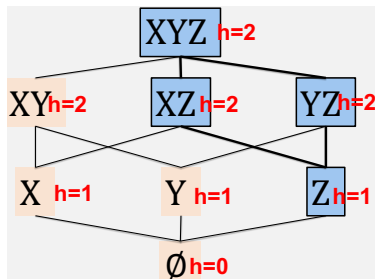
$$I_h(X; Y|Z) =$$

Example: The Parity Function

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

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

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



$$h(Y|Z) = 1$$

$$h(Y|XZ) = 0 \text{ Always decreases}$$

$$I_h(X; Y|\emptyset) = 0$$

$$I_h(X; Y|Z) = 1 \text{ May increase or decrease}$$

Properties of Entropic Vectors

Prove these in the Homework, using the definition $\sum p_i \log p_i$

- $0 \leq h(X) \leq \log N$
- Monotonicity: $h(\mathbf{U}) \leq h(\mathbf{UV})$
- Submodularity: $h(\mathbf{U}) + h(\mathbf{V}) \geq h(\mathbf{U} \cup \mathbf{V}) + h(\mathbf{U} \cap \mathbf{V})$.
- Conditional: $h(\mathbf{V}|\mathbf{U}) = \mathbb{E}_{\mathbf{u}}[h(\mathbf{V}|\mathbf{U} = \mathbf{u})]$
- Conditional Independence: $\mathbf{V} \perp \mathbf{W}|\mathbf{U}$ iff $I_h(\mathbf{V}; \mathbf{W}|\mathbf{U}) = 0$.

Once these are establish, we no longer need the definition $\sum p_i \log p_i$.

Information Inequalities v.s. Databases

Informally: $h(XY) \sim \log |\Pi_{XY}(R)|$. What do inequalities say about R ?

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Information Inequalities v.s. Databases

Informally: $h(XY) \sim \log |\Pi_{XY}(R)|$. What do inequalities say about R ?

- $h(X) \leq h(XY) \leq h(XYZ)$

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Information Inequalities v.s. Databases

Informally: $h(XY) \sim \log |\Pi_{XY}(R)|$. What do inequalities say about R ?

- $h(X) \leq h(XY) \leq h(XYZ)$
Says $|\Pi_X(R)| \leq |\Pi_{XY}(R)| \leq |R|$.

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Information Inequalities v.s. Databases

Informally: $h(XY) \sim \log |\Pi_{XY}(R)|$. What do inequalities say about R ?

- $h(X) \leq h(XY) \leq h(XYZ)$
Says $|\Pi_X(R)| \leq |\Pi_{XY}(R)| \leq |R|$.
- $h(XY) + h(Z) \geq h(XYZ)$

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Information Inequalities v.s. Databases

Informally: $h(XY) \sim \log |\Pi_{XY}(R)|$. What do inequalities say about R ?

- $h(X) \leq h(XY) \leq h(XYZ)$
Says $|\Pi_X(R)| \leq |\Pi_{XY}(R)| \leq |R|$.
- $h(XY) + h(Z) \geq h(XYZ)$
Says $|\Pi_{XY}(R)| \cdot |\Pi_Z(R)| \geq |R|$.

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Information Inequalities v.s. Databases

Informally: $h(XY) \sim \log |\Pi_{XY}(R)|$. What do inequalities say about R ?

- $h(X) \leq h(XY) \leq h(XYZ)$
Says $|\Pi_X(R)| \leq |\Pi_{XY}(R)| \leq |R|$.
- $h(XY) + h(Z) \geq h(XYZ)$
Says $|\Pi_{XY}(R)| \cdot |\Pi_Z(R)| \geq |R|$.
- $h(XYZ|X) \geq h(XYZ|XY)$

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Information Inequalities v.s. Databases

Informally: $h(XY) \sim \log |\Pi_{XY}(R)|$. What do inequalities say about R ?

- $h(X) \leq h(XY) \leq h(XYZ)$
Says $|\Pi_X(R)| \leq |\Pi_{XY}(R)| \leq |R|$.
- $h(XY) + h(Z) \geq h(XYZ)$
Says $|\Pi_{XY}(R)| \cdot |\Pi_Z(R)| \geq |R|$.
- $h(XYZ|X) \geq h(XYZ|XY)$
Max frequency(X) is \geq max frequency(XY).

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Information Inequalities v.s. Databases

Informally: $h(XY) \sim \log |\Pi_{XY}(R)|$. What do inequalities say about R ?

- $h(X) \leq h(XY) \leq h(XYZ)$
Says $|\Pi_X(R)| \leq |\Pi_{XY}(R)| \leq |R|$.
- $h(XY) + h(Z) \geq h(XYZ)$
Says $|\Pi_{XY}(R)| \cdot |\Pi_Z(R)| \geq |R|$.
- $h(XYZ|X) \geq h(XYZ|XY)$
Max frequency(X) is \geq max frequency(XY).
- **Careful!** $h(XZ) + h(YZ) \geq h(XYZ) + h(Z)$,
but $|\Pi_{XZ}(R)| \cdot |\Pi_{YZ}(R)| \not\geq |R| \cdot |\Pi_Z(R)|$

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Information Inequalities v.s. Databases

Informally: $h(XY) \sim \log |\Pi_{XY}(R)|$. What do inequalities say about R ?

- $h(X) \leq h(XY) \leq h(XYZ)$
Says $|\Pi_X(R)| \leq |\Pi_{XY}(R)| \leq |R|$.
- $h(XY) + h(Z) \geq h(XYZ)$
Says $|\Pi_{XY}(R)| \cdot |\Pi_Z(R)| \geq |R|$.
- $h(XYZ|X) \geq h(XYZ|XY)$
Max frequency(X) is \geq max frequency(XY).
- **Careful!** $h(XZ) + h(YZ) \geq h(XYZ) + h(Z)$,
but $\underbrace{|\Pi_{XZ}(R)|}_3 \cdot \underbrace{|\Pi_{YZ}(R)|}_3 \not\geq \underbrace{|R|}_5 \cdot \underbrace{|\Pi_Z(R)|}_2$

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Discussion

- We view entropies as a vector in $\mathbb{R}_+^{2^{[n]}}$.
- After you do the homework: forget the formula $\sum p_i \log p_i$, but remember its (simple!) consequences.
- We use entropies to compute query upper bounds (next), and to reason about database constraints (later).

Generalized Query Upper Bound

Motivation

- The AGM bound uses only cardinalities. Massive overapproximation, e.g. join $R(X, Y) \bowtie S(Y, Z)$.
- To use additional statistics (max degrees, ℓ_p -norms) we need to rely on information inequalities.

Recap: From Statistics to Upper Bound

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

Given an input instance $\mathbf{D} = (R^D, S^D, T^D)$,
define the uniform distribution on the output $Q(\mathbf{D})$:

$$Q(\mathbf{D}) =$$

X	Y	Z	p
a	b	c	$1/ Q $
a	b	d	$1/ Q $
	...		

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

Expressing Statistics Using the Entropy Vector

For any probability distribution on $R(X, Y)$, its entropy satisfies:

- $\boxed{h(XY) \leq \log |R|}.$
- $\boxed{h(Y|X) \leq \log \max \deg_R(Y|X)}.$
- For $p \in \mathbb{N}$, $p \geq 1$: $\boxed{h(X) + p \cdot h(Y|X) \leq \log \|\deg_R(Y|X)\|_p^p}$
(This is not obvious! Exercise)

This generalizes naturally to more attributes: $R(X, Y, Z, \dots)$

Example of Statistics:

 $R =$

$$\deg_R(VW|U) = (4, 2, 1)$$

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

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
c	1	m

$$\deg_R(VW|U) = (4, 2, 1)$$

$$h(VW|U) \leq \log \max \deg_R(VW|U) = \log 4$$

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
c	1	m

$$\deg_R(VW|U) = (4, 2, 1)$$

$$h(VW|U) \leq \log \max \deg_R(VW|U) = \log 4$$

$$\|\deg_R(VW|U)\|_2^2 = 4^2 + 2^2 + 1^2 = 21$$

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
c	1	m

$$\deg_R(VW|U) = (4, 2, 1)$$

$$h(VW|U) \leq \log \max \deg_R(VW|U) = \log 4$$

$$\|\deg_R(VW|U)\|_2^2 = 4^2 + 2^2 + 1^2 = 21$$

$$h(U) + 2 \cdot h(VW|U) \leq \log \|\deg_R(VW|U)\|_2^2 = \log 21$$

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
c	1	m

$$\deg_R(VW|U) = (4, 2, 1)$$

$$h(VW|U) \leq \log \max \deg_R(VW|U) = \log 4$$

$$\|\deg_R(VW|U)\|_2^2 = 4^2 + 2^2 + 1^2 = 21$$

$$h(U) + 2 \cdot h(VW|U) \leq \log \|\deg_R(VW|U)\|_2^2 = \log 21$$

$$\deg_R(V|U) = (3, 2, 1)$$

...

Example: Upper Bound with Max Degrees or FDs

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

Assume $|R| = |S| = |T| = N$, $|A| = |B| = \infty$

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

Example: Upper Bound with Max Degrees or FDs

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

Assume $|R| = |S| = |T| = N$, $|A| = |B| = \infty$

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

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

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

Example: Upper Bound with Max Degrees or FDs

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

Assume $|R| = |S| = |T| = N$, $|A| = |B| = \infty$

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

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: Upper Bound with Max Degrees or FDs

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

Assume $|R| = |S| = |T| = N$, $|A| = |B| = \infty$

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

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: Upper Bound with Max Degrees or FDs

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

Assume $|R| = |S| = |T| = N$, $|A| = |B| = \infty$

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

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: Upper Bound with Max Degrees or FDs

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

Assume $|R| = |S| = |T| = N$, $|A| = |B| = \infty$

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

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: Upper Bound with Max Degrees or FDs

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

Assume $|R| = |S| = |T| = N$, $|A| = |B| = \infty$

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

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: Upper Bound with Max Degrees or FDs

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

Assume $|R| = |S| = |T| = N$, $|A| = |B| = \infty$

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

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: Upper Bound with Max Degrees or FDs

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

Assume $|R| = |S| = |T| = N$, $|A| = |B| = \infty$

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

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: Upper Bound with Max Degrees or FDs

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

Assume $|R| = |S| = |T| = N$, $|A| = |B| = \infty$

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

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

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

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

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

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

Example: Upper Bound with ℓ_p -Norm of the Degrees

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

$$\text{Then } |Q| \leq (\|\deg_R(Y|X)\|_2^2 \cdot \|\deg_S(Z|Y)\|_2^2 \cdot \|\deg_T(X|Z)\|_2^2)^{1/3}.$$

Example: Upper Bound with ℓ_p -Norm of the Degrees

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

$$\text{Then } |Q| \leq (||\deg_R(Y|X)||_2^2 \cdot ||\deg_S(Z|Y)||_2^2 \cdot ||\deg_T(X|Z)||_2^2)^{1/3}.$$

Proof:

$$\log ||\deg_R(Y|X)||_2^2 + \log ||\deg_S(Z|Y)||_2^2 + \log ||\deg_T(X|Z)||_2^2 \geq$$

Example: Upper Bound with ℓ_p -Norm of the Degrees

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

$$\text{Then } |Q| \leq (||\deg_R(Y|X)||_2^2 \cdot ||\deg_S(Z|Y)||_2^2 \cdot ||\deg_T(X|Z)||_2^2)^{1/3}.$$

Proof:

$$\begin{aligned} \log ||\deg_R(Y|X)||_2^2 + \log ||\deg_S(Z|Y)||_2^2 + \log ||\deg_T(X|Z)||_2^2 &\geq \\ &\geq h(X) + 2h(Y|X) + h(Y) + 2h(Z|Y) + h(Z) + 2h(X|Z) \end{aligned}$$

Example: Upper Bound with ℓ_p -Norm of the Degrees

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

$$\text{Then } |Q| \leq (||\deg_R(Y|X)||_2^2 \cdot ||\deg_S(Z|Y)||_2^2 \cdot ||\deg_T(X|Z)||_2^2)^{1/3}.$$

Proof:

$$\begin{aligned} \log ||\deg_R(Y|X)||_2^2 + \log ||\deg_S(Z|Y)||_2^2 + \log ||\deg_T(X|Z)||_2^2 &\geq \\ &\geq h(X) + 2h(Y|X) + h(Y) + 2h(Z|Y) + h(Z) + 2h(X|Z) \\ &= h(XY) + h(Y|X) + h(YZ) + h(Z|Y) + h(XZ) + h(X|Z) \end{aligned}$$

Example: Upper Bound with ℓ_p -Norm of the Degrees

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

$$\text{Then } |Q| \leq (||\deg_R(Y|X)||_2^2 \cdot ||\deg_S(Z|Y)||_2^2 \cdot ||\deg_T(X|Z)||_2^2)^{1/3}.$$

Proof:

$$\begin{aligned} & \log ||\deg_R(Y|X)||_2^2 + \log ||\deg_S(Z|Y)||_2^2 + \log ||\deg_T(X|Z)||_2^2 \geq \\ & \geq h(X) + 2h(Y|X) + h(Y) + 2h(Z|Y) + h(Z) + 2h(X|Z) \\ & = h(XY) + \underline{h(Y|X)} + h(YZ) + \underline{h(Z|Y)} + h(XZ) + \underline{h(X|Z)} \\ & \geq h(XY) + h(Y|XZ) + h(YZ) + h(Z|XY) + h(XZ) + h(X|YZ) \end{aligned}$$

Example: Upper Bound with ℓ_p -Norm of the Degrees

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

$$\text{Then } |Q| \leq (||\deg_R(Y|X)||_2^2 \cdot ||\deg_S(Z|Y)||_2^2 \cdot ||\deg_T(X|Z)||_2^2)^{1/3}.$$

Proof:

$$\begin{aligned} & \log ||\deg_R(Y|X)||_2^2 + \log ||\deg_S(Z|Y)||_2^2 + \log ||\deg_T(X|Z)||_2^2 \geq \\ & \geq h(X) + 2h(Y|X) + h(Y) + 2h(Z|Y) + h(Z) + 2h(X|Z) \\ & = h(XY) + \underline{h(Y|X)} + h(YZ) + \underline{h(Z|Y)} + h(XZ) + \underline{h(X|Z)} \\ & \geq h(XY) + h(Y|XZ) + h(YZ) + h(Z|XY) + h(XZ) + h(X|YZ) \\ & = 3h(XYZ) = 3 \log |Q| \end{aligned}$$

Discussion

- Current systems: use cardinalities, average degrees.
- Upper bound: uses cardinalities, max degrees, and ℓ_p -norms.

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

$$|Q| \leq \|\deg_R(X|Y)\|_2 \cdot \|\deg_S(Z|Y)\|_2$$

$$\text{for all } p, q \geq 2: |Q| \leq \|\deg_R(X|Y)\|_p \cdot |\text{Dom}(Y)|^{1 - \frac{1}{p} - \frac{1}{q}} \cdot \|\deg_S(Z|Y)\|_q$$

- Predicates (equality, range, like) don't require new math, but lots of engineering to incorporate these stats into histograms.

Computing the Upper Bound

Motivation

- The AGM bound is defined by a linear optimization program, is computed in PTIME, and is tight.
- How do we compute the generalized upper bound?
Using an exponential-size linear optimization program.
- Is it tight? **Yes** for practical queries, **no** in general.

The Linear Program

$Q(\mathbf{X}) = \bigwedge_j R_j(\mathbf{X}_j)$, m atoms, n variables.

Construct the following linear program:

- There are 2^n variables, denoted $h(\mathbf{U})$ for every $\mathbf{U} \subseteq \mathbf{X}$.

The Linear Program

$Q(\mathbf{X}) = \bigwedge_j R_j(\mathbf{X}_j)$, m atoms, n variables.

Construct the following linear program:

- There are 2^n variables, denoted $h(\mathbf{U})$ for every $\mathbf{U} \subseteq \mathbf{X}$.
- For each stats add the corresponding constraint:

$$h(\mathbf{X}_j) \leq \log |R_j|$$

$$h(\mathbf{V}|\mathbf{U}) \leq \log \max \deg(\mathbf{V}|\mathbf{U})$$

$$h(\mathbf{U}) + ph(\mathbf{V}|\mathbf{U}) \leq \log \|\deg(\mathbf{V}|\mathbf{U})\|_p^p$$

The Linear Program

$Q(\mathbf{X}) = \bigwedge_j R_j(\mathbf{X}_j)$, m atoms, n variables.

Construct the following linear program:

- There are 2^n variables, denoted $h(\mathbf{U})$ for every $\mathbf{U} \subseteq \mathbf{X}$.
- For each stats add the corresponding constraint:

$$h(\mathbf{X}_j) \leq \log |R_j|$$

$$h(\mathbf{V}|\mathbf{U}) \leq \log \max \deg(\mathbf{V}|\mathbf{U})$$

$$h(\mathbf{U}) + p h(\mathbf{V}|\mathbf{U}) \leq \log \|\deg(\mathbf{V}|\mathbf{U})\|_p^p$$

- Add all Shannon inequalities as constraints:

$$-h(XY) - h(YZ) + h(XYZ) + h(Y) \leq 0$$

...

The Linear Program

$Q(\mathbf{X}) = \bigwedge_j R_j(\mathbf{X}_j)$, m atoms, n variables.

Construct the following linear program:

- There are 2^n variables, denoted $h(\mathbf{U})$ for every $\mathbf{U} \subseteq \mathbf{X}$.
- For each stats add the corresponding constraint:

$$h(\mathbf{X}_j) \leq \log |R_j|$$

$$h(\mathbf{V}|\mathbf{U}) \leq \log \max \deg(\mathbf{V}|\mathbf{U})$$

$$h(\mathbf{U}) + p h(\mathbf{V}|\mathbf{U}) \leq \log \|\deg(\mathbf{V}|\mathbf{U})\|_p^p$$

- Add all Shannon inequalities as constraints:

$$-h(XY) - h(YZ) + h(XYZ) + h(Y) \leq 0$$

...

- Maximize $h(\mathbf{X})$.

Example

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

Maximize $h(XYZ)$, where:

Example

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

Maximize $h(XYZ)$, where:

$c_1 :$ $h(XY) \leq \log |R|$

Example

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

Maximize $h(XYZ)$, where:

$$c_1 : h(XY) \leq \log |R|$$

$$c_2 : h(YZ) \leq \log |S|$$

Example

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

Maximize $h(XYZ)$, where:

$$c_1 : h(XY) \leq \log |R|$$

$$c_2 : h(YZ) \leq \log |S|$$

$$c_3 : h(XZ) \leq \log |T|$$

Example

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

Maximize $h(XYZ)$, where:

$$c_1 : h(XY) \leq \log |R|$$

$$c_2 : h(YZ) \leq \log |S|$$

$$c_3 : h(XZ) \leq \log |T|$$

$$\begin{aligned} \sigma_1 : & -h(XY) - h(YZ) \\ & +h(XYZ) + h(Y) \leq 0 \end{aligned}$$

Example

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

Maximize $h(XYZ)$, where:

$$c_1 : h(XY) \leq \log |R|$$

$$c_2 : h(YZ) \leq \log |S|$$

$$c_3 : h(XZ) \leq \log |T|$$

$$\sigma_1 : \begin{aligned} & -h(XY) - h(YZ) \\ & +h(XYZ) + h(Y) \leq 0 \end{aligned}$$

$$\sigma_2 : \begin{aligned} & -h(Y) - h(XZ) \\ & +h(XYZ) \leq 0 \end{aligned}$$

...

Example

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

Dual:

Maximize $h(XYZ)$, where:

$$c_1 : h(XY) \leq \log |R|$$

$$c_2 : h(YZ) \leq \log |S|$$

$$c_3 : h(XZ) \leq \log |T|$$

$$\sigma_1 : \begin{aligned} & -h(XY) - h(YZ) \\ & +h(XYZ) + h(Y) \leq 0 \end{aligned}$$

$$\sigma_2 : \begin{aligned} & -h(Y) - h(XZ) \\ & +h(XYZ) \leq 0 \end{aligned}$$

...

$$\sigma_{18} : \dots \leq 0$$

Example

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

Primal:

Minimize $c_1 \log |R| + c_2 \log |S| + c_3 \log |T|$
where:

Dual:

Maximize $h(XYZ)$, where:

$$c_1 : \quad h(XY) \leq \log |R|$$

$$c_2 : \quad h(YZ) \leq \log |S|$$

$$c_3 : \quad h(XZ) \leq \log |T|$$

$$\begin{aligned} \sigma_1 : \quad & -h(XY) - h(YZ) \\ & + h(XYZ) + h(Y) \leq 0 \end{aligned}$$

$$\begin{aligned} \sigma_2 : \quad & -h(Y) - h(XZ) \\ & + h(XYZ) \leq 0 \end{aligned}$$

...

$$\sigma_{18} : \quad \dots \leq 0$$

Example

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

Primal:

Minimize $c_1 \log |R| + c_2 \log |S| + c_3 \log |T|$
where:

$$h(XYZ) : \quad \sigma_1 + \sigma_2 + \dots \geq 1$$

Dual:

Maximize $h(XYZ)$, where:

$$c_1 : \quad h(XY) \leq \log |R|$$

$$c_2 : \quad h(YZ) \leq \log |S|$$

$$c_3 : \quad h(XZ) \leq \log |T|$$

$$\begin{aligned} \sigma_1 : \quad & -h(XY) - h(YZ) \\ & + h(XYZ) + h(Y) \leq 0 \end{aligned}$$

$$\begin{aligned} \sigma_2 : \quad & -h(Y) - h(XZ) \\ & + h(XYZ) \leq 0 \end{aligned}$$

...

$$\sigma_{18} : \quad \dots \leq 0$$

Example

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

Primal:

Minimize $c_1 \log |R| + c_2 \log |S| + c_3 \log |T|$
where:

$$h(XYZ) : \quad \sigma_1 + \sigma_2 + \dots \geq 1$$

$$h(XY) : \quad c_1 - \sigma_1 + \dots \geq 0$$

$$h(YZ) : \quad c_2 - \sigma_1 + \dots \geq 0$$

$$h(XZ) : \quad c_3 - \sigma_2 + \dots \geq 0$$

Dual:

Maximize $h(XYZ)$, where:

$$c_1 : \quad h(XY) \leq \log |R|$$

$$c_2 : \quad h(YZ) \leq \log |S|$$

$$c_3 : \quad h(XZ) \leq \log |T|$$

$$\sigma_1 : \quad -h(XY) - h(YZ) \\ + h(XYZ) + h(Y) \leq 0$$

$$\sigma_2 : \quad -h(Y) - h(XZ) \\ + h(XYZ) \leq 0$$

...

$$\sigma_{18} : \quad \dots \leq 0$$

Example

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

Primal:

Minimize $c_1 \log |R| + c_2 \log |S| + c_3 \log |T|$
where:

$$h(XYZ) : \sigma_1 + \sigma_2 + \dots \geq 1$$

$$h(XY) : c_1 - \sigma_1 + \dots \geq 0$$

$$h(YZ) : c_2 - \sigma_1 + \dots \geq 0$$

$$h(XZ) : c_3 - \sigma_2 + \dots \geq 0$$

$$h(X) : \dots \geq 0$$

$$h(Y) : \sigma_1 - \sigma_2 + \dots \geq 0$$

$$h(Z) : \dots \geq 0$$

Dual:

Maximize $h(XYZ)$, where:

$$c_1 : h(XY) \leq \log |R|$$

$$c_2 : h(YZ) \leq \log |S|$$

$$c_3 : h(XZ) \leq \log |T|$$

$$\sigma_1 : -h(XY) - h(YZ) + h(XYZ) + h(Y) \leq 0$$

$$\sigma_2 : -h(Y) - h(XZ) + h(XYZ) \leq 0$$

...

$$\sigma_{18} : \dots \leq 0$$

Example

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

Primal:

Minimize $c_1 \log |R| + c_2 \log |S| + c_3 \log |T|$
where:

$$h(XYZ) : \quad \sigma_1 + \sigma_2 + \dots \geq 1$$

$$h(XY) : \quad c_1 - \sigma_1 + \dots \geq 0$$

$$h(YZ) : \quad c_2 - \sigma_1 + \dots \geq 0$$

$$h(XZ) : \quad c_3 - \sigma_2 + \dots \geq 0$$

$$h(X) : \quad \dots \geq 0$$

$$h(Y) : \quad \sigma_1 - \sigma_2 + \dots \geq 0$$

$$h(Z) : \quad \dots \geq 0$$

Dual:

Maximize $h(XYZ)$, where:

$$c_1 : \quad h(XY) \leq \log |R|$$

$$c_2 : \quad h(YZ) \leq \log |S|$$

$$c_3 : \quad h(XZ) \leq \log |T|$$

$$\sigma_1 : \quad -h(XY) - h(YZ) \\ + h(XYZ) + h(Y) \leq 0$$

$$\sigma_2 : \quad -h(Y) - h(XZ) \\ + h(XYZ) \leq 0$$

...

$$\sigma_{18} : \quad \dots \leq 0$$

Correctness: any feasible solution $c_1, c_2, c_3, \sigma_1, \dots, \sigma_{18}$ of the primal defines a Shannon inequality $c_1 h(XY) + c_2 h(YZ) + c_3 h(XZ) \geq h(XYZ)$.

Correctness Proof – Will Skip This Slide

Theorem

Any feasible solution $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \sigma_1, \dots, \sigma_{18}$ of the primal defines a Shannon inequality $\mathbf{c}_1 h(XY) + \mathbf{c}_2 h(YZ) + \mathbf{c}_3 h(XZ) \geq h(XYZ)$.

Proof: Multiply each inequality with its h -term and add them:

$$h(XYZ)(\sigma_1 + \dots) + h(XY)(\mathbf{c}_1 - \sigma_1 + \dots) + \dots \geq h(XYZ)$$

Group by the coefficients $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \sigma_1, \sigma_2, \dots$

$$\mathbf{c}_1 h(XY) + \mathbf{c}_2 h(YZ) + \mathbf{c}_3 h(XZ) + \sigma_1(\dots) + \dots \geq h(XYZ)$$

By design, the co-factor of σ_i is the LHS of a Shannon inequality,

$$\text{e.g. } \sigma_1(-h(XY) - h(YZ) + h(XYZ) + h(Y))$$

Shannon inequalities $-h(XY) - h(YZ) + h(XYZ) + h(Y) \leq 0$ imply:

$$\mathbf{c}_1 h(XY) + \mathbf{c}_2 h(YZ) + \mathbf{c}_3 h(XZ) \geq h(XYZ)$$

Discussion

AGM bound:

- Primal: a frac. edge cover, upper bound $|Q| \leq \dots$
- Dual: a frac. vertex cover, worst case database instance.

Discussion

AGM bound:

- Primal: a frac. edge cover, upper bound $|Q| \leq \dots$
- Dual: a frac. vertex cover, worst case database instance.

General bound:

- Primal: upper bound $\log |Q| \leq c_1 \log |R| + c_2 \log \max \deg(Y|X) + \dots$
- Dual: worst-case vector $\mathbf{h} \in \mathbb{R}_+^{2^n}$; but no database instance in general.

Discussion

AGM bound:

- Primal: a frac. edge cover, upper bound $|Q| \leq \dots$
- Dual: a frac. vertex cover, worst case database instance.

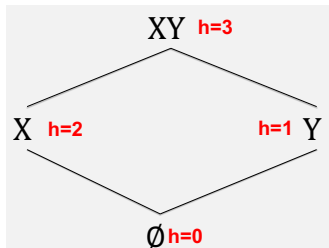
General bound:

- Primal: upper bound $\log |Q| \leq c_1 \log |R| + c_2 \log \max \deg(Y|X) + \dots$
- Dual: worst-case vector $\mathbf{h} \in \mathbb{R}_+^{2^n}$; but no database instance in general.
- Special case: all stats are cardinalities, then \mathbf{h} is **modular**; \mathbf{h} defines a worst-case **product database**. **Homework**
- Special case: all degree sequences are **simple**, then \mathbf{h} is **normal**; \mathbf{h} defines a worst-case **normal database** [Suciu, 2023].

Modular Functions

$h \in \mathbb{R}_+^{2^n}$ is called *modular* if $h(\mathbf{U}) + h(\mathbf{V}) = h(\mathbf{UV})$ for all $\mathbf{U} \cap \mathbf{V} = \emptyset$.

X	Y	p
1	a	$1/8$
1	b	$1/8$
2	a	$1/8$
2	b	$1/8$
3	a	$1/8$
3	b	$1/8$
4	a	$1/8$
4	b	$1/8$



h is **modular** iff it is the entropic vector of n independent random variables

Discussion

On the homework:

- If all statistics are cardinality constraints (i.e. no conditionals $h(\mathbf{V}|\mathbf{U})$) then the dual LP has an optimal solution \mathbf{h} that is a **modular function**:
 - ▶ Can compute in PTIME (only n variables).
 - ▶ Can construct a product worst-case instance.
- This explains why the AGM is much simpler than the general case.

Not on the homework: if conditionals are **simple**: the dual has a **normal** optimal solution: need EXPTIME but admits a **domain-product** worst case instance (next lecture).



Atserias, A., Grohe, M., and Marx, D. (2013).

Size bounds and query plans for relational joins.

SIAM J. Comput., 42(4):1737–1767.



Suciu, D. (2023).

Applications of information inequalities to database theory problems.

In *LICS*, pages 1–30.