Cota AGM

Clase 25

IIC 3413

Prof. Cristian Riveros

Tamaño del output de una consulta

Considere la consulta conjuntiva:

$$Q \;:=\; R(x,y), S(y,z)$$

Suponga que $|R(D)| = |S(D)| = N$.

¿cuál es el tamaño de $|Q(D)|$ según N **en el peor caso**?

Considere ahora la consulta conjuntiva:

$$Q' \;:=\; R(x,y), S(y,z), T(z,x)$$

Suponga que $|R(D)| = |S(D)| = |T(D)| = N$.

¿y ahora? ¿cuál es el tamaño de $Q'(D)$?

Outline

Cubrimientos

Cota AGM

Peor caso

Outline

Cubrimientos

Cota AGM

Peor caso

En busqueda de una cota para el tamaño del output

Considere la consulta conjuntiva (sin proyección ni constantes, solo joins):

$$Q(y_1, \ldots, y_m) \;:=\; R_1(\bar{x}_1), \ldots, R_n(\bar{x}_n)$$

y D una base de datos tal que $N_i = |R_i(D)|$ para todo $i \leq n$.

Objetivo principal

Buscar una **cota** $C(N_1, \ldots, N_n)$ tal que
para toda base de datos D con $N_i = |R_i(D)|$, se cumple que:

$$|Q(D)| \;\leq\; C(N_1, \ldots, N_n)$$

¿cuál es una **buena cota** para $|Q(D)|$?

¿cuál es una buena cota para $|Q(D)|$?

- $Q := R_1(\bar{x}_1), \ldots, R_n(\bar{x}_n)$ con $y_1, \ldots, y_m$ todas las variables en Q.
- D una base de datos tal que $N_i = |R_i(D)|$ para todo $i \leq n$.

Algunos casos

1. Si hay un átomo $R_i(\bar{x}_i)$ tal que $\bar{x}_i = \{y_1, \ldots, y_m\}$, entonces:

$$|Q(D)| \leq N_i \qquad \text{(¿por qué?)}$$

2. Si hay átomos $R_{i_1}(\bar{x}_{i_1}), \ldots, R_{i_k}(\bar{x}_{i_k})$ tal que $\bar{x}_{i_1} \cup \ldots \cup \bar{x}_{i_k} = \{y_1, \ldots, y_m\}$ entonces:

$$|Q(D)| \leq \prod_{j=1}^{k} N_{i_j} \qquad \text{(¿por qué?)}$$

En otras palabras, buscamos **cubrir** las variables con algunos átomos.

Cubrimiento de un hipergrafo

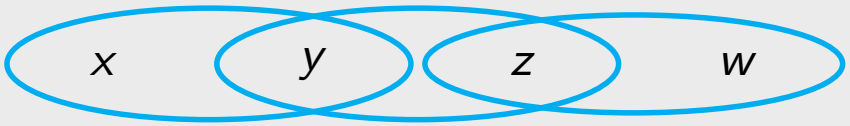
Sea $\mathcal{H} = (V, E)$ un hipergrafo
con V el conjunto de **vértices** y $E \subseteq 2^V$ el conjunto de **hiperaristas**.

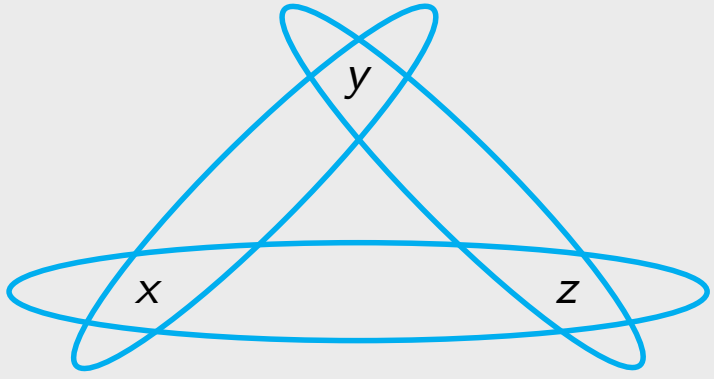
Definición

Decimos que $C \subseteq E$ es un **cubrimiento** de $\mathcal{H}$ ssi:

$$V = \bigcup_{h \in C} h$$

¿cuál es un cubrimiento para cada hipergrafo?





Cubrimiento de un hipergrafo

Sea $\mathcal{H} = (V, E)$ un hipergrafo
con V el conjunto de **vértices** y $E \subseteq 2^V$ el conjunto de **hiperaristas**.

Definición

Decimos que $C \subseteq E$ es un **cubrimiento** de $\mathcal{H}$ ssi:

$$V = \bigcup_{h \in C} h$$

Problema **clásico** en teoría de la computación y optimización:

`PROBLEMA:`	Cubrimiento de un hipergrafo
`INPUT:`	Un hipergrafo $\mathcal{H} = (V, E)$
`OUTPUT:`	Un cubrimiento C de $\mathcal{H}$ de tamaño mínimo en términos de $\lvert C \rvert$.

¿cómo nos sirve este problema **para encontrar una cota** para $\lvert Q(D) \rvert$?

Primera cota para $|Q(D)|$

- $Q := R_1(\bar{x}_1), \ldots, R_n(\bar{x}_n)$ con $y_1, \ldots, y_m$ todas las variables en Q.
- D una base de datos tal que $N_i = |R_i(D)|$ para todo $i \leq n$.

Definición (recordatorio)

Definimos el **hipergrafo** de Q como $\mathcal{H}_Q = (V, E)$ tal que:

- $V = \{y_1, \ldots, y_m\}$
- $E = \{\{x_1, \ldots, x_k\} \subseteq 2^V \mid \exists i.\ x_1, \ldots, x_k \text{ son todas las variables de } R_i\}$.

Desde ahora usaremos $R_i = \{x_1, \ldots, x_k\}$, $E = \{R_1, \ldots, R_n\}$ y $N_{R_i} = N_i$.

Cota de cubrimiento

$$|Q(D)| \leq \min_{C \text{ cubrimiento de } \mathcal{H}_Q} \left\{ \prod_{R \in C} N_R \right\}$$

¿es una "buena cota" 'para $|Q(D)|$?

Outline

Cubrimientos

Cota AGM

Peor caso

Programa entero para cota de cubrimiento

- $Q := R_1(\bar{x}_1), \ldots, R_n(\bar{x}_n)$ con $y_1, \ldots, y_m$ todas las variables en Q.
- $\mathcal{H}_Q = (V, E)$ es el hipergrafo con $V = \{y_1, \ldots, y_m\}$ y $E = \{R_1, \ldots, R_n\}$.
- D una base de datos tal que $N_R = |R(D)|$ para todo $R \in E$.

$$|Q(D)| \leq \min_{C \text{ cubrimiento de } \mathcal{H}_Q} \left\{ \prod_{R \in C} N_R \right\}$$

Cota de cubrimiento versión en programación entera

$$\mathcal{P}_{Q,D}: \quad \min \quad \prod_{R \in E} (N_R)^{c_R}$$

$$\text{tal que:} \quad \sum_{R: y \in R} c_R \geq 1 \quad \text{para cada variable } y \in V$$

$$c_R \in \{0, 1\} \quad \text{para cada relación } R \in E$$

Programa entero para cota de cubrimiento

Cota de cubrimiento versión en programación entera

$$\mathcal{P}_{Q,D}: \quad \min \quad \prod_{R \in E} (N_R)^{c_R}$$

$$\text{tal que:} \quad \sum_{R: y \in R} c_R \geq 1 \quad \text{para cada variable } y \in V$$

$$c_R \in \{0,1\} \quad \text{para cada relación } R \in E$$

Ejemplo $Q := R(x,y), S(y,z), T(z,x)$

$$\min \quad (N_R)^{c_R} \cdot (N_S)^{c_S} \cdot (N_T)^{c_T}$$

$$\begin{aligned}
\text{tal que:} \quad & c_R + c_T \geq 1 && \text{(restricción de } x\text{)} \\
& c_R + c_S \geq 1 && \text{(restricción de } y\text{)} \\
& c_S + c_T \geq 1 && \text{(restricción de } z\text{)} \\
& c_R, c_S, c_T \in \{0,1\} && \text{para cada } i \leq n
\end{aligned}$$

Programa entero para cota de cubrimiento

Cota de cubrimiento versión en programación entera

$$\mathcal{P}_{Q,D}: \quad \min \quad \prod_{R \in E} (N_R)^{c_R}$$

$$\text{tal que:} \quad \sum_{R: y \in R} c_R \geq 1 \quad \text{para cada variable } y \in V$$

$$c_R \in \{0,1\} \quad \text{para cada relación } R \in E$$

El programa anterior es **equivalente** a minimizar:

$$\mathcal{P}^*_{Q,D}: \quad \min \quad \sum_{R \in E} \log_2(N_R) \cdot c_R$$

$$\text{tal que:} \quad \sum_{R: y \in R} c_R \geq 1 \quad \text{para cada variable } y \in V$$

$$c_R \in \{0,1\} \quad \text{para cada relación } R \in E$$

tal que si O^* es el valor óptimo para $\mathcal{P}_{Q,D}$, entonces $|Q(D)| \leq 2^{O^*}$.

¿es posible mejorar la cota encontrada por el **programa entero**?

Relajación de programa entero

Podemos relajar el programa anterior desde los enteros a **los racionales**:

$$\mathcal{P}^*_{Q,D}: \quad \min \quad \sum_{R \in E} \log_2(N_R) \cdot c_R$$

$$\text{tal que:} \quad \sum_{R: y \in R} c_R \geq 1 \qquad \text{para cada variable } y \in V$$

$$0 \leq c_R \leq 1 \qquad \text{para cada relación } R \in E$$

¿cuál es la **interpretación** del programa lineal $\mathcal{P}^*_{Q,D}$?

Cubrimiento fraccionario de un hipergrafo

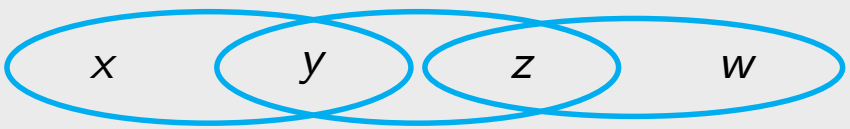
Sea $\mathcal{H} = (V, E)$ un hipergrafo
con V el conjunto de **vértices** y $E \subseteq 2^V$ el conjunto de **hiperaristas**.

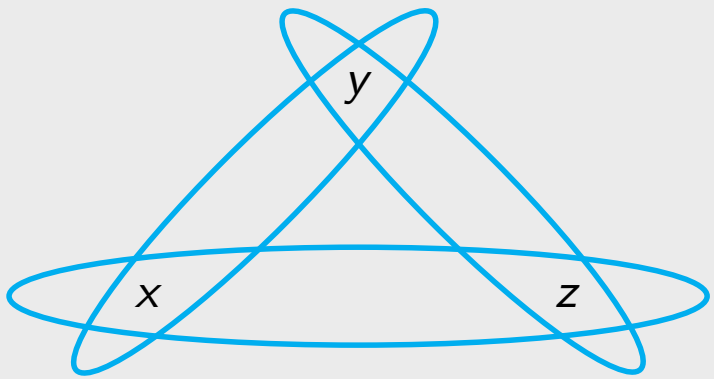
Definición

Decimos que $C : E \to [0, 1]$ es un **cubrimiento fraccionario** de $\mathcal{H}$ ssi :

$$\text{para todo } y \in V, \quad \sum_{h : y \in h} C(h) \geq 1$$

¿cuál es un cubrimiento fraccionario para cada hipergrafo?





Cubrimiento fraccionario de un hipergrafo

Sea $\mathcal{H} = (V, E)$ un hipergrafo
con V el conjunto de **vértices** y $E \subseteq 2^V$ el conjunto de **hiperaristas**.

Definición

Decimos que $C : E \to [0, 1]$ es un **cubrimiento fraccionario** de $\mathcal{H}$ ssi :

$$\text{para todo } y \in V, \quad \sum_{h : y \in h} C(h) \geq 1$$

`PROBLEMA:`	Cubrimiento fraccionario de un hipergrafo
`INPUT:`	Un hipergrafo $\mathcal{H} = (V, E)$
`OUTPUT:`	Un cubrimiento fraccionario $C : E \to [0, 1]$ de $\mathcal{H}$ de "tamaño mínimo" en términos de C.

¿cuál es la función de **tamaño** que estamos minimizando para $|Q(D)|$?

Relajación de programa entero

Podemos relajar el programa anterior desde los enteros a **los racionales**:

$$\mathcal{P}^*_{Q,D}: \quad \min \quad \sum_{R \in E} \log_2(N_R) \cdot c_R$$

$$\text{tal que:} \quad \sum_{R: y \in R} c_R \geq 1 \qquad \text{para cada variable } y \in V$$

$$0 \leq c_R \leq 1 \qquad \text{para cada relación } R \in E$$

¿cuál es la **interpretación** del programa lineal $\mathcal{P}^*_{Q,D}$?

- Estamos buscando un cubrimiento fraccionario de $\mathcal{H}_Q$.
- Si $c^*_{R_1}, \ldots, c^*_{R_n}$ es una solución para el **programa entero** $\mathcal{P}_{Q,D}$, entonces $c^*_{R_1}, \ldots, c^*_{R_n}$ será una solución para el **programa lineal** $\mathcal{P}^*_{Q,D}$.

(pero no necesariamente al revez)

Cota AGM (Atserias-Grohe-Marx)

Podemos relajar el programa anterior desde los enteros a **los racionales**:

$$\mathcal{P}^*_{Q,D}: \quad \min \quad \sum_{R \in E} \log_2(N_R) \cdot c_R$$

$$\text{tal que:} \quad \sum_{R: y \in R} c_R \geq 1 \qquad \text{para cada variable } y \in V$$

$$0 \leq c_R \leq 1 \qquad \text{para cada relación } R \in E$$

Teorem (Cota AGM)

Para toda consulta conjuntiva Q y base de datos D, si $O^*_{Q,D}$ es **el valor óptimo para el programa lineal** $\mathcal{P}^*_{Q,D}$, entonces:

$$|Q(D)| \leq 2^{O^*_{Q,D}}$$

y existen BD D **arbitrariamente grandes** tal que $|Q(D)| = 2^{O^*_{Q,D}}$.

La cota $2^{O^*_{Q,D}}$ es **óptima***

Cota AGM (Atserias-Grohe-Marx)

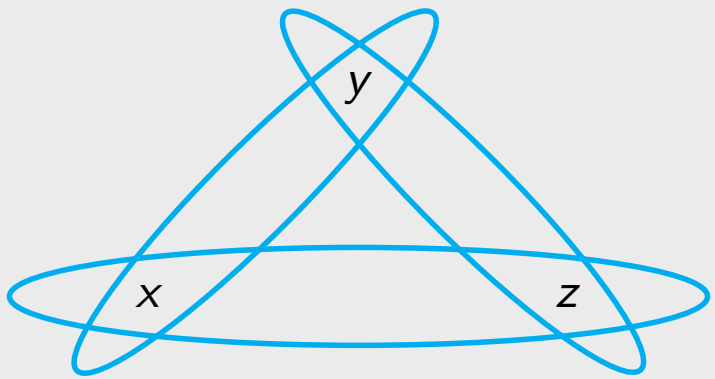
Teorem (Cota AGM)

Para toda consulta conjuntiva Q y base de datos D, si $O^*_{Q,D}$ es **el valor óptimo para el programa lineal** $\mathcal{P}^*_{Q,D}$, entonces:

$$|Q(D)| \leq 2^{O^*_{Q,D}}$$

y existen BD D **arbitrariamente grandes** tal que $|Q(D)| = 2^{O^*_{Q,D}}$.

Ejemplo: $R(x,y), S(y,z), T(z,x)$ con $N_R = N_S = N_T = N$



$$\begin{aligned} \min \quad & N^{c_R} \cdot N^{c_S} \cdot N^{c_T} \\ \text{tal que:} \quad & c_R + c_T \geq 1 \\ & c_R + c_S \geq 1 \\ & c_S + c_T \geq 1 \\ & c_R, c_S, c_T \in [0,1] \end{aligned}$$

La cota para la consulta de triangulo es $N^{\frac{3}{2}}$!

Outline

Cubrimientos

Cota AGM

Peor caso

¿cómo encontramos el peor caso de la cota AGM?

Programa lineal:

$$\begin{aligned} \min \quad & c^t \cdot \bar{x} \\ \text{tal que:} \quad & A \cdot \bar{x} \geq b \\ & \bar{x} \geq 0 \end{aligned}$$

Programa dual:

$$\begin{aligned} \max \quad & b^t \cdot \bar{y} \\ \text{tal que:} \quad & A^t \cdot \bar{y} \leq c \\ & \bar{y} \geq 0 \end{aligned}$$

Del programa lineal al programa dual

$\mathcal{P}^*_{Q,D}$:

$$\begin{aligned} \min \quad & \sum_{R \in E} \log_2(N_R) \cdot c_R \\ \text{tq:} \quad & \sum_{R: y \in R} c_R \geq 1 \quad \forall y \in V \\ & 0 \leq c_R \leq 1 \quad \forall R \in E \end{aligned}$$

¿cómo encontramos el peor caso de la cota AGM?

Programa lineal:

$$\begin{aligned} \min \quad & c^t \cdot \bar{x} \\ \text{tal que:} \quad & A \cdot \bar{x} \geq b \\ & \bar{x} \geq 0 \end{aligned}$$

Programa dual:

$$\begin{aligned} \max \quad & b^t \cdot \bar{y} \\ \text{tal que:} \quad & A^t \cdot \bar{y} \leq c \\ & \bar{y} \geq 0 \end{aligned}$$

Del programa lineal al programa dual

$\mathcal{P}^*_{Q,D}$:

$$\begin{aligned} \min \quad & \sum_{R \in E} \log_2(N_R) \cdot c_R \\ \text{tq:} \quad & \sum_{R: y \in R} c_R \geq 1 \quad \forall y \in V \\ & 0 \leq c_R \quad \forall R \in E \end{aligned}$$

$\mathcal{D}^*_{Q,D}$:

$$\begin{aligned} \max \quad & \sum_{y \in V} d_y \\ \text{tq:} \quad & \sum_{y: y \in R} d_y \leq \log_2(N_R) \quad \forall R \in E \\ & 0 \leq d_y \quad \forall y \in V \end{aligned}$$

tal que el valor óptimo de $\mathcal{P}^*_{Q,D}$ es igual al valor óptimo de $\mathcal{D}^*_{Q,D}$.

¿cómo encontramos el peor caso de la cota AGM?

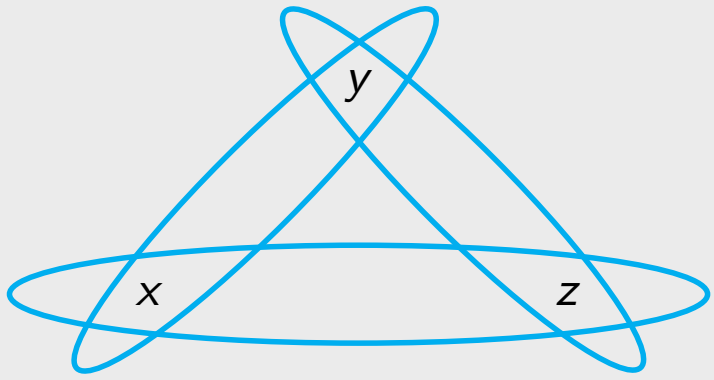
$$\begin{aligned} \min \quad & \sum_{R \in E} \log_2(N_R) \cdot c_R \\ \text{tq:} \quad & \sum_{R: y \in R} c_R \geq 1 \quad \forall y \in V \\ & 0 \leq c_R \quad \forall R \in E \end{aligned} \quad \rightarrow \quad \begin{aligned} \max \quad & \sum_{y \in V} d_y \\ \text{tq:} \quad & \sum_{y: y \in R} d_y \leq \log_2(N_R) \quad \forall R \in E \\ & 0 \leq d_y \quad \forall y \in V \end{aligned}$$

Ejemplo: $R(x,y), S(y,z), T(z,x)$ con $N_R = N_S = N_T = N$

$$\begin{aligned} \min \quad & (c_R + c_S + c_T) \cdot \log_2(N) \\ \text{tq:} \quad & c_R + c_T \geq 1 \\ & c_R + c_S \geq 1 \\ & c_S + c_T \geq 1 \\ & c_R, c_S, c_T \geq 0 \end{aligned} \quad \rightarrow \quad \begin{aligned} \max \quad & d_x + d_y + d_z \\ \text{tq:} \quad & d_x + d_y \leq \log_2(N) \\ & d_y + d_z \leq \log_2(N) \\ & d_z + d_x \leq \log_2(N) \\ & d_x, d_y, d_z \geq 0 \end{aligned}$$

¿qué representan las **nuevas variables** d_y para cada variable y?

Conclusión sobre la cota AGM



$$\begin{aligned} \min \quad & N^{c_R} \cdot N^{c_S} \cdot N^{c_T} \\ \text{tal que:} \quad & c_R + c_T \geq 1 \\ & c_R + c_S \geq 1 \\ & c_S + c_T \geq 1 \\ & c_R, c_S, c_T \in [0,1] \end{aligned}$$

Conclusión sobre $Q_\Delta := R(x,y), S(y,z), T(z,x)$

- El tamaño de $|Q_\Delta(D)|$ es a lo más $N^{\frac{3}{2}}$.
- El tamaño de $|R \bowtie S|$, $|R \bowtie T|$, o $|T \bowtie S|$ puede ser N^2.

¿es posible calcular $Q_\Delta(D)$ en tiempo $\mathcal{O}(N^{\frac{3}{2}})$?
Para toda Q y D, ¿es posible calcular $Q(D)$ en tiempo a lo más $\mathcal{O}(2^{O^*_{Q,D}})$?