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 $\boldsymbol{w}$

Minimize $\sum_e w_e$, where:

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

Fractional Vertex Packing $\boldsymbol{v}$

Maximize $\sum_x v_x$, where:

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

Weak duality: $\sum_e w_e$

Fractional Edge Cover / Vertex Packing

Hypergraph $G = (V, E)$

Fractional Edge Cover $\boldsymbol{w}$

Minimize $\sum_e w_e$, where:

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

$$w_e \geq 0$$

Fractional Vertex Packing $\boldsymbol{v}$

Maximize $\sum_x v_x$, where:

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

$$v_x \geq 0$$

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 $\boldsymbol{w}$

Minimize $\sum_e w_e$, where:

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

$$w_e \geq 0$$

Fractional Vertex Packing $\boldsymbol{v}$

Maximize $\sum_x v_x$, where:

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

$$v_x \geq 0$$

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 $\boldsymbol{w}$

Minimize $\sum_e w_e$, where:

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

$$w_e \geq 0$$

Fractional Vertex Packing $\boldsymbol{v}$

Maximize $\sum_x v_x$, where:

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

$$v_x \geq 0$$

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 $\boldsymbol{w}$

Minimize $\sum_e w_e$, where:

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

Fractional Vertex Packing $\boldsymbol{v}$

Maximize $\sum_x v_x$, where:

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

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: $\min_{\boldsymbol{w}} \sum_e w_e = \min_{\boldsymbol{v}} \sum_x v_x \stackrel{\text{def}}{=} \rho^*$

Fractional edge covering number

The AGM Bound

[Atserias et al., 2013]

$Q(\boldsymbol{x}) = \bigwedge_j R_j(\boldsymbol{x}_j)$ 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 $\boldsymbol{v}^* =$ optimal vertex packing.

The AGM Bound

[Atserias et al., 2013]

$Q(\boldsymbol{x}) = \bigwedge_j R_j(\boldsymbol{x}_j)$ 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 $\boldsymbol{v}^*$ = optimal vertex packing.

The AGM Bound

[Atserias et al., 2013]

$Q(\boldsymbol{x}) = \bigwedge_j R_j(\boldsymbol{x}_j)$ 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 $\boldsymbol{v}^* =$ 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: $\boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$
$\boldsymbol{w}^* = (1, 1, 0, 1)$, $\boldsymbol{v}^* = (1, 0, 1, 0, 1)$.
$AGM = N^3, \quad A_1, \ldots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$

Examples

L_5: $\boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$
$\boldsymbol{w}^* = (1, 1, 0, 1)$, $\boldsymbol{v}^* = (1, 0, 1, 0, 1)$.
$AGM = N^3, \quad A_1, \ldots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$

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: $\boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$
$\boldsymbol{w}^* = (1, 1, 0, 1)$, $\boldsymbol{v}^* = (1, 0, 1, 0, 1)$.
$AGM = N^3, \quad A_1, \ldots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$

C_5: $\boxed{A_{12}(x_1, x_2) \wedge A_{23}(x_2, x_3) \wedge A_{34}(x_3, x_4) \wedge A_{45}(x_4, x_5) \wedge A_{51}(x_5, x_1)}$
$\boldsymbol{w}^* = (1/2, \ldots, 1/2)$, $\boldsymbol{v}^* = (1/2, \ldots, 1/2)$.
$AGM = N^{5/2}$; $A_{12} = A_{23} = \cdots = [N^{1/2}] \times [N^{1/2}]$

Examples

L_5: $\boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$
$\boldsymbol{w}^* = (1, 1, 0, 1)$, $\boldsymbol{v}^* = (1, 0, 1, 0, 1)$.
$AGM = N^3$, $\quad A_1, \ldots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$

C_5: $\boxed{A_{12}(x_1, x_2) \wedge A_{23}(x_2, x_3) \wedge A_{34}(x_3, x_4) \wedge A_{45}(x_4, x_5) \wedge A_{51}(x_5, x_1)}$
$\boldsymbol{w}^* = (1/2, \ldots, 1/2)$, $\boldsymbol{v}^* = (1/2, \ldots, 1/2)$.
$AGM = N^{5/2}$; $A_{12} = A_{23} = \cdots = [N^{1/2}] \times [N^{1/2}]$

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

Examples

L_5: $\boxed{A_1(x_1,x_2) \wedge A_2(x_2,x_3) \wedge A_3(x_3,x_4) \wedge A_4(x_4,x_5)}$
$\boldsymbol{w}^* = (1,1,0,1)$, $\boldsymbol{v}^* = (1,0,1,0,1)$.
$AGM = N^3$, $\quad A_1, \ldots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$

C_5: $\boxed{A_{12}(x_1,x_2) \wedge A_{23}(x_2,x_3) \wedge A_{34}(x_3,x_4) \wedge A_{45}(x_4,x_5) \wedge A_{51}(x_5,x_1)}$
$\boldsymbol{w}^* = (1/2, \ldots, 1/2)$, $\boldsymbol{v}^* = (1/2, \ldots, 1/2)$.
$AGM = N^{5/2}$; $A_{12} = A_{23} = \cdots = [N^{1/2}] \times [N^{1/2}]$

K_5: $\boxed{\bigwedge_{1 \leq i < j \leq 5} A_{ij}(x_i, x_j)}$
$\boldsymbol{w}^* = (1/4, \ldots, 1/4)$, $\boldsymbol{v}^* = (1/2, 1/2, 1/2, 1/2, 1/2)$
$AGM = N^{5/2}$; $A_{12} = A_{23} = \cdots = [N^{1/2}] \times [N^{1/2}]$

Examples

L_5: $\boxed{A_1(x_1, x_2) \wedge A_2(x_2, x_3) \wedge A_3(x_3, x_4) \wedge A_4(x_4, x_5)}$
$\boldsymbol{w}^* = (1, 1, 0, 1)$, $\boldsymbol{v}^* = (1, 0, 1, 0, 1)$.
$AGM = N^3, \quad A_1, \ldots, A_4 = [N] \times [1], \quad [1] \times [N], \quad [N] \times [1], \quad [1] \times [N]$

C_5: $\boxed{A_{12}(x_1, x_2) \wedge A_{23}(x_2, x_3) \wedge A_{34}(x_3, x_4) \wedge A_{45}(x_4, x_5) \wedge A_{51}(x_5, x_1)}$
$\boldsymbol{w}^* = (1/2, \ldots, 1/2)$, $\boldsymbol{v}^* = (1/2, \ldots, 1/2)$.
$AGM = N^{5/2}$; $A_{12} = A_{23} = \cdots = [N^{1/2}] \times [N^{1/2}]$

K_5: $\boxed{\bigwedge_{1 \leq i < j \leq 5} A_{ij}(x_i, x_j)}$
$\boldsymbol{w}^* = (1/4, \ldots, 1/4)$, $\boldsymbol{v}^* = (1/2, 1/2, 1/2, 1/2, 1/2)$
$AGM = N^{5/2}$; $A_{12} = A_{23} = \cdots = [N^{1/2}] \times [N^{1/2}]$

Loomis-Whitney:
$\boxed{A_1(x_2, x_3, x_4, x_5) \wedge A_2(x_1, x_3, x_4, x_5) \wedge \cdots \wedge A_5(x_1, x_2, x_3, x_4)}$

Examples

L_5: $\boxed{A_1(x_1,x_2) \wedge A_2(x_2,x_3) \wedge A_3(x_3,x_4) \wedge A_4(x_4,x_5)}$
$\boldsymbol{w}^* = (1,1,0,1)$, $\boldsymbol{v}^* = (1,0,1,0,1)$.
$AGM = N^3$, $\quad A_1,\ldots,A_4 = [N]\times[1], \quad [1]\times[N], \quad [N]\times[1], \quad [1]\times[N]$

C_5: $\boxed{A_{12}(x_1,x_2) \wedge A_{23}(x_2,x_3) \wedge A_{34}(x_3,x_4) \wedge A_{45}(x_4,x_5) \wedge A_{51}(x_5,x_1)}$
$\boldsymbol{w}^* = (1/2,\ldots,1/2)$, $\boldsymbol{v}^* = (1/2,\ldots,1/2)$.
$AGM = N^{5/2}$; $A_{12} = A_{23} = \cdots = [N^{1/2}]\times[N^{1/2}]$

K_5: $\boxed{\bigwedge_{1\leq i<j\leq 5} A_{ij}(x_i,x_j)}$
$\boldsymbol{w}^* = (1/4,\ldots,1/4)$, $\boldsymbol{v}^* = (1/2,1/2,1/2,1/2,1/2)$
$AGM = N^{5/2}$; $A_{12} = A_{23} = \cdots = [N^{1/2}]\times[N^{1/2}]$

Loomis-Whitney:
$\boxed{A_1(x_2,x_3,x_4,x_5) \wedge A_2(x_1,x_3,x_4,x_5) \wedge \cdots \wedge A_5(x_1,x_2,x_3,x_4)}$
$AGM = N^{5/4}$, $A_1 = A_2 = \cdots = [N^{1/4}]\times[N^{1/4}]\times[N^{1/4}]\times[N^{1/4}]$

Arbitrary Cardinalities

- Each relation has a different cardinality $|R|, |S|, \ldots$
- AGM is no longer N^{ρ^*}, but some function of $|R|, |S|, \ldots$
- Need to consider multiple fractional vertex cover: AGM is a $\min(\cdots)$.
- 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

$Q(\boldsymbol{x}) = \bigwedge_j R_j(\boldsymbol{x}_j)$ Full CQ with m relations, n variables

Upper bound:

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

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

$$w_j \geq 0$$

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

Lower bound:

Maximize $\sum_i v_i$ where:

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

$$v_i \geq 0$$

Forall $\boldsymbol{v}$, $\exists$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: $\min_{\boldsymbol{w}} \sum_j w_j \log |R_j| = \min_{\boldsymbol{v}} \sum_i v_i \stackrel{\text{def}}{=} \log (\text{AGM})$

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: $\boxed{h(X) \stackrel{\text{def}}{=} -\sum_i p_i \log p_i}$

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

Derived quantities:

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

Chain rule: $h(\boldsymbol{U}) + h(\boldsymbol{V}|\boldsymbol{V}) = h(\boldsymbol{UV})$

Conditional Mutual Information:

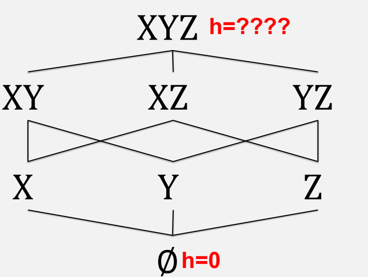
$$\boxed{I_h(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U}) \stackrel{\text{def}}{=} h(\boldsymbol{UV}) + h(\boldsymbol{UW}) - h(\boldsymbol{UVW}) - h(\boldsymbol{U})}$$

Example: The Parity Function

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

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

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

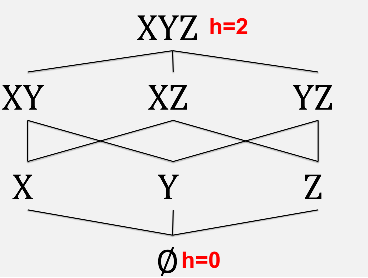


Example: The Parity Function

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

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

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

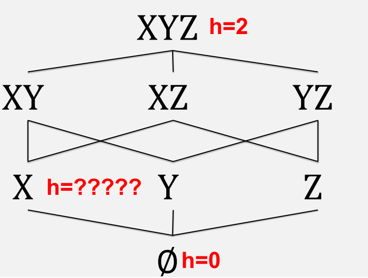


Example: The Parity Function

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

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

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

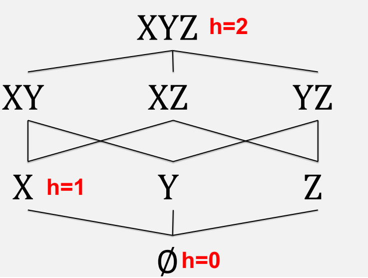


Example: The Parity Function

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

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

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

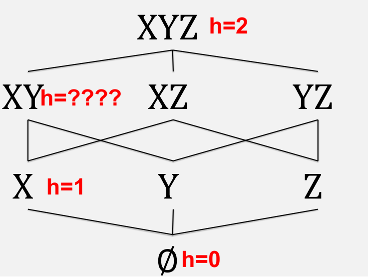


Example: The Parity Function

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

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

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

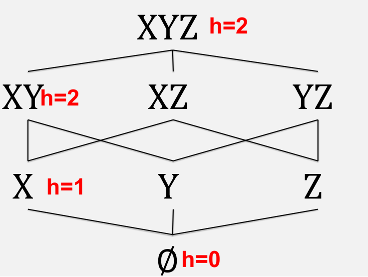


Example: The Parity Function

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

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

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

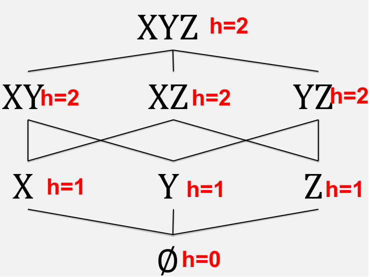


Example: The Parity Function

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

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

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

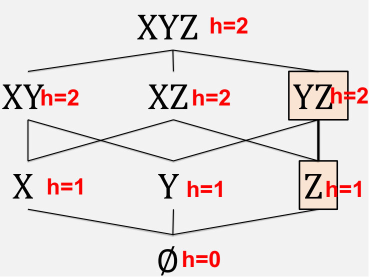


Example: The Parity Function

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

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

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



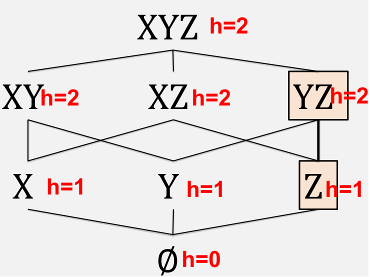
$h(YZ) =$

Example: The Parity Function

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

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

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



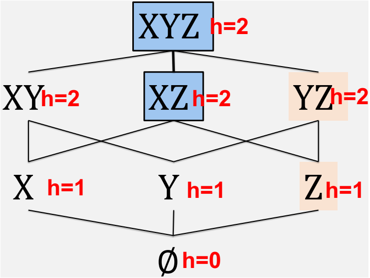
$h(YZ) = 1$

Example: The Parity Function

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

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

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



$h(YZ) = 1$

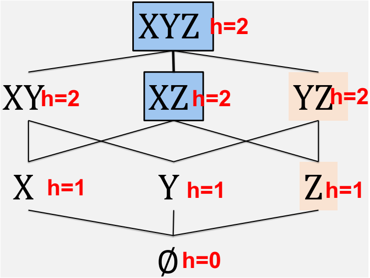
$h(Y|XZ) =$

Example: The Parity Function

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

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

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



$h(YZ) = 1$

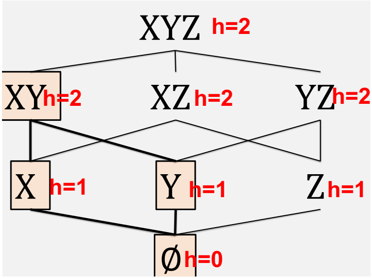
$h(Y|XZ) = 0$ Always decreases

Example: The Parity Function

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

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

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



$h(YZ) = 1$

$h(Y|XZ) = 0$ Always decreases

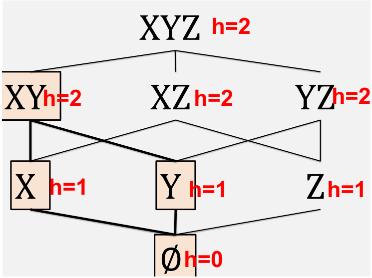
$I_h(X;Y|\emptyset) =$

Example: The Parity Function

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

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

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



$h(YZ) = 1$

$h(Y|XZ) = 0$ Always decreases

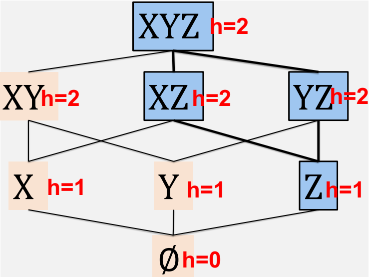
$I_h(X;Y|\emptyset) = 0$

Example: The Parity Function

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

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

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



$h(YZ) = 1$

$h(Y|XZ) = 0$ Always decreases

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

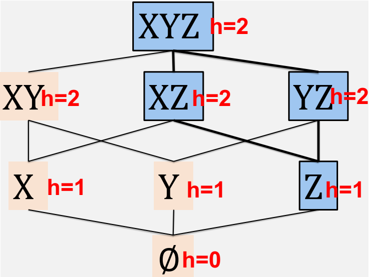
$I_h(X;Y|Z) =$

Example: The Parity Function

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

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

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



$h(YZ) = 1$

$h(Y|XZ) = 0$ Always decreases

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

$I_h(X;Y|Z) = 1$ 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: $h(\boldsymbol{U}) \leq h(\boldsymbol{UV})$
- Submodularity: $h(\boldsymbol{U}) + h(\boldsymbol{V}) \geq h(\boldsymbol{U} \cup \boldsymbol{V}) + h(\boldsymbol{U} \cap \boldsymbol{V})$.
- Conditional: $h(\boldsymbol{V}|\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{u}}[h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})]$
- Conditional Independence: $\boldsymbol{V} \perp \boldsymbol{W}|\boldsymbol{U}$ iff $I_h(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U}) = 0$.

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 $\boldsymbol{D} = (R^D, S^D, T^D)$, define the uniform distribution on the output $Q(\boldsymbol{D})$:

$Q(\boldsymbol{D}) =$

X	Y	Z	p
a	b	c	$1/\lvert Q\rvert$
a	b	d	$1/\lvert Q\rvert$
	$\dots$		

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

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: $R(X, Y, Z, \ldots)$

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

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

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

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

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

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

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

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

$$\ldots$$

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 $|R| = |S| = |T| = N$, $|A| = |B| = \infty$ $\quad AGM(Q) = N^2$.

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

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 $|R| = |S| = |T| = N$, $|A| = |B| = \infty$ $\qquad AGM(Q) = N^2$.
If the FDs $XZ \to U$ and $YU \to X$ hold: $\qquad |Q| \leq N^{3/2}$.

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

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

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 $|R| = |S| = |T| = N$, $|A| = |B| = \infty$ $\quad AGM(Q) = N^2$.

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

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

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

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 $|R| = |S| = |T| = N$, $|A| = |B| = \infty$ $\quad AGM(Q) = N^2$.

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

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

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

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

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 $|R| = |S| = |T| = N$, $|A| = |B| = \infty$ $\quad AGM(Q) = N^2$.

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

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

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

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

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 $|R| = |S| = |T| = N$, $|A| = |B| = \infty$ $\quad AGM(Q) = N^2$.

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

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

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

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

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

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

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 $|R| = |S| = |T| = N$, $|A| = |B| = \infty$ $\qquad AGM(Q) = N^2$.

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

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

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

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 $|R| = |S| = |T| = N$, $|A| = |B| = \infty$ $\qquad AGM(Q) = N^2$.

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

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

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

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

Example: Upper Bound with ℓ_p-Norm of the Degrees

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

Then $|Q| \leq \left(||\text{deg}_R(Y|X)||_2^2 \cdot ||\text{deg}_S(Z|Y)||_2^2 \cdot ||\text{deg}_T(X|Z)||_2^2\right)^{1/3}$.

Example: Upper Bound with ℓ_p-Norm of the Degrees

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$
Then $|Q| \leq \left(||\deg_R(Y|X)||_2^2 \cdot ||\deg_S(Z|Y)||_2^2 \cdot ||\deg_T(X|Z)||_2^2\right)^{1/3}$.

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

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$
Then $|Q| \leq \left(||\deg_R(Y|X)||_2^2 \cdot ||\deg_S(Z|Y)||_2^2 \cdot ||\deg_T(X|Z)||_2^2\right)^{1/3}$.

Proof:

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

Example: Upper Bound with ℓ_p-Norm of the Degrees

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$
Then $|Q| \leq \left(||\text{deg}_R(Y|X)||_2^2 \cdot ||\text{deg}_S(Z|Y)||_2^2 \cdot ||\text{deg}_T(X|Z)||_2^2\right)^{1/3}$.

Proof:

$$\begin{aligned}
\log ||\text{deg}_R(Y|X)||_2^2 + \log ||\text{deg}_S(Z|Y)||_2^2 + \log ||\text{deg}_T(X|Z)||_2^2 \geq \\
\geq & h(X) + 2h(Y|X) + h(Y) + 2h(Z|Y) + h(Z) + 2h(X|Z) \\
= & h(XY) + h(Y|X) + h(YZ) + h(Z|Y) + h(XZ) + h(X|Z)
\end{aligned}$$

Example: Upper Bound with ℓ_p-Norm of the Degrees

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$
Then $|Q| \leq \left(||\text{deg}_R(Y|X)||_2^2 \cdot ||\text{deg}_S(Z|Y)||_2^2 \cdot ||\text{deg}_T(X|Z)||_2^2\right)^{1/3}$.

Proof:

$$\begin{aligned}
\log ||\text{deg}_R(Y|X)||_2^2 + \log ||\text{deg}_S(Z|Y)||_2^2 + \log ||\text{deg}_T(X|Z)||_2^2 \geq \\
\geq & h(X) + 2h(Y|X) + h(Y) + 2h(Z|Y) + h(Z) + 2h(X|Z) \\
= & h(XY) + \underline{h(Y|X)} + h(YZ) + \underline{h(Z|Y)} + h(XZ) + \underline{h(X|Z)} \\
\geq & h(XY) + h(Y|XZ) + h(YZ) + h(Z|XY) + h(XZ) + h(X|YZ)
\end{aligned}$$

Example: Upper Bound with ℓ_p-Norm of the Degrees

$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X)$
Then $|Q| \leq \left(||\text{deg}_R(Y|X)||_2^2 \cdot ||\text{deg}_S(Z|Y)||_2^2 \cdot ||\text{deg}_T(X|Z)||_2^2\right)^{1/3}$.

Proof:

$$
\begin{aligned}
&\log ||\text{deg}_R(Y|X)||_2^2 + \log ||\text{deg}_S(Z|Y)||_2^2 + \log ||\text{deg}_T(X|Z)||_2^2 \geq \\
&\quad \geq h(X) + 2h(Y|X) + h(Y) + 2h(Z|Y) + h(Z) + 2h(X|Z) \\
&\quad = h(XY) + \underline{h(Y|X)} + h(YZ) + \underline{h(Z|Y)} + h(XZ) + \underline{h(X|Z)} \\
&\quad \geq h(XY) + h(Y|XZ) + h(YZ) + h(Z|XY) + h(XZ) + h(X|YZ) \\
&\quad = 3h(XYZ) = 3 \log |Q|
\end{aligned}
$$

Discussion

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

  $Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z)$: $|Q| \leq ||\text{deg}_R(X|Y)||_2 \cdot ||\text{deg}_S(Z|Y)||_2$

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

- 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

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

Construct the following linear program:

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

The Linear Program

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

Construct the following linear program:

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

$$\begin{aligned} h(\boldsymbol{X}_j) &\leq \log |R_j| \\ h(\boldsymbol{V}|\boldsymbol{U}) &\leq \log \max \deg(\boldsymbol{V}|\boldsymbol{U}) \\ h(\boldsymbol{U}) + p h(\boldsymbol{V}|\boldsymbol{U}) &\leq \log ||\deg(\boldsymbol{V}|\boldsymbol{U})||_p^p \end{aligned}$$

The Linear Program

$Q(\boldsymbol{X}) = \bigwedge_j R_j(\boldsymbol{X}_j)$, m atoms, n variables.
Construct the following linear program:

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

$$\begin{aligned} h(\boldsymbol{X}_j) &\leq \log|R_j| \\ h(\boldsymbol{V}|\boldsymbol{U}) &\leq \log \max \deg(\boldsymbol{V}|\boldsymbol{U}) \\ h(\boldsymbol{U}) + p h(\boldsymbol{V}|\boldsymbol{U}) &\leq \log ||\deg(\boldsymbol{V}|\boldsymbol{U})||_p^p \end{aligned}$$

- Add all Shannon inequalities as constraints:

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

The Linear Program

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

Construct the following linear program:

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

$$\begin{aligned} h(\boldsymbol{X}_j) &\leq \log |R_j| \\ h(\boldsymbol{V}|\boldsymbol{U}) &\leq \log \max \deg(\boldsymbol{V}|\boldsymbol{U}) \\ h(\boldsymbol{U}) + p h(\boldsymbol{V}|\boldsymbol{U}) &\leq \log ||\deg(\boldsymbol{V}|\boldsymbol{U})||_p^p \end{aligned}$$

- Add all Shannon inequalities as constraints:

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

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

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:

$$c_1 : \quad h(XY) \leq \log |R|$$
$$c_2 : \quad h(YZ) \leq \log |S|$$

Example

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

Maximize $h(XYZ)$, where:

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

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

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

Example

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

Maximize $h(XYZ)$, where:

$$\begin{aligned}
c_1 &: & h(XY) &\leq \log |R| \\
c_2 &: & h(YZ) &\leq \log |S| \\
c_3 &: & h(XZ) &\leq \log |T| \\
\sigma_1 &: & -h(XY) - h(YZ) & \\
& & +h(XYZ) + h(Y) &\leq 0
\end{aligned}$$

Example

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

Maximize $h(XYZ)$, where:

$$\begin{aligned}
c_1 :& \quad h(XY) \leq \log|R| \\
c_2 :& \quad h(YZ) \leq \log|S| \\
c_3 :& \quad h(XZ) \leq \log|T| \\
\sigma_1 :& \quad -h(XY) - h(YZ) \\
& \quad +h(XYZ) + h(Y) \leq 0 \\
\sigma_2 :& \quad -h(Y) - h(XZ) \\
& \quad +h(XYZ) \leq 0 \\
& \quad \ldots
\end{aligned}$$

Example

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

Dual:

Maximize $h(XYZ)$, where:

$$
\begin{aligned}
c_1 &: & h(XY) &\leq \log|R| \\
c_2 &: & h(YZ) &\leq \log|S| \\
c_3 &: & h(XZ) &\leq \log|T| \\
\sigma_1 &: & -h(XY) - h(YZ) & \\
& & +h(XYZ) + h(Y) &\leq 0 \\
\sigma_2 &: & -h(Y) - h(XZ) & \\
& & +h(XYZ) &\leq 0 \\
& & & \cdots \\
\sigma_{18} &: & \cdots &\leq 0
\end{aligned}
$$

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:

$$\begin{aligned}
c_1 &: & h(XY) &\leq \log |R| \\
c_2 &: & h(YZ) &\leq \log |S| \\
c_3 &: & h(XZ) &\leq \log |T| \\
\sigma_1 &: & -h(XY) - h(YZ) & \\
& & +h(XYZ) + h(Y) &\leq 0 \\
\sigma_2 &: & -h(Y) - h(XZ) & \\
& & +h(XYZ) &\leq 0 \\
& & & \cdots \\
\sigma_{18} &: & \cdots &\leq 0
\end{aligned}$$

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:

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

Dual:

Maximize $h(XYZ)$, where:

$$\begin{aligned}
c_1 &: & h(XY) &\leq \log |R| \\
c_2 &: & h(YZ) &\leq \log |S| \\
c_3 &: & h(XZ) &\leq \log |T| \\
\sigma_1 &: & -h(XY) - h(YZ) & \\
& & +h(XYZ) + h(Y) &\leq 0 \\
\sigma_2 &: & -h(Y) - h(XZ) & \\
& & +h(XYZ) &\leq 0 \\
& & & \cdots \\
\sigma_{18} &: & \cdots &\leq 0
\end{aligned}$$

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:

$$\begin{aligned}
h(XYZ): &\quad \sigma_1 + \sigma_2 + \cdots \geq 1\\
h(XY): &\quad c_1 - \sigma_1 + \cdots \geq 0\\
h(YZ): &\quad c_2 - \sigma_1 + \cdots \geq 0\\
h(XZ): &\quad c_3 - \sigma_2 + \cdots \geq 0
\end{aligned}$$

Dual:

Maximize $h(XYZ)$, where:

$$\begin{aligned}
c_1: &\quad h(XY) \leq \log|R|\\
c_2: &\quad h(YZ) \leq \log|S|\\
c_3: &\quad h(XZ) \leq \log|T|\\
\sigma_1: &\quad -h(XY) - h(YZ)\\
&\quad +h(XYZ) + h(Y) \leq 0\\
\sigma_2: &\quad -h(Y) - h(XZ)\\
&\quad +h(XYZ) \leq 0\\
&\quad \cdots\\
\sigma_{18}: &\quad \cdots \leq 0
\end{aligned}$$

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:

$$\begin{aligned}
h(XYZ): &\quad \sigma_1 + \sigma_2 + \cdots \geq 1\\
h(XY): &\quad c_1 - \sigma_1 + \cdots \geq 0\\
h(YZ): &\quad c_2 - \sigma_1 + \cdots \geq 0\\
h(XZ): &\quad c_3 - \sigma_2 + \cdots \geq 0\\
h(X): &\quad \cdots \geq 0\\
h(Y): &\quad \sigma_1 - \sigma_2 + \cdots \geq 0\\
h(Z): &\quad \cdots \geq 0
\end{aligned}$$

Dual:

Maximize $h(XYZ)$, where:

$$\begin{aligned}
c_1: &\quad h(XY) \leq \log|R|\\
c_2: &\quad h(YZ) \leq \log|S|\\
c_3: &\quad h(XZ) \leq \log|T|\\
\sigma_1: &\quad -h(XY) - h(YZ)\\
&\quad +h(XYZ) + h(Y) \leq 0\\
\sigma_2: &\quad -h(Y) - h(XZ)\\
&\quad +h(XYZ) \leq 0\\
&\quad \cdots\\
\sigma_{18}: &\quad \cdots \leq 0
\end{aligned}$$

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:

$$
\begin{aligned}
h(XYZ) &: & \sigma_1 + \sigma_2 + \cdots &\geq 1 \\
h(XY) &: & c_1 - \sigma_1 + \cdots &\geq 0 \\
h(YZ) &: & c_2 - \sigma_1 + \cdots &\geq 0 \\
h(XZ) &: & c_3 - \sigma_2 + \cdots &\geq 0 \\
h(X) &: & \cdots &\geq 0 \\
h(Y) &: & \sigma_1 - \sigma_2 + \cdots &\geq 0 \\
h(Z) &: & \cdots &\geq 0
\end{aligned}
$$

Dual:

Maximize $h(XYZ)$, where:

$$
\begin{aligned}
c_1 &: & h(XY) &\leq \log|R| \\
c_2 &: & h(YZ) &\leq \log|S| \\
c_3 &: & h(XZ) &\leq \log|T| \\
\sigma_1 &: & -h(XY) - h(YZ) & \\
& & +h(XYZ) + h(Y) &\leq 0 \\
\sigma_2 &: & -h(Y) - h(XZ) & \\
& & +h(XYZ) &\leq 0 \\
& & &\cdots \\
\sigma_{18} &: & \cdots &\leq 0
\end{aligned}
$$

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

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

$$h(XYZ)(\sigma_1 + \cdots) + h(XY)(c_1 - \sigma_1 + \cdots) + \cdots \geq h(XYZ)$$

Group by the coefficients $c_1, c_2, c_3, \sigma_1, \sigma_2, \ldots$

$$c_1 h(XY) + c_2 h(YZ) + c_3 h(XZ) + \sigma_1(\cdots) + \cdots \geq h(XYZ)$$

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:

$$c_1 h(XY) + c_2 h(YZ) + c_3 h(XZ) \geq h(XYZ)$$

Discussion

AGM bound:

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

Discussion

AGM bound:

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

General bound:

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

Discussion

AGM bound:

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

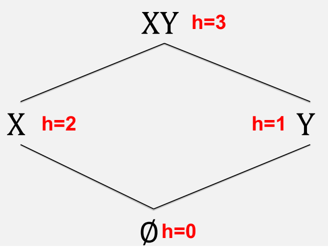
General bound:

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

Modular Functions

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

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 $h(\boldsymbol{V}|\boldsymbol{U})$) then the dual LP has an optimal solution $\boldsymbol{h}$ 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.