Efficient Algorithms for Frequently Asked Questions

7. Worst-Case Optimal Size Bounds for Joins

Prof. Dan Olteanu

DaST
Data • (Systems+Theory)

University of Zurich UZH

May 2, 2022

https://lms.uzh.ch/url/RepositoryEntry/17185308706

Agenda for This Lecture

Worst-case optimal size bounds for joins

- Key parameter: The fractional edge cover number ρ^*
- Mentioned it several times in the previous lectures

Upper bound via an information-theoretic argument

- Warm-up: Triangle join
- General Case using Shearer's Lemma

Lower bound

- Warm-up: Triangle join
- General case via dual linear program for fractional edge cover number

The effect of the size of input factors: Same size vs different sizes

The Upper Bound Argument

Upper Bound on Join Output Size

Consider the join (all variables free, no marginalisation)

$$\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and input factor sizes $|\psi_S| = N_S$ for $S \in \mathcal{E}$

Upper Bound on Join Output Size

Consider the join (all variables free, no marginalisation)

$$\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and input factor sizes $|\psi_S| = N_S$ for $S \in \mathcal{E}$

- Let $(w_S)_{S \in \mathcal{E}}$ be any feasible solution to the linear program computing $\rho^*(\mathcal{H})$ with minimisation objective $\prod_{S \in \mathcal{E}} N_S^{w_S}$
- We will show that the output size $|\Phi|$ is upper-bounded by $\prod_{S \in \mathcal{E}} N_S^{w_S}$

Upper Bound on Join Output Size

Consider the join (all variables free, no marginalisation)

$$\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and input factor sizes $|\psi_S| = N_S$ for $S \in \mathcal{E}$

- Let $(w_S)_{S \in \mathcal{E}}$ be any feasible solution to the linear program computing $\rho^*(\mathcal{H})$ with minimisation objective $\prod_{S \in \mathcal{E}} N_S^{w_S}$
- We will show that the output size $|\Phi|$ is upper-bounded by $\prod_{S \in \mathcal{E}} N_S^{w_S}$
- By choosing $N = \max_{S \in \mathcal{E}} N_S$, this implies

$$|\Phi| \leq \prod_{S \in \mathcal{E}} N_S^{w_S} \leq \prod_{S \in \mathcal{E}} N^{w_S} = N^{\sum_{S \in \mathcal{E}} w_S} = N^{\rho^*(\mathcal{H})}$$

Upper Bound on Join Output Size

Consider the join (all variables free, no marginalisation)

$$\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and input factor sizes $|\psi_S| = N_S$ for $S \in \mathcal{E}$

- Let $(w_S)_{S \in \mathcal{E}}$ be any feasible solution to the linear program computing $\rho^*(\mathcal{H})$ with minimisation objective $\prod_{S \in \mathcal{E}} N_S^{w_S}$
- We will show that the output size $|\Phi|$ is upper-bounded by $\prod_{S \in \mathcal{E}} N_S^{w_S}$
- By choosing $N = \max_{S \in \mathcal{E}} N_S$, this implies

$$|\Phi| \leq \prod_{S \in \mathcal{E}} N_S^{w_S} \leq \prod_{S \in \mathcal{E}} N^{w_S} = N^{\sum_{S \in \mathcal{E}} w_S} = N^{\rho^*(\mathcal{H})}$$

- We will sketch a proof based on information theory
- Warm-up first: Triangle join with input factor sizes N

Warm-Up: Size Bound for Triangle Join

Upper Bound on Triangle Join Output Size

$$\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$$

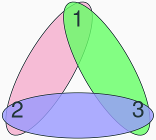
with input factor sizes $|\psi_{12}| = |\psi_{23}| = |\psi_{13}| = N$

Upper Bound on Triangle Join Output Size

$$\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$$

with input factor sizes $|\psi_{12}| = |\psi_{23}| = |\psi_{13}| = N$

Hypergraph $\mathcal{H}$



Linear program computing $\rho^*(\mathcal{H})$

minimise $w_{12} + w_{23} + w_{13}$
subject to

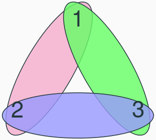
$$\begin{array}{lccccccl}
1: & w_{12} & + & w_{23} & & & \geq & 1 \\
2: & w_{12} & & & + & w_{13} & \geq & 1 \\
3: & & & w_{23} & + & w_{13} & \geq & 1 \\
 & w_{12} \geq 0 & & w_{23} \geq 0 & & w_{13} \geq 0 & &
\end{array}$$

Upper Bound on Triangle Join Output Size

$$\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$$

with input factor sizes $|\psi_{12}| = |\psi_{23}| = |\psi_{13}| = N$

Hypergraph $\mathcal{H}$



Linear program computing $\rho^*(\mathcal{H})$

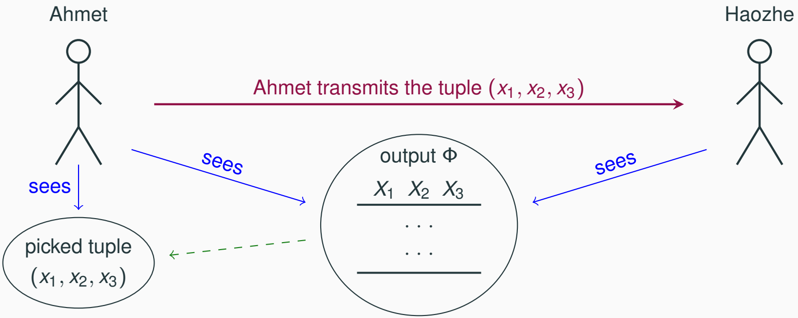
$$\begin{aligned}
&\textit{minimise } w_{12} + w_{23} + w_{13} \\
&\textit{subject to} \\
&1: \quad w_{12} + w_{23} \geq 1 \\
&2: \quad w_{12} + w_{13} \geq 1 \\
&3: \quad w_{23} + w_{13} \geq 1 \\
&\quad w_{12} \geq 0 \quad w_{23} \geq 0 \quad w_{13} \geq 0
\end{aligned}$$

- The optimal solution to the above program is $w_{12} = w_{23} = w_{13} = \frac{1}{2}$
- We will show that $|\Phi| \leq N^{\frac{3}{2}}$

A Two-Player Game

Consider a two-player game between Ahmet and Haozhe

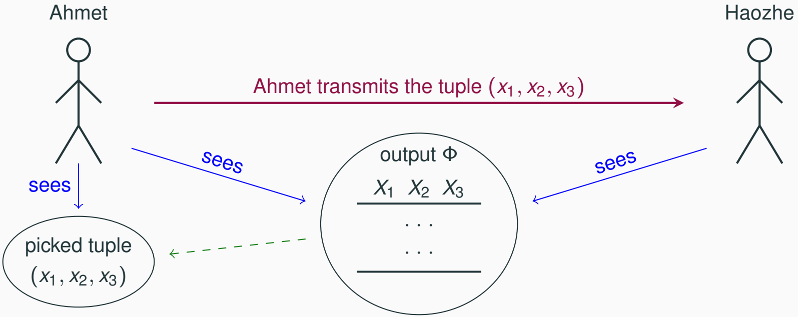
- Both players know the output of the triangle query
- Ahmet picks an arbitrary tuple from the output and transmits it to Haozhe



A Two-Player Game

Consider a two-player game between Ahmet and Haozhe

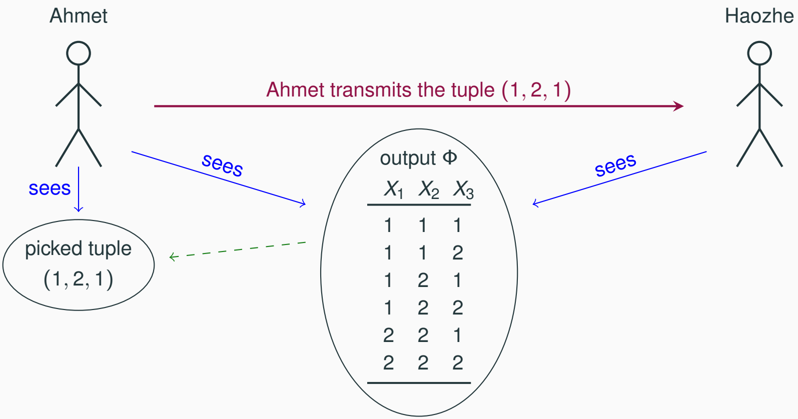
- Both players know the output of the triangle query
- Ahmet picks an arbitrary tuple from the output and transmits it to Haozhe



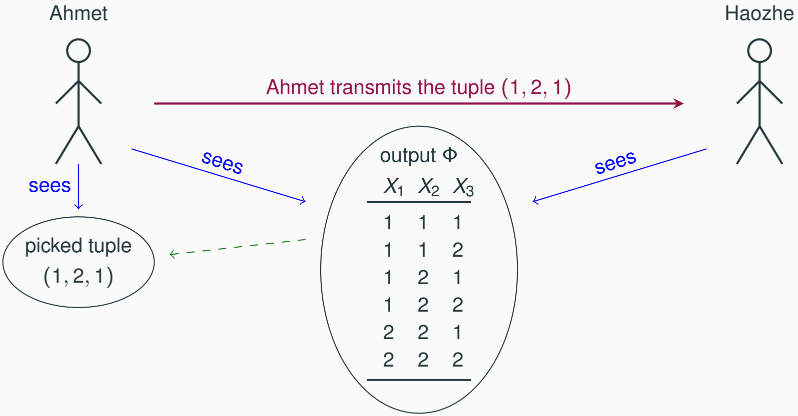
- Assume that the players have agreed on a binary coding system

How many bits does Ahmet need on avg to inform Haozhe which tuple he picked?

Two-Player Game Example



Two-Player Game Example



The best Ahmet and Haozhe can do is:

- Assign to each of the N tuples an index from 0 to $N - 1$
- Ahmet transmits to Haozhe the index of the picked tuple in binary

In the above example: $\log|\Phi| = \log 6$ bits are needed

Information Theoretic Perspective

- Ahmet picking an arbitrary tuple can be considered an experiment with random variable O
- The values of O are the output tuples in Φ
- The avg number of bits needed to transmit tuples depends on the uncertainty about O

Information Theoretic Perspective

- Ahmet picking an arbitrary tuple can be considered an experiment with random variable O
- The values of O are the output tuples in Φ
- The avg number of bits needed to transmit tuples depends on the uncertainty about O

Special cases:

- If O takes on a tuple with probability 1 (there is only one tuple), then there is no uncertainty and the avg number of needed bits is 0
- If the tuples are uniformly distributed, then the uncertainty is maximal and the avg number of needed bits is $\log |\Phi|$

Information Theoretic Perspective

- Ahmet picking an arbitrary tuple can be considered an experiment with random variable O
- The values of O are the output tuples in Φ
- The avg number of bits needed to transmit tuples depends on the uncertainty about O

Special cases:

- If O takes on a tuple with probability 1 (there is only one tuple), then there is no uncertainty and the avg number of needed bits is 0
- If the tuples are uniformly distributed, then the uncertainty is maximal and the avg number of needed bits is $\log |\Phi|$

The avg number of needed bits is the entropy $H(O)$ of O

Quick Recap: Entropy

The entropy of a random variable O with n possible outcomes $v_1, \ldots, v_n$:

$$H(O) = -\sum_{i \in [n]} \mathrm{P}(v_i) \cdot \log \mathrm{P}(v_i)$$

Quick Recap: Entropy

The entropy of a random variable O with n possible outcomes $v_1, \ldots, v_n$:

$$H(O) = -\sum_{i \in [n]} \mathrm{P}(v_i) \cdot \log \mathrm{P}(v_i)$$

- Special case 1: If O takes on a tuple with probability 1 (there is only one tuple), then there is no uncertainty and the avg number of needed bits is 0

  Only one outcome means $n = 1$. Then,

$$H(O) = -\mathrm{P}(v_1) \cdot \log \mathrm{P}(v_1) = -1 \cdot \log 1 = 0$$

Quick Recap: Entropy

The entropy of a random variable O with n possible outcomes $v_1, \ldots, v_n$:

$$H(O) = -\sum_{i \in [n]} \mathrm{P}(v_i) \cdot \log \mathrm{P}(v_i)$$

- Special case 1: If O takes on a tuple with probability 1 (there is only one tuple), then there is no uncertainty and the avg number of needed bits is 0

  Only one outcome means $n = 1$. Then,

  $$H(O) = -\mathrm{P}(v_1) \cdot \log \mathrm{P}(v_1) = -1 \cdot \log 1 = 0$$

- Special case 2: If the tuples are uniformly distributed, then the uncertainty is maximal and the avg number of needed bits is $\log |\Phi|$

  Uniform distribution means $P(v_i) = \frac{1}{n}, \forall i \in [n]$. Then,

  $$H(O) = -\sum_{i \in [n]} \mathrm{P}(v_i) \cdot \log \mathrm{P}(v_i) = -n \cdot (\frac{1}{n} \cdot \log \frac{1}{n}) = -\log \frac{1}{n} = -(\log 1 - \log n) = \log n$$

Our Goal

- We assume that Ahmet picks a tuple from the output uniformly at random

  $\Longrightarrow H(O) = \log |\Phi|$

Our Goal

- We assume that Ahmet picks a tuple from the output uniformly at random
  $\implies H(O) = \log|\Phi|$
- Assume that I_{12}, I_{23}, and I_{13} are random variables where each I_{ij} takes on a tuple from ψ_{ij} uniformly at random
  $\implies H(I_{ij}) = \log|\psi_{ij}| = \log N$

Our Goal

- We assume that Ahmet picks a tuple from the output uniformly at random $\implies H(O) = \log|\Phi|$
- Assume that I_{12}, I_{23}, and I_{13} are random variables where each I_{ij} takes on a tuple from ψ_{ij} uniformly at random $\implies H(I_{ij}) = \log|\psi_{ij}| = \log N$

Our goal is to show: $2H(O) \leq H(I_{12}) + H(I_{23}) + H(I_{13})$

Our Goal

- We assume that Ahmet picks a tuple from the output uniformly at random $\implies H(O) = \log|\Phi|$
- Assume that I_{12}, I_{23}, and I_{13} are random variables where each I_{ij} takes on a tuple from ψ_{ij} uniformly at random $\implies H(I_{ij}) = \log|\psi_{ij}| = \log N$

Our goal is to show: $2H(O) \leq H(I_{12}) + H(I_{23}) + H(I_{13})$

This implies:

$$2\log|\Phi| \leq \log N + \log N + \log N$$
$$\implies 2\log|\Phi| \leq 3\log N$$

Our Goal

- We assume that Ahmet picks a tuple from the output uniformly at random
  $\implies H(O) = \log|\Phi|$
- Assume that I_{12}, I_{23}, and I_{13} are random variables where each I_{ij} takes on a tuple from ψ_{ij} uniformly at random
  $\implies H(I_{ij}) = \log|\psi_{ij}| = \log N$

Our goal is to show: $2H(O) \leq H(I_{12}) + H(I_{23}) + H(I_{13})$

This implies:

$$\begin{aligned} 2\log|\Phi| &\leq \log N + \log N + \log N \\ \implies 2\log|\Phi| &\leq 3\log N \\ \implies \log|\Phi| &\leq \frac{3}{2}\log N \end{aligned}$$

Our Goal

- We assume that Ahmet picks a tuple from the output uniformly at random $\implies H(O) = \log|\Phi|$
- Assume that I_{12}, I_{23}, and I_{13} are random variables where each I_{ij} takes on a tuple from ψ_{ij} uniformly at random $\implies H(I_{ij}) = \log|\psi_{ij}| = \log N$

Our goal is to show: $2H(O) \leq H(I_{12}) + H(I_{23}) + H(I_{13})$

This implies:

$$
\begin{aligned}
& 2\log|\Phi| \leq \log N + \log N + \log N \\
\implies & 2\log|\Phi| \leq 3\log N \\
\implies & \log|\Phi| \leq \frac{3}{2}\log N \\
\implies & \log|\Phi| \leq \log N^{\frac{3}{2}}
\end{aligned}
$$

Our Goal

- We assume that Ahmet picks a tuple from the output uniformly at random
  $\implies H(O) = \log|\Phi|$
- Assume that I_{12}, I_{23}, and I_{13} are random variables where each I_{ij} takes on a tuple from ψ_{ij} uniformly at random
  $\implies H(I_{ij}) = \log|\psi_{ij}| = \log N$

Our goal is to show: $2H(O) \leq H(I_{12}) + H(I_{23}) + H(I_{13})$

This implies:

$$
\begin{aligned}
&2\log|\Phi| \leq \log N + \log N + \log N \\
\implies &2\log|\Phi| \leq 3\log N \\
\implies &\log|\Phi| \leq \frac{3}{2}\log N \\
\implies &\log|\Phi| \leq \log N^{\frac{3}{2}} \\
\implies &|\Phi| \leq N^{\frac{3}{2}}
\end{aligned}
$$

Our Goal

- We assume that Ahmet picks a tuple from the output uniformly at random
  $\implies H(O) = \log|\Phi|$
- Assume that I_{12}, I_{23}, and I_{13} are random variables where each I_{ij} takes on a tuple from ψ_{ij} uniformly at random
  $\implies H(I_{ij}) = \log|\psi_{ij}| = \log N$

Our goal is to show: $2H(O) \leq H(I_{12}) + H(I_{23}) + H(I_{13})$

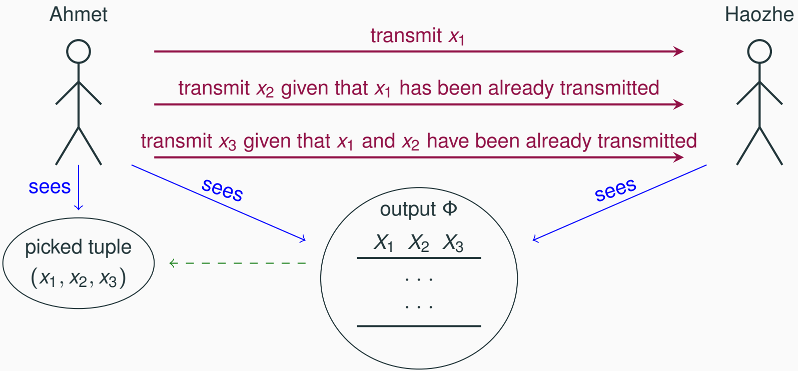
This implies:

$$
\begin{aligned}
& 2\log|\Phi| \leq \log N + \log N + \log N \\
\implies & 2\log|\Phi| \leq 3\log N \\
\implies & \log|\Phi| \leq \frac{3}{2}\log N \\
\implies & \log|\Phi| \leq \log N^{\frac{3}{2}} \\
\implies & |\Phi| \leq N^{\frac{3}{2}}
\end{aligned}
$$

Next: a strategy for Ahmet that helps to express $H(O)$ in terms of $H(I_{12})$, $H(I_{23})$, and $H(I_{13})$

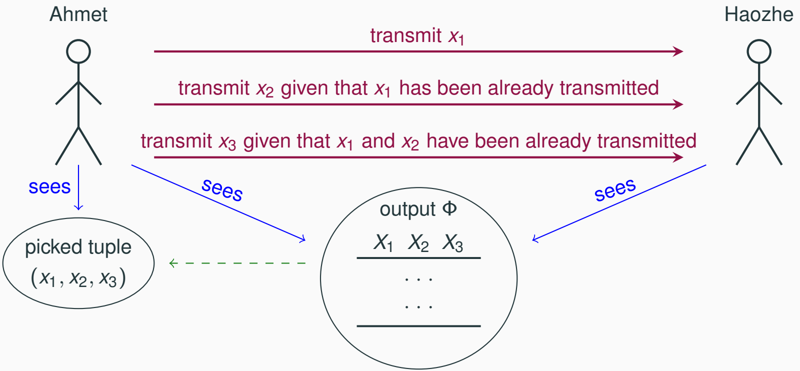
Alternative Strategy

Ahmet transmits the picked tuple in three steps



Alternative Strategy

Ahmet transmits the picked tuple in three steps



- In each step, Ahmet uses an optimal encoding given that Haozhe knows the values transmitted before

How many bits does Ahmet need on avg at each step?

Information Theoretic Perspective

We write O as a triple $O = (O_1, O_2, O_3)$ where each O_i is a random variable that takes on an X_i value

Information Theoretic Perspective

We write O as a triple $O = (O_1, O_2, O_3)$ where each O_i is a random variable that takes on an X_i value

- O_1, O_2, and O_3 are not uniformly distributed and are not independent!

Information Theoretic Perspective

We write O as a triple $O = (O_1, O_2, O_3)$ where each O_i is a random variable that takes on an X_i value

- O_1, O_2, and O_3 are not uniformly distributed and are not independent!

transmitting x_1

$H(O_1)$

Information Theoretic Perspective

We write O as a triple $O = (O_1, O_2, O_3)$ where each O_i is a random variable that takes on an X_i value

- O_1, O_2, and O_3 are not uniformly distributed and are not independent!

transmitting x_1	transmitting x_2 given x_1
$H(O_1)$	$H(O_2 \mid O_1)$

Information Theoretic Perspective

We write O as a triple $O = (O_1, O_2, O_3)$ where each O_i is a random variable that takes on an X_i value

- O_1, O_2, and O_3 are not uniformly distributed and are not independent!

transmitting x_1	transmitting x_2 given x_1	transmitting x_3 given x_1 and x_2
$H(O_1)$	$H(O_2 \mid O_1)$	$H(O_3 \mid O_1, O_2)$

Information Theoretic Perspective

We write O as a triple $O = (O_1, O_2, O_3)$ where each O_i is a random variable that takes on an X_i value

- O_1, O_2, and O_3 are not uniformly distributed and are not independent!

transmitting x_1	transmitting x_2 given x_1	transmitting x_3 given x_1 and x_2
$H(O_1)$	$H(O_2 \mid O_1)$	$H(O_3 \mid O_1, O_2)$

$$H(O) = H(O_1, O_2, O_3) = H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)$$

Information Theoretic Perspective

We write O as a triple $O = (O_1, O_2, O_3)$ where each O_i is a random variable that takes on an X_i value

- O_1, O_2, and O_3 are not uniformly distributed and are not independent!

transmitting x_1	transmitting x_2 given x_1	transmitting x_3 given x_1 and x_2
$H(O_1)$	$H(O_2 \mid O_1)$	$H(O_3 \mid O_1, O_2)$

$$H(O) = H(O_1, O_2, O_3) = H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)$$

- Conditional entropy $H(O_2 \mid O_1)$ gives the avg number of bits needed to transmit x_2 given that x_1 has been already transmitted
- Conditional entropy $H(O_3 \mid O_1, O_2)$ gives the avg number of bits needed to transmit x_3 given that x_1 and x_2 have been already transmitted

Information Theoretic Perspective

We write O as a triple $O = (O_1, O_2, O_3)$ where each O_i is a random variable that takes on an X_i value

- O_1, O_2, and O_3 are not uniformly distributed and are not independent!

transmitting x_1	transmitting x_2 given x_1	transmitting x_3 given x_1 and x_2
$H(O_1)$	$H(O_2 \mid O_1)$	$H(O_3 \mid O_1, O_2)$

$$H(O) = H(O_1, O_2, O_3) = H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)$$

- Conditional entropy $H(O_2 \mid O_1)$ gives the avg number of bits needed to transmit x_2 given that x_1 has been already transmitted
- Conditional entropy $H(O_3 \mid O_1, O_2)$ gives the avg number of bits needed to transmit x_3 given that x_1 and x_2 have been already transmitted
- We have $H(O_i, O_j) = H(O_i) + H(O_j \mid O_i)$

Information Theoretic Perspective

We write O as a triple $O = (O_1, O_2, O_3)$ where each O_i is a random variable that takes on an X_i value

- O_1, O_2, and O_3 are not uniformly distributed and are not independent!

transmitting x_1	transmitting x_2 given x_1	transmitting x_3 given x_1 and x_2
$H(O_1)$	$H(O_2 \mid O_1)$	$H(O_3 \mid O_1, O_2)$

$$H(O) = H(O_1, O_2, O_3) = H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)$$

- Conditional entropy $H(O_2 \mid O_1)$ gives the avg number of bits needed to transmit x_2 given that x_1 has been already transmitted
- Conditional entropy $H(O_3 \mid O_1, O_2)$ gives the avg number of bits needed to transmit x_3 given that x_1 and x_2 have been already transmitted
- We have $H(O_i, O_j) = H(O_i) + H(O_j \mid O_i)$

Next, we look closer at the relationship between $H(O_i, O_j)$ and $H(I_{ij})$

Observation 1

Transmitting (x_1, x_2) such that there is an x_3 with $(x_1, x_2, x_3) \in \Phi$ does not require more bits than transmitting $(x_1, x_2) \in \psi_{12}$ chosen uniformly at random

$$H(O_1) + H(O_2 \mid O_1) = H(O_1, O_2) \leq H(I_{12})$$

Observation 1

Transmitting (x_1, x_2) such that there is an x_3 with $(x_1, x_2, x_3) \in \Phi$ does not require more bits than transmitting $(x_1, x_2) \in \psi_{12}$ chosen uniformly at random

$$H(O_1) + H(O_2 \mid O_1) = H(O_1, O_2) \leq H(I_{12})$$

Example

input ψ_{12}

X_1	X_2
1	1
1	2
2	2
2	5
2	6

input ψ_{23}

X_2	X_3
1	1
1	2
2	1
2	2
3	1

input ψ_{13}

X_1	X_3
1	1
1	2
2	1
2	2
1	5

output Φ

X_1	X_2	X_3
1	1	1
1	1	2
1	2	1
1	2	2
2	2	1
2	2	2

Observation 1

Transmitting (x_1, x_2) such that there is an x_3 with $(x_1, x_2, x_3) \in \Phi$ does not require more bits than transmitting $(x_1, x_2) \in \psi_{12}$ chosen uniformly at random

$$H(O_1) + H(O_2 \mid O_1) = H(O_1, O_2) \leq H(I_{12})$$

Example

input ψ_{12}

X_1	X_2	
1	1	1/5
1	2	1/5
2	2	1/5
2	5	1/5
2	6	1/5

input ψ_{23}

X_2	X_3
1	1
1	2
2	1
2	2
3	1

input ψ_{13}

X_1	X_3
1	1
1	2
2	1
2	2
1	5

output Φ

X_1	X_2	X_3	
1	1	1	1/6
1	1	2	1/6
1	2	1	1/6
1	2	2	1/6
2	2	1	1/6
2	2	2	1/6

marginalised output $\bigoplus_{x_3} \Phi$

X_1	X_2	
1	1	1/3
1	2	1/3
2	2	1/3

$$H(O_1, O_2) = \log 3 \leq \log 5 = H(I_{12})$$

Observation 2

Transmitting (x_2, x_3) such that there is an x_1 with $(x_1, x_2, x_3) \in \Phi$ does not require more bits than transmitting $(x_2, x_3) \in \psi_{23}$ chosen uniformly at random

$$H(O_2) + H(O_3 \mid O_2) = H(O_2, O_3) \leq H(I_{23})$$

Observation 2

Transmitting (x_2, x_3) such that there is an x_1 with $(x_1, x_2, x_3) \in \Phi$ does not require more bits than transmitting $(x_2, x_3) \in \psi_{23}$ chosen uniformly at random

$$H(O_2) + H(O_3 \mid O_2) = H(O_2, O_3) \leq H(I_{23})$$

Example

input ψ_{12}

X_1	X_2
1	1
1	2
2	2
2	5
2	6

input ψ_{23}

X_2	X_3
1	1
1	2
2	1
2	2
3	1

input ψ_{13}

X_1	X_3
1	1
1	2
2	1
2	2
1	5

output Φ

X_1	X_2	X_3
1	1	1
1	1	2
1	2	1
1	2	2
2	2	1
2	2	2

Observation 2

Transmitting (x_2, x_3) such that there is an x_1 with $(x_1, x_2, x_3) \in \Phi$ does not require more bits than transmitting $(x_2, x_3) \in \psi_{23}$ chosen uniformly at random

$$H(O_2) + H(O_3 \mid O_2) = H(O_2, O_3) \leq H(I_{23})$$

Example

input ψ_{12}

X_1	X_2
1	1
1	2
2	2
2	5
2	6

input ψ_{23}

X_2	X_3	
1	1	1/5
1	2	1/5
2	1	1/5
2	2	1/5
3	1	1/5

input ψ_{13}

X_1	X_3
1	1
1	2
2	1
2	2
1	5

output Φ

X_1	X_2	X_3	
1	1	1	1/6
1	1	2	1/6
1	2	1	1/6
1	2	2	1/6
2	2	1	1/6
2	2	2	1/6

marginalised output $\bigoplus_{x_1} \Phi$

X_2	X_3	
1	1	1/6
1	2	1/6
2	1	1/3
2	2	1/3

$$H(O_2, O_3) = \frac{2}{6} \log 6 + \frac{2}{3} \log 3 \leq \log 5 = H(I_{23})$$

Observation 3

Similar to the other Observations

$$H(O_1) + H(O_3 \mid O_1) = H(O_1, O_3) \leq H(I_{13})$$

Observation 3

Similar to the other Observations

$$H(O_1) + H(O_3 \mid O_1) = H(O_1, O_3) \leq H(I_{13})$$

Example

input ψ_{12}

X_1	X_2
1	1
1	2
2	2
2	5
2	6

input ψ_{23}

X_2	X_3
1	1
1	2
2	1
2	2
3	1

input ψ_{13}

X_1	X_3
1	1
1	2
2	1
2	2
1	5

output Φ

X_1	X_2	X_3
1	1	1
1	1	2
1	2	1
1	2	2
2	2	1
2	2	2

Observation 3

Similar to the other Observations

$$H(O_1) + H(O_3 \mid O_1) = H(O_1, O_3) \leq H(I_{13})$$

Example

input ψ_{12}

X_1	X_2
1	1
1	2
2	2
2	5
2	6

input ψ_{23}

X_2	X_3
1	1
1	2
2	1
2	2
3	1

input ψ_{13}

X_1	X_3	
1	1	1/5
1	2	1/5
2	1	1/5
2	2	1/5
1	5	1/5

output Φ

X_1	X_2	X_3	
1	1	1	1/6
1	1	2	1/6
1	2	1	1/6
1	2	2	1/6
2	2	1	1/6
2	2	2	1/6

marginalised output $\bigoplus_{x_2} \Phi$

X_1	X_3	
1	1	1/3
1	2	1/3
2	1	1/6
2	2	1/6

$$H(O_1, O_3) = \frac{2}{3}\log 3 + \frac{2}{6}\log 6 \leq \log 5 = H(I_{13})$$

Putting Things Together

$$2 \log |\Phi| = 2H(O)$$ output tuples uniformly distributed

Putting Things Together

$$2 \log |\Phi| = 2H(O)$$ output tuples uniformly distributed

$$= 2\Big[H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big]$$

Putting Things Together

$$2\log|\Phi| = 2H(O) \qquad \text{output tuples uniformly distributed}$$

$$= 2\Big[H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big]$$

$$= \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big] +$$

$$\Big[H(O_1) + H(O_3 \mid O_1, O_2)\Big]$$

Putting Things Together

$$
\begin{aligned}
& 2\log|\Phi| = 2H(O) && \text{output tuples uniformly distributed} \\
&= 2\Big[H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big] \\
&= \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big] + \\
&\quad \Big[H(O_1) + H(O_3 \mid O_1, O_2)\Big] \\
&\leq \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2) + H(O_3 \mid O_2)\Big] + \\
&\quad \Big[H(O_1) + H(O_3 \mid O_1)\Big] && \text{dropping information cannot decrease entropy}
\end{aligned}
$$

Putting Things Together

$$
\begin{aligned}
& 2\log|\Phi| = 2H(O) && \text{output tuples uniformly distributed}\\
&= 2\Big[H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big]\\
&= \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big] +\\
&\quad \Big[H(O_1) + H(O_3 \mid O_1, O_2)\Big]\\
&\le \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2) + H(O_3 \mid O_2)\Big] +\\
&\quad \Big[H(O_1) + H(O_3 \mid O_1)\Big] && \text{dropping information cannot decrease entropy}\\
&= H(O_1, O_2) + H(O_2, O_3) + H(O_1, O_3) && \text{conditional entropies}
\end{aligned}
$$

Putting Things Together

$$
\begin{aligned}
2\log|\Phi| &= 2H(O) && \text{output tuples uniformly distributed}\\
&= 2\Big[H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big]\\
&= \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big] +\\
&\quad \Big[H(O_1) + H(O_3 \mid O_1, O_2)\Big]\\
&\le \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2) + H(O_3 \mid O_2)\Big] +\\
&\quad \Big[H(O_1) + H(O_3 \mid O_1)\Big] && \text{dropping information cannot decrease entropy}\\
&= H(O_1, O_2) + H(O_2, O_3) + H(O_1, O_3) && \text{conditional entropies}\\
&\le H(I_{12}) + H(I_{23}) + H(I_{13}) && \text{Observations 1, 2, and 3}
\end{aligned}
$$

Putting Things Together

$$
\begin{aligned}
& 2\log|\Phi| = 2H(O) && \text{output tuples uniformly distributed} \\
&= 2\Big[H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big] \\
&= \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big] + \\
&\quad \Big[H(O_1) + H(O_3 \mid O_1, O_2)\Big] \\
&\le \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2) + H(O_3 \mid O_2)\Big] + \\
&\quad \Big[H(O_1) + H(O_3 \mid O_1)\Big] && \text{dropping information cannot decrease entropy} \\
&= H(O_1, O_2) + H(O_2, O_3) + H(O_1, O_3) && \text{conditional entropies} \\
&\le H(I_{12}) + H(I_{23}) + H(I_{13}) && \text{Observations 1, 2, and 3} \\
&= \log N + \log N + \log N && \text{input tuples uniformly distributed}
\end{aligned}
$$

Putting Things Together

$$
\begin{aligned}
& 2\log|\Phi| = 2H(O) && \text{output tuples uniformly distributed} \\
&= 2\Big[H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big] \\
&= \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big] + \\
&\quad \Big[H(O_1) + H(O_3 \mid O_1, O_2)\Big] \\
&\le \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2) + H(O_3 \mid O_2)\Big] + \\
&\quad \Big[H(O_1) + H(O_3 \mid O_1)\Big] && \text{dropping information cannot decrease entropy} \\
&= H(O_1, O_2) + H(O_2, O_3) + H(O_1, O_3) && \text{conditional entropies} \\
&\le H(I_{12}) + H(I_{23}) + H(I_{13}) && \text{Observations 1, 2, and 3} \\
&= \log N + \log N + \log N && \text{input tuples uniformly distributed}
\end{aligned}
$$

$$\Longrightarrow |\Phi| \le N^{\frac{3}{2}} \qquad \text{as explained before}$$

Putting Things Together

$$
\begin{aligned}
2\log|\Phi| &= 2H(O) && \text{output tuples uniformly distributed}\\
&= 2\Big[H(O_1) + H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big]\\
&= \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2 \mid O_1) + H(O_3 \mid O_1, O_2)\Big] +\\
&\quad \Big[H(O_1) + H(O_3 \mid O_1, O_2)\Big]\\
&\le \Big[H(O_1) + H(O_2 \mid O_1)\Big] + \Big[H(O_2) + H(O_3 \mid O_2)\Big] +\\
&\quad \Big[H(O_1) + H(O_3 \mid O_1)\Big] && \text{dropping information cannot decrease entropy}\\
&= H(O_1, O_2) + H(O_2, O_3) + H(O_1, O_3) && \text{conditional entropies}\\
&\le H(I_{12}) + H(I_{23}) + H(I_{13}) && \text{Observations 1, 2, and 3}\\
&= \log N + \log N + \log N && \text{input tuples uniformly distributed}
\end{aligned}
$$

$$\implies |\Phi| \le N^{\frac{3}{2}} \qquad \text{as explained before}$$

We next generalise the approach taken in this example to arbitrary joins

General Case: Size Bound for Any Join

Quick Recap on Random Variables over Discrete Domains

- $\text{Dom}(X)$ is the domain of variable X
- For each $x \in \text{Dom}(X)$, we have a probability $P(X = x)$
- Joint Probability of random variables X and Y:

  Let $x \in \text{Dom}(X)$, $y \in \text{Dom}(Y)$.

  $P(X = x, Y = y)$ gives the joint probability of $X = x$ and $Y = y$

Quick Recap on Random Variables over Discrete Domains

- $\text{Dom}(X)$ is the domain of variable X
- For each $x \in \text{Dom}(X)$, we have a probability $P(X = x)$
- Joint Probability of random variables X and Y:

  Let $x \in \text{Dom}(X)$, $y \in \text{Dom}(Y)$.

  $P(X = x, Y = y)$ gives the joint probability of $X = x$ and $Y = y$
- Marginalised probability:

$$P(X = x) = \sum_{y} P(X = x, Y = y)$$

- Conditional probability: Assuming $P(Y = y) \neq 0$,

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

Entropy of Random Variable

- Entropy of a random variable X:

$$H(X) = -\sum_{x} P(X = x) \cdot \log P(X = x)$$

Intuitively: $H(X)$ measures the uncertainty about X

Entropy of Random Variable

- Entropy of a random variable X:

$$H(X) = -\sum_{x} P(X = x) \cdot \log P(X = x)$$

  Intuitively: $H(X)$ measures the uncertainty about X

- Joint entropy:

$$H(X, Y) = -\sum_{x,y} P(X = x, Y = y) \cdot \log P(X = x, Y = y)$$

- Conditional entropy: Assuming $P(Y = y) \neq 0$,

$$H(X|Y = y) = -\sum_{x} P(X = x|Y = y) \cdot \log P(X = x|Y = y)$$

$$H(X|Y) = \sum_{y} P(Y = y) \cdot H(X|Y = y)$$

Observations

Observation 1: The joint entropy of $\mathbf{X}_{[n]} = (X_1, \ldots, X_n)$ can be expressed as the sum of the entropies of each X_i conditioned on $\mathbf{X}_{[i-1]} = (X_1, \ldots, X_{i-1})$

$$H(\mathbf{X}_{[n]}) = H(X_1) + H(X_2 | X_1) + \ldots + H(X_n \mid \mathbf{X}_{[n-1]})$$

Observations

Observation 1: The joint entropy of $\mathbf{X}_{[n]} = (X_1, \ldots, X_n)$ can be expressed as the sum of the entropies of each X_i conditioned on $\mathbf{X}_{[i-1]} = (X_1, \ldots, X_{i-1})$

$$H(\mathbf{X}_{[n]}) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n \mid \mathbf{X}_{[n-1]})$$

Observation 2: The entropy of X conditioned on $\mathbf{X}_{[n]} = (X_1, \ldots, X_n)$ is not larger than the entropy of X conditioned on a subset $\mathbf{X}_J$ of $\mathbf{X}_{[n]}$

$$H(X \mid \mathbf{X}_{[n]}) \leq H(X \mid \mathbf{X}_J) \text{ for all } J \subseteq [n]$$

Shearer's Lemma

Let

- $\mathbf{X}_{[n]} = (X_1, \ldots, X_n)$ are random variables
- $\mathcal{J} \subseteq 2^{[n]}$ is multiset such that each $i \in [n]$ is in at least q members of $\mathcal{J}$
  - $2^{[n]}$ is the set of all possible subsets of $[n] = \{1, \ldots, n\}$
  - $\mathcal{J}$ is a subset of $2^{[n]}$, but possibly with repetitions (hence, multiset)
  - $\mathcal{J}$ is like the set of hyperedges of a multi-hypergraph whose set of nodes is $[n]$

Then,

$$q \cdot H(\mathbf{X}_{[n]}) \leq \sum_{J \in \mathcal{J}} H(\mathbf{X}_J)$$

Example

Triangle Query $\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and output:

output Φ

X_1	X_2	X_3
1	1	1
1	1	2
1	2	1
1	2	2
2	2	1
2	2	2

Example

Triangle Query $\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and output:

output Φ

X_1	X_2	X_3
1	1	1
1	1	2
1	2	1
1	2	2
2	2	1
2	2	2

- Choose $\mathcal{J} = \mathcal{E} = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}$
- Each $i \in [3]$ occurs in at least two members of $\mathcal{J}$

Example

Triangle Query $\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and output:

output Φ

X_1	X_2	X_3	
1	1	1	$1/6$
1	1	2	$1/6$
1	2	1	$1/6$
1	2	2	$1/6$
2	2	1	$1/6$
2	2	2	$1/6$

- Choose $\mathcal{J} = \mathcal{E} = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}$
- Each $i \in [3]$ occurs in at least two members of $\mathcal{J}$

$2H(O) = 2 \log 6 \approx 1.56$

Example

Triangle Query $\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and output:

output Φ

X_1	X_2	X_3	
1	1	1	$1/6$
1	1	2	$1/6$
1	2	1	$1/6$
1	2	2	$1/6$
2	2	1	$1/6$
2	2	2	$1/6$

marginalised output $\bigoplus_{x_3} \Phi$

X_1	X_2	
1	1	$1/3$
1	2	$1/3$
2	2	$1/3$

- Choose $\mathcal{J} = \mathcal{E} = \{\{1,2\},\{2,3\},\{1,3\}\}$
- Each $i \in [3]$ occurs in at least two members of $\mathcal{J}$

$2H(O) = 2\log 6 \approx 1.56$ $\log 3$

$H(O_1, O_2)$

Example

Triangle Query $\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and output:

output Φ

X_1	X_2	X_3	
1	1	1	$1/6$
1	1	2	$1/6$
1	2	1	$1/6$
1	2	2	$1/6$
2	2	1	$1/6$
2	2	2	$1/6$

marginalised output $\bigoplus_{x_3} \Phi$

X_1	X_2	
1	1	$1/3$
1	2	$1/3$
2	2	$1/3$

marginalised output $\bigoplus_{x_1} \Phi$

X_2	X_3	
1	1	$1/6$
1	2	$1/6$
2	1	$1/3$
2	2	$1/3$

- Choose $\mathcal{J} = \mathcal{E} = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}$
- Each $i \in [3]$ occurs in at least two members of $\mathcal{J}$

$2H(O) = 2 \log 6 \approx 1.56$

$\log 3 + \frac{2}{6} \log 6 + \frac{2}{3} \log 3$

$H(O_1, O_2)$ $\quad$ $H(O_2, O_3)$

Example

Triangle Query $\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and output:

output Φ

X_1	X_2	X_3	
1	1	1	$1/6$
1	1	2	$1/6$
1	2	1	$1/6$
1	2	2	$1/6$
2	2	1	$1/6$
2	2	2	$1/6$

marginalised output $\bigoplus_{x_3} \Phi$

X_1	X_2	
1	1	$1/3$
1	2	$1/3$
2	2	$1/3$

marginalised output $\bigoplus_{x_1} \Phi$

X_2	X_3	
1	1	$1/6$
1	2	$1/6$
2	1	$1/3$
2	2	$1/3$

marginalised output $\bigoplus_{x_2} \Phi$

X_1	X_3	
1	1	$1/3$
1	2	$1/3$
2	1	$1/6$
2	2	$1/6$

- Choose $\mathcal{J} = \mathcal{E} = \{\{1,2\},\{2,3\},\{1,3\}\}$
- Each $i \in [3]$ occurs in at least two members of $\mathcal{J}$

$2H(O) = 2\log 6 \approx 1.56$

$$\underbrace{\log 3}_{H(O_1, O_2)} + \underbrace{\frac{2}{6}\log 6 + \frac{2}{3}\log 3}_{H(O_2, O_3)} + \underbrace{\frac{2}{6}\log 6 + \frac{2}{3}\log 3}_{H(O_1, O_3)}$$

Example

Triangle Query $\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$

with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and output:

output Φ

X_1	X_2	X_3	
1	1	1	$1/6$
1	1	2	$1/6$
1	2	1	$1/6$
1	2	2	$1/6$
2	2	1	$1/6$
2	2	2	$1/6$

marginalised output $\bigoplus_{x_3} \Phi$

X_1	X_2	
1	1	$1/3$
1	2	$1/3$
2	2	$1/3$

marginalised output $\bigoplus_{x_1} \Phi$

X_2	X_3	
1	1	$1/6$
1	2	$1/6$
2	1	$1/3$
2	2	$1/3$

marginalised output $\bigoplus_{x_2} \Phi$

X_1	X_3	
1	1	$1/3$
1	2	$1/3$
2	1	$1/6$
2	2	$1/6$

- Choose $\mathcal{J} = \mathcal{E} = \{\{1,2\},\{2,3\},\{1,3\}\}$
- Each $i \in [3]$ occurs in at least two members of $\mathcal{J}$

$$2H(O) = 2\log 6 \approx 1.56 \leq 1.63 \approx \underbrace{\log 3}_{H(O_1,O_2)} + \underbrace{\frac{2}{6}\log 6 + \frac{2}{3}\log 3}_{H(O_2,O_3)} + \underbrace{\frac{2}{6}\log 6 + \frac{2}{3}\log 3}_{H(O_1,O_3)}$$

Proof of Shearer's Lemma

$$
\begin{aligned}
& q \cdot H(\mathbf{X}_{[n]}) \\
&= q \cdot \sum_{i \in [n]} H(X_i \mid \mathbf{X}_{[i-1]}) \quad \text{Observation 1 on chain rule for joint entropy}
\end{aligned}
$$

Proof of Shearer's Lemma

$$
\begin{aligned}
& q \cdot H(\mathbf{X}_{[n]}) \\
&= q \cdot \sum_{i \in [n]} H(X_i \mid \mathbf{X}_{[i-1]}) \quad \text{Observation 1 on chain rule for joint entropy} \\
&= \quad q \cdot H(X_1) \; + \quad q \cdot H(X_2 \mid X_1) \; + \ldots + \quad q \cdot H(X_n \mid \mathbf{X}_{[n-1]})
\end{aligned}
$$

Proof of Shearer's Lemma

$$q \cdot H(\mathbf{X}_{[n]})$$

$$= q \cdot \sum_{i \in [n]} H(X_i \mid \mathbf{X}_{[i-1]}) \quad \text{Observation 1 on chain rule for joint entropy}$$

$$= \quad q \cdot H(X_1) \quad + \quad q \cdot H(X_2 \mid X_1) \quad + \ldots + \quad q \cdot H(X_n \mid \mathbf{X}_{[n-1]})$$

$$\qquad \rotatebox{90}{$\leq$} \qquad\qquad\qquad \rotatebox{90}{$\leq$} \qquad\qquad\qquad \rotatebox{90}{$\leq$}$$

$$\leq \sum_{J \in \mathcal{J} : 1 \in J} H(X_1) + \sum_{J \in \mathcal{J} : 2 \in J} H(X_2 \mid X_1) + \ldots + \sum_{J \in \mathcal{J} : n \in J} H(X_n \mid \mathbf{X}_{[n-1]})$$

Since each i appears in at least q sets

Proof of Shearer's Lemma

$$q \cdot H(\mathbf{X}_{[n]})$$

$$= q \cdot \sum_{i \in [n]} H(X_i \mid \mathbf{X}_{[i-1]}) \quad \text{Observation 1 on chain rule for joint entropy}$$

$$= \quad q \cdot H(X_1) \quad + \quad q \cdot H(X_2 \mid X_1) \quad + \ldots + \quad q \cdot H(X_n \mid \mathbf{X}_{[n-1]})$$

$$\qquad \rotatebox{90}{$\geq$} \qquad\qquad\qquad \rotatebox{90}{$\geq$} \qquad\qquad\qquad \rotatebox{90}{$\geq$}$$

$$\leq \sum_{J \in \mathcal{J} : 1 \in J} H(X_1) + \sum_{J \in \mathcal{J} : 2 \in J} H(X_2 \mid X_1) + \ldots + \sum_{J \in \mathcal{J} : n \in J} H(X_n \mid \mathbf{X}_{[n-1]})$$

Since each i appears in at least q sets

$$\leq \sum_{J \in \mathcal{J} : 1 \in J} H(X_1) + \sum_{J \in \mathcal{J} : 2 \in J} H(X_2 \mid X_{\{1\} \cap J}) + \ldots + \sum_{J \in \mathcal{J} : n \in J} H(X_n \mid \mathbf{X}_{[n-1] \cap J})$$

Observation 2: Conditioning on less variables does not decrease entropy

Proof of Shearer's Lemma

$$q \cdot H(\mathbf{X}_{[n]})$$

$$= q \cdot \sum_{i \in [n]} H(X_i \mid \mathbf{X}_{[i-1]})$$ Observation 1 on chain rule for joint entropy

$$= \quad q \cdot H(X_1) \quad + \quad q \cdot H(X_2 \mid X_1) \quad + \ldots + \quad q \cdot H(X_n \mid \mathbf{X}_{[n-1]})$$

$$\quad \rotatebox{90}{$\geq$} \qquad\qquad \rotatebox{90}{$\geq$} \qquad\qquad \rotatebox{90}{$\geq$}$$

$$\leq \sum_{J \in \mathcal{J}: 1 \in J} H(X_1) + \sum_{J \in \mathcal{J}: 2 \in J} H(X_2 \mid X_1) + \ldots + \sum_{J \in \mathcal{J}: n \in J} H(X_n \mid \mathbf{X}_{[n-1]})$$

Since each i appears in at least q sets

$$\leq \sum_{J \in \mathcal{J}: 1 \in J} H(X_1) + \sum_{J \in \mathcal{J}: 2 \in J} H(X_2 \mid X_{\{1\} \cap J}) + \ldots + \sum_{J \in \mathcal{J}: n \in J} H(X_n \mid \mathbf{X}_{[n-1] \cap J})$$

Observation 2: Conditioning on less variables does not decrease entropy

$$= \sum_{J \in \mathcal{J}} \sum_{i \in J} H(X_i \mid \mathbf{X}_{[i-1] \cap J})$$

Proof of Shearer's Lemma

$$q \cdot H(\mathbf{X}_{[n]})$$

$$= q \cdot \sum_{i \in [n]} H(X_i \mid \mathbf{X}_{[i-1]})$$ Observation 1 on chain rule for joint entropy

$$= \quad q \cdot H(X_1) \quad + \quad q \cdot H(X_2 \mid X_1) \quad + \ldots + \quad q \cdot H(X_n \mid \mathbf{X}_{[n-1]})$$

$$\leq \qquad\qquad\qquad \leq \qquad\qquad\qquad \leq$$

$$\leq \sum_{J \in \mathcal{J} : 1 \in J} H(X_1) + \sum_{J \in \mathcal{J} : 2 \in J} H(X_2 \mid X_1) + \ldots + \sum_{J \in \mathcal{J} : n \in J} H(X_n \mid \mathbf{X}_{[n-1]})$$

Since each i appears in at least q sets

$$\leq \sum_{J \in \mathcal{J} : 1 \in J} H(X_1) + \sum_{J \in \mathcal{J} : 2 \in J} H(X_2 \mid X_{\{1\} \cap J}) + \ldots + \sum_{J \in \mathcal{J} : n \in J} H(X_n \mid \mathbf{X}_{[n-1] \cap J})$$

Observation 2: Conditioning on less variables does not decrease entropy

$$= \sum_{J \in \mathcal{J}} \sum_{i \in J} H(X_i \mid \mathbf{X}_{[i-1] \cap J}) = \sum_{J \in \mathcal{J}} H(\mathbf{X}_J)$$ Observation 1 on chain rule

Connection to Join Output Size

FAQ $\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and factor sizes $(N_S)_{S \in \mathcal{E}}$

- Let $(w_S)_{S \in \mathcal{E}}$ be any feasible solution to the linear program computing $\rho^*(\mathcal{H})$ with minimisation objective $\prod_{S \in \mathcal{E}} N_S^{w_S}$

Why can we apply Shearer's lemma in our case?

Connection to Join Output Size

FAQ $\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and factor sizes $(N_S)_{S \in \mathcal{E}}$

- Let $(w_S)_{S \in \mathcal{E}}$ be any feasible solution to the linear program computing $\rho^*(\mathcal{H})$ with minimisation objective $\prod_{S \in \mathcal{E}} N_S^{w_S}$

Why can we apply Shearer's lemma in our case?

- Each factor ψ_S = joint distribution over the random variables in S
- Hyperedges $S \in \mathcal{E}$ = sets $J \in \mathcal{J}$ in Shearer's lemma; more precisely:
  - Choose natural numbers q and $(p_S)_{S \in \mathcal{E}}$ such that $w_S = \frac{p_S}{q}$ for all $S \in \mathcal{E}$
  - Let $\mathcal{J} \subseteq 2^{[n]}$ be a multiset that consists of p_S copies of each $S \in \mathcal{E}$
- We still need to hold: every $i \in [n]$ occurs in at least q sets in $\mathcal{J}$

Connection to Join Output Size

FAQ $\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and factor sizes $(N_S)_{S \in \mathcal{E}}$

- Let $(w_S)_{S \in \mathcal{E}}$ be any feasible solution to the linear program computing $\rho^*(\mathcal{H})$ with minimisation objective $\prod_{S \in \mathcal{E}} N_S^{w_S}$

Why can we apply Shearer's lemma in our case?

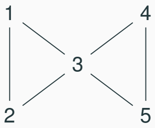
- Each factor ψ_S = joint distribution over the random variables in S
- Hyperedges $S \in \mathcal{E}$ = sets $J \in \mathcal{J}$ in Shearer's lemma; more precisely:
  - Choose natural numbers q and $(p_S)_{S \in \mathcal{E}}$ such that $w_S = \frac{p_S}{q}$ for all $S \in \mathcal{E}$
  - Let $\mathcal{J} \subseteq 2^{[n]}$ be a multiset that consists of p_S copies of each $S \in \mathcal{E}$
- We still need to hold: every $i \in [n]$ occurs in at least q sets in $\mathcal{J}$

  This holds because the number of sets containing i is:

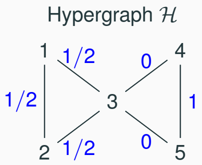
$$\sum_{S \in \mathcal{J}: i \in S} p_S = \sum_{S \in \mathcal{J}: i \in S} q \cdot w_S = q \cdot \underbrace{\sum_{S \in \mathcal{J}: i \in S} w_S}_{\geq 1 \text{ due to linear program}} \geq q$$

Example Connecting Shearer Setup with Feasible Solution for ρ^*

Hypergraph $\mathcal{H}$



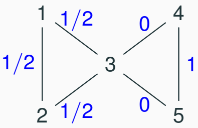
Example Connecting Shearer Setup with Feasible Solution for ρ^*



- Feasible solution to the linear program computing $\rho^*(\mathcal{H})$:
  $w_{12} = w_{23} = w_{13} = \frac{1}{2}$, $w_{34} = w_{35} = 0$, $w_{45} = 1$

Example Connecting Shearer Setup with Feasible Solution for ρ^*

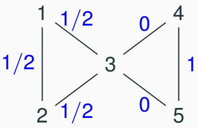
Hypergraph $\mathcal{H}$



- Feasible solution to the linear program computing $\rho^*(\mathcal{H})$:
  $w_{12} = w_{23} = w_{13} = \frac{1}{2}$, $w_{34} = w_{35} = 0$, $w_{45} = 1$
- We can choose $q = 2$, $p_{12} = p_{23} = p_{13} = 1$, $p_{34} = p_{35} = 0$, and $p_{45} = 2$, since $w_{12} = w_{23} = w_{13} = \frac{1}{2}$, $w_{34} = w_{35} = \frac{0}{2}$, and $w_{45} = \frac{2}{2}$

Example Connecting Shearer Setup with Feasible Solution for ρ^*

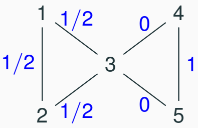
Hypergraph $\mathcal{H}$



- Feasible solution to the linear program computing $\rho^*(\mathcal{H})$: $w_{12} = w_{23} = w_{13} = \frac{1}{2}$, $w_{34} = w_{35} = 0$, $w_{45} = 1$
- We can choose $q = 2$, $p_{12} = p_{23} = p_{13} = 1$, $p_{34} = p_{35} = 0$, and $p_{45} = 2$, since $w_{12} = w_{23} = w_{13} = \frac{1}{2}$, $w_{34} = w_{35} = \frac{0}{2}$, and $w_{45} = \frac{2}{2}$
- Then, $\mathcal{J} = \{\{1,2\},\{2,3\},\{1,3\},\{4,5\},\{4,5\}\}$

Example Connecting Shearer Setup with Feasible Solution for ρ^*

Hypergraph $\mathcal{H}$



- Feasible solution to the linear program computing $\rho^*(\mathcal{H})$:
  $w_{12} = w_{23} = w_{13} = \frac{1}{2}$, $w_{34} = w_{35} = 0$, $w_{45} = 1$
- We can choose $q = 2$, $p_{12} = p_{23} = p_{13} = 1$, $p_{34} = p_{35} = 0$, and $p_{45} = 2$, since $w_{12} = w_{23} = w_{13} = \frac{1}{2}$, $w_{34} = w_{35} = \frac{0}{2}$, and $w_{45} = \frac{2}{2}$
- Then, $\mathcal{J} = \{\{1,2\},\{2,3\},\{1,3\},\{4,5\},\{4,5\}\}$

$\implies$ Every $i \in [5]$ occurs in 2 sets in $\mathcal{J}$.

Putting Things Together

W.l.o.g assume $|\Phi| \neq 0$, otherwise the size bound trivially holds

Let $X = (\mathbf{X}_{[n]})$ be uniformly distributed over the output Φ

Putting Things Together

W.l.o.g assume $|\Phi| \neq 0$, otherwise the size bound trivially holds

Let $X = (\mathbf{X}_{[n]})$ be uniformly distributed over the output Φ

$$\log |\Phi| = H(X) \qquad X \text{ is uniformly distributed}$$

Putting Things Together

W.l.o.g assume $|\Phi| \neq 0$, otherwise the size bound trivially holds

Let $X = (\mathbf{X}_{[n]})$ be uniformly distributed over the output Φ

$$\begin{aligned} \log|\Phi| = &\ H(X) && X \text{ is uniformly distributed} \\ \leq &\ \frac{1}{q} \cdot \sum_{J \in \mathcal{J}} H(\mathbf{X}_J) && \text{Shearer's Lemma} \end{aligned}$$

Putting Things Together

W.l.o.g assume $|\Phi| \neq 0$, otherwise the size bound trivially holds

Let $X = (\mathbf{X}_{[n]})$ be uniformly distributed over the output Φ

$$
\begin{aligned}
\log|\Phi| = \ & H(X) && X \text{ is uniformly distributed} \\
\leq \ & \frac{1}{q} \cdot \sum_{J \in \mathcal{J}} H(\mathbf{X}_J) && \text{Shearer's Lemma} \\
= \ & \frac{1}{q} \cdot \sum_{S \in \mathcal{E}} p_S \cdot H(\mathbf{X}_S) && \mathcal{J} \text{ consists of } p_S \text{ copies of each } S \in \mathcal{E}
\end{aligned}
$$

Putting Things Together

W.l.o.g assume $|\Phi| \neq 0$, otherwise the size bound trivially holds

Let $X = (\mathbf{X}_{[n]})$ be uniformly distributed over the output Φ

$$\begin{aligned}
\log|\Phi| &= H(X) && X \text{ is uniformly distributed} \\
&\leq \frac{1}{q} \cdot \sum_{J \in \mathcal{J}} H(\mathbf{X}_J) && \text{Shearer's Lemma} \\
&= \frac{1}{q} \cdot \sum_{S \in \mathcal{E}} p_S \cdot H(\mathbf{X}_S) && \mathcal{J} \text{ consists of } p_S \text{ copies of each } S \in \mathcal{E} \\
&\leq \sum_{S \in \mathcal{E}} w_S \cdot H(\mathbf{X}_S) && w_s = \frac{p_S}{q}
\end{aligned}$$

Putting Things Together

W.l.o.g assume $|\Phi| \neq 0$, otherwise the size bound trivially holds

Let $X = (\mathbf{X}_{[n]})$ be uniformly distributed over the output Φ

$$
\begin{aligned}
\log|\Phi| = \; & H(X) && X \text{ is uniformly distributed} \\
\leq \; & \frac{1}{q} \cdot \sum_{J \in \mathcal{J}} H(\mathbf{X}_J) && \text{Shearer's Lemma} \\
= \; & \frac{1}{q} \cdot \sum_{S \in \mathcal{E}} p_S \cdot H(\mathbf{X}_S) && \mathcal{J} \text{ consists of } p_S \text{ copies of each } S \in \mathcal{E} \\
\leq \; & \sum_{S \in \mathcal{E}} w_S \cdot H(\mathbf{X}_S) && w_s = \frac{p_S}{q} \\
\leq \; & \sum_{S \in \mathcal{E}} w_S \cdot \log N_S && H(\mathbf{X}_S) \leq \log N_S
\end{aligned}
$$

Putting Things Together

W.l.o.g assume $|\Phi| \neq 0$, otherwise the size bound trivially holds

Let $X = (\mathbf{X}_{[n]})$ be uniformly distributed over the output Φ

$$\begin{aligned}
\log|\Phi| = \; & H(X) && X \text{ is uniformly distributed} \\
\leq \; & \frac{1}{q} \cdot \sum_{J \in \mathcal{J}} H(\mathbf{X}_J) && \text{Shearer's Lemma} \\
= \; & \frac{1}{q} \cdot \sum_{S \in \mathcal{E}} p_S \cdot H(\mathbf{X}_S) && \mathcal{J} \text{ consists of } p_S \text{ copies of each } S \in \mathcal{E} \\
\leq \; & \sum_{S \in \mathcal{E}} w_S \cdot H(\mathbf{X}_S) && w_s = \frac{p_S}{q} \\
\leq \; & \sum_{S \in \mathcal{E}} w_S \cdot \log N_S && H(\mathbf{X}_S) \leq \log N_S
\end{aligned}$$

This implies:

$$\log|\Phi| \leq \sum_{S \in \mathcal{E}} \log N_S^{w_S}$$

Putting Things Together

W.l.o.g assume $|\Phi| \neq 0$, otherwise the size bound trivially holds

Let $X = (\mathbf{X}_{[n]})$ be uniformly distributed over the output Φ

$$
\begin{aligned}
\log|\Phi| = \ & H(X) && X \text{ is uniformly distributed} \\
\leq \ & \frac{1}{q} \cdot \sum_{J \in \mathcal{J}} H(\mathbf{X}_J) && \text{Shearer's Lemma} \\
= \ & \frac{1}{q} \cdot \sum_{S \in \mathcal{E}} p_S \cdot H(\mathbf{X}_S) && \mathcal{J} \text{ consists of } p_S \text{ copies of each } S \in \mathcal{E} \\
\leq \ & \sum_{S \in \mathcal{E}} w_S \cdot H(\mathbf{X}_S) && w_s = \frac{p_S}{q} \\
\leq \ & \sum_{S \in \mathcal{E}} w_S \cdot \log N_S && H(\mathbf{X}_S) \leq \log N_S
\end{aligned}
$$

This implies:

$$
\log|\Phi| \leq \sum_{S \in \mathcal{E}} \log N_S^{w_S} \Leftrightarrow \log|\Phi| \leq \log \prod_{S \in \mathcal{E}} N_S^{w_S}
$$

Putting Things Together

W.l.o.g assume $|\Phi| \neq 0$, otherwise the size bound trivially holds

Let $X = (\mathbf{X}_{[n]})$ be uniformly distributed over the output Φ

$$\begin{aligned}
\log|\Phi| &= H(X) && X \text{ is uniformly distributed} \\
&\leq \frac{1}{q} \cdot \sum_{J \in \mathcal{J}} H(\mathbf{X}_J) && \text{Shearer's Lemma} \\
&= \frac{1}{q} \cdot \sum_{S \in \mathcal{E}} p_S \cdot H(\mathbf{X}_S) && \mathcal{J} \text{ consists of } p_S \text{ copies of each } S \in \mathcal{E} \\
&\leq \sum_{S \in \mathcal{E}} w_S \cdot H(\mathbf{X}_S) && w_s = \frac{p_S}{q} \\
&\leq \sum_{S \in \mathcal{E}} w_S \cdot \log N_S && H(\mathbf{X}_S) \leq \log N_S
\end{aligned}$$

This implies:

$$\log|\Phi| \leq \sum_{S \in \mathcal{E}} \log N_S^{w_S} \Leftrightarrow \log|\Phi| \leq \log \prod_{S \in \mathcal{E}} N_S^{w_S} \Leftrightarrow |\Phi| \leq \prod_{S \in \mathcal{E}} N_S^{w_S}$$

The Lower Bound Argument

Lower Bound for Join Output Size

Consider an FAQ join $\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$

We have shown:

- If input factors ψ_S are of size N, then $|\Phi| \leq N^{\rho^*(\mathcal{H})}$

Lower Bound for Join Output Size

Consider an FAQ join $\Phi(\mathbf{x}) = \bigotimes_{\mathcal{S}\in\mathcal{E}} \psi_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}})$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$

We have shown:

- If input factors $\psi_{\mathcal{S}}$ are of size N, then $|\Phi| \leq N^{\rho^*(\mathcal{H})}$

What we would like to show in the ideal case:

- If input factors $\psi_{\mathcal{S}}$ are of size N, then $|\Phi| \geq N^{\rho^*(\mathcal{H})}$
- This is not always possible

Lower Bound for Join Output Size

Consider an FAQ join $\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$

We have shown:

- If input factors ψ_S are of size N, then $|\Phi| \leq N^{\rho^*(\mathcal{H})}$

What we would like to show in the ideal case:

- If input factors ψ_S are of size N, then $|\Phi| \geq N^{\rho^*(\mathcal{H})}$
- This is not always possible

We can however show:

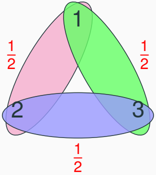
- For every N_0, we construct factors of size $N \geq N_0$ such that $|\Phi| \geq N^{\rho^*(\mathcal{H})}$
- This lower bound extends to factors of different sizes

Warm-Up: Size Bound for Triangle Join

Lower Bound on Triangle Join Output Size (1/2)

$$\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$$

Hypergraph $\mathcal{H}$



$$\rho^*(\mathcal{H}) = \frac{3}{2}$$

Lower Bound on Triangle Join Output Size (2/2)

$$\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$$

- We can construct input factors ψ_{ij} of size 4 with $|\Phi| = 4^{\frac{3}{2}} = 8$.

input ψ_{12}

X_1	X_2
1	1
1	2
2	1
2	2

= [2] $\times$ [2]

input ψ_{23}

X_2	X_3
1	1
1	2
2	1
2	2

= [2] $\times$ [2]

input ψ_{13}

X_1	X_3
1	1
1	2
2	1
2	2

= [2] $\times$ [2]

output Φ

X_1	X_2	X_3
1	1	1
1	1	2
1	2	1
1	2	2
2	1	1
2	1	2
2	2	1
2	2	2

Lower Bound on Triangle Join Output Size (2/2)

$$\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$$

- We can construct input factors ψ_{ij} of size 4 with $|\Phi| = 4^{\frac{3}{2}} = 8$.

input ψ_{12}

X_1	X_2
1	1
1	2
2	1
2	2

= [2] $\times$ [2]

input ψ_{23}

X_2	X_3
1	1
1	2
2	1
2	2

= [2] $\times$ [2]

input ψ_{13}

X_1	X_3
1	1
1	2
2	1
2	2

= [2] $\times$ [2]

output Φ

X_1	X_2	X_3
1	1	1
1	1	2
1	2	1
1	2	2
2	1	1
2	1	2
2	2	1
2	2	2

- We next generalise the idea of this construction

Dual Linear Program

The dual of the linear program computing the fractional edge cover number ρ^*

LP for $\rho^*(\mathcal{H})$	Dual LP for $D(\mathcal{H})$
minimise $\sum_{S\in\mathcal{E}} w_S$	maximise $\sum_{i\in[n]} v_i$
subject to $\sum_{S\in\mathcal{E}:v\in S} w_S \geq 1 \quad \forall v \in \mathcal{V},$	subject to $\sum_{i\in S} v_i \leq 1 \quad \forall S \in \mathcal{E},$
$0 \leq w_S \leq 1 \quad \forall S \in \mathcal{E}$	$0 \leq v_i \leq 1 \quad \forall i \in [n]$

- Left: Weights w_S assigned to hyperedges
- Right: Weights v_i assigned to nodes

Dual Linear Program

The dual of the linear program computing the fractional edge cover number ρ^*

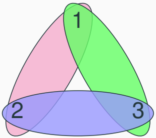
LP for $\rho^*(\mathcal{H})$	Dual LP for $D(\mathcal{H})$
minimise $\sum_{S\in\mathcal{E}} w_S$	maximise $\sum_{i\in[n]} v_i$
subject to $\sum_{S\in\mathcal{E}:v\in S} w_S \geq 1 \quad \forall v \in \mathcal{V},$	subject to $\sum_{i\in S} v_i \leq 1 \quad \forall S \in \mathcal{E},$
$0 \leq w_S \leq 1 \quad \forall S \in \mathcal{E}$	$0 \leq v_i \leq 1 \quad \forall i \in [n]$

- Left: Weights w_S assigned to hyperedges
- Right: Weights v_i assigned to nodes

By linear program duality: $\rho^*(\mathcal{H}) = D(\mathcal{H})$

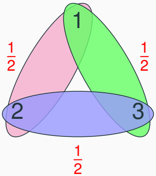
Dual LP for Triangle Join

$$\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$$



Dual LP for Triangle Join

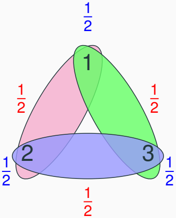
$$\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$$



$$\rho^*(\mathcal{H}) = \frac{3}{2}$$

Dual LP for Triangle Join

$$\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$$

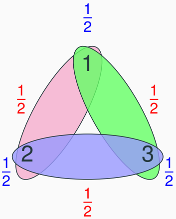


$$\rho^*(\mathcal{H}) = \frac{3}{2}$$

$$D(\mathcal{H}) = \frac{3}{2}$$

Dual LP for Triangle Join

$$\Phi(x_1, x_2, x_3) = \psi_{12}(x_1, x_2) \otimes \psi_{23}(x_2, x_3) \otimes \psi_{13}(x_1, x_3)$$



$\rho^*(\mathcal{H}) = \frac{3}{2}$

$D(\mathcal{H}) = \frac{3}{2}$

For factors size N_0, take $N \geq N_0$ a power of 2.

Choose $p, q \in \mathbb{N}$ such that $\frac{1}{2} \cdot \log N = \frac{p}{q}$.

We construct $\psi_{12} = \psi_{13} = \psi_{23} = [2^p] \times [2^p]$ and then

- $|\psi_{12}| = |\psi_{13}| = |\psi_{23}| = 2^{2p} = 2^{q \log N} = (2^{\log N})^q = N^q$
- $|\Phi| = 2^{3p} = 2^{3q\frac{1}{2} \log N} = (2^{\log N})^{q\frac{3}{2}} = N^{q\frac{3}{2}} = (N^q)^{\frac{3}{2}}$

Size Lower Bound for Any Join

Construction of Input Factors

- Consider an optimal solution $(v_i)_{i\in[n]}$ to the linear program computing $D(\mathcal{H})$
- Choose natural numbers q, $(p_i)_{i\in[n]}$ such that $v_i \cdot \log N = \frac{p_i}{q}$
  - This works if $N \geq N_0$ is a power of 2, so $\log N$ is a natural number
- We construct in two steps input factors ψ_S of size N^q such that

$$|\Phi| \geq (N^q)^{\rho^*(\mathcal{H})}$$

Construction of Input Factors: Step 1

For each $S \in \mathcal{E}$, construct ψ'_S as the Cartesian product

$$\psi'_S = \times_{i \in S}[2^{p_i}]$$

Construction of Input Factors: Step 1

For each $S \in \mathcal{E}$, construct ψ'_S as the Cartesian product

$$\psi'_S = \times_{i \in S}[2^{p_i}]$$

This implies

$$|\psi'_S| = \prod_{i \in S} 2^{p_i} = \prod_{i \in S} 2^{q \cdot v_i \cdot \log N} \qquad\qquad p_i = q \cdot v_i \cdot \log N$$

Construction of Input Factors: Step 1

For each $S \in \mathcal{E}$, construct ψ'_S as the Cartesian product

$$\psi'_S = \times_{i \in S}[2^{p_i}]$$

This implies

$$|\psi'_S| = \prod_{i \in S} 2^{p_i} = \prod_{i \in S} 2^{q \cdot v_i \cdot \log N} \qquad p_i = q \cdot v_i \cdot \log N$$

$$= \prod_{i \in S} 2^{\log N^{q \cdot v_i}}$$

Construction of Input Factors: Step 1

For each $S \in \mathcal{E}$, construct ψ'_S as the Cartesian product

$$\psi'_S = \times_{i \in S}[2^{p_i}]$$

This implies

$$|\psi'_S| = \prod_{i \in S} 2^{p_i} = \prod_{i \in S} 2^{q \cdot v_i \cdot \log N} \qquad p_i = q \cdot v_i \cdot \log N$$

$$= \prod_{i \in S} 2^{\log N^{q \cdot v_i}}$$

$$= \prod_{i \in S} N^{q \cdot v_i}$$

Construction of Input Factors: Step 1

For each $S \in \mathcal{E}$, construct ψ'_S as the Cartesian product

$$\psi'_S = \times_{i \in S}[2^{p_i}]$$

This implies

$$|\psi'_S| = \prod_{i \in S} 2^{p_i} = \prod_{i \in S} 2^{q \cdot v_i \cdot \log N} \qquad p_i = q \cdot v_i \cdot \log N$$

$$= \prod_{i \in S} 2^{\log N^{q \cdot v_i}}$$

$$= \prod_{i \in S} N^{q \cdot v_i}$$

$$= \prod_{i \in S} (N^q)^{v_i}$$

Construction of Input Factors: Step 1

For each $S \in \mathcal{E}$, construct ψ'_S as the Cartesian product

$$\psi'_S = \times_{i \in S}[2^{p_i}]$$

This implies

$$|\psi'_S| = \prod_{i \in S} 2^{p_i} = \prod_{i \in S} 2^{q \cdot v_i \cdot \log N} \qquad p_i = q \cdot v_i \cdot \log N$$

$$= \prod_{i \in S} 2^{\log N^{q \cdot v_i}}$$

$$= \prod_{i \in S} N^{q \cdot v_i}$$

$$= \prod_{i \in S} (N^q)^{v_i}$$

$$= (N^q)^{\sum_{i \in S} v_i}$$

Construction of Input Factors: Step 1

For each $S \in \mathcal{E}$, construct ψ'_S as the Cartesian product

$$\psi'_S = \times_{i \in S}[2^{p_i}]$$

This implies

$$\begin{aligned}
|\psi'_S| = \prod_{i \in S} 2^{p_i} &= \prod_{i \in S} 2^{q \cdot v_i \cdot \log N} && p_i = q \cdot v_i \cdot \log N \\
&= \prod_{i \in S} 2^{\log N^{q \cdot v_i}} \\
&= \prod_{i \in S} N^{q \cdot v_i} \\
&= \prod_{i \in S} (N^q)^{v_i} \\
&= (N^q)^{\sum_{i \in S} v_i} \\
&\leq N^q && \sum_{i \in S} v_i \leq 1
\end{aligned}$$

Construction of Input Factors: Step 2

For each $S \in \mathcal{E}$, construct an arbitrary ψ_S with $\psi_S \supseteq \psi'_S$ and $|\psi_S| = N^q$

Construction of Input Factors: Step 2

For each $S \in \mathcal{E}$, construct an arbitrary ψ_S with $\psi_S \supseteq \psi'_S$ and $|\psi_S| = N^q$

This implies

$$\Phi \supseteq \times_{i\in[n]}[2^{p_i}]$$

Construction of Input Factors: Step 2

For each $S \in \mathcal{E}$, construct an arbitrary ψ_S with $\psi_S \supseteq \psi'_S$ and $|\psi_S| = N^q$

This implies

$$\Phi \supseteq \times_{i \in [n]} [2^{p_i}]$$

Hence,

$$\begin{aligned} |\Phi| &\geq \prod_{i \in [n]} 2^{p_i} \\ &= \prod_{i \in [n]} 2^{q \cdot v_i \cdot \log N} \qquad\qquad p_i = q \cdot v_i \cdot \log N \end{aligned}$$

Construction of Input Factors: Step 2

For each $S \in \mathcal{E}$, construct an arbitrary ψ_S with $\psi_S \supseteq \psi'_S$ and $|\psi_S| = N^q$

This implies

$$\Phi \supseteq \times_{i \in [n]} [2^{p_i}]$$

Hence,

$$\begin{aligned} |\Phi| &\geq \prod_{i \in [n]} 2^{p_i} \\ &= \prod_{i \in [n]} 2^{q \cdot v_i \cdot \log N} && p_i = q \cdot v_i \cdot \log N \\ &= (N^q)^{\sum_{i \in [n]} v_i} && \text{analogous to previous slide} \end{aligned}$$

Construction of Input Factors: Step 2

For each $S \in \mathcal{E}$, construct an arbitrary ψ_S with $\psi_S \supseteq \psi'_S$ and $|\psi_S| = N^q$

This implies

$$\Phi \supseteq \times_{i\in[n]}[2^{p_i}]$$

Hence,

$$\begin{aligned}
|\Phi| &\geq \prod_{i\in[n]} 2^{p_i} \\
&= \prod_{i\in[n]} 2^{q\cdot v_i \cdot \log N} && p_i = q \cdot v_i \cdot \log N \\
&= (N^q)^{\sum_{i\in[n]} v_i} && \text{analogous to previous slide} \\
&= (N^q)^{D(\mathcal{H})} && (v_i)_{i\in[n]} \text{ is optimal solution to dual program}
\end{aligned}$$

Construction of Input Factors: Step 2

For each $S \in \mathcal{E}$, construct an arbitrary ψ_S with $\psi_S \supseteq \psi'_S$ and $|\psi_S| = N^q$

This implies

$$\Phi \supseteq \times_{i \in [n]} [2^{p_i}]$$

Hence,

$$\begin{aligned}
|\Phi| &\geq \prod_{i \in [n]} 2^{p_i} && \\
&= \prod_{i \in [n]} 2^{q \cdot v_i \cdot \log N} && p_i = q \cdot v_i \cdot \log N \\
&= (N^q)^{\sum_{i \in [n]} v_i} && \text{analogous to previous slide} \\
&= (N^q)^{D(\mathcal{H})} && (v_i)_{i \in [n]} \text{ is optimal solution to dual program} \\
&= (N^q)^{\rho^*(\mathcal{H})} && \text{linear program duality}
\end{aligned}$$

Lower Bound in Case of Input Factors with Different Sizes

Given a join $\Phi(\mathbf{x}) = \bigotimes_{S\in\mathcal{E}} \psi_S(\mathbf{x}_S)$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and input factor sizes N_S for $S \in \mathcal{E}$, the dual linear program extends to

$$
\begin{aligned}
\text{maximise} \quad & \textstyle\sum_{i\in[n]} v_i \\
\text{subject to} \quad & \textstyle\sum_{i\in S} v_i \leq \log N_S && \forall S \in \mathcal{E}, \\
& v_i \geq 0 && \forall i \in [n]
\end{aligned}
$$

Lower Bound in Case of Input Factors with Different Sizes

Given a join $\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and input factor sizes N_S for $S \in \mathcal{E}$, the dual linear program extends to

$$\begin{aligned} \text{maximise} \quad & \textstyle\sum_{i \in [n]} v_i \\ \text{subject to} \quad & \textstyle\sum_{i \in S} v_i \leq \log N_S & \forall S \in \mathcal{E}, \\ & v_i \geq 0 & \forall i \in [n] \end{aligned}$$

- Given an optimal solution $(v_i)_{i \in [n]}$ to the above program, we choose natural numbers q, $(p_i)_{i \in [n]}$ such that $v_i = \frac{p_i}{q}$

Lower Bound in Case of Input Factors with Different Sizes

Given a join $\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and input factor sizes N_S for $S \in \mathcal{E}$, the dual linear program extends to

$$\begin{aligned} \text{maximise} \quad & \textstyle\sum_{i \in [n]} v_i \\ \text{subject to} \quad & \textstyle\sum_{i \in S} v_i \leq \log N_S && \forall S \in \mathcal{E}, \\ & v_i \geq 0 && \forall i \in [n] \end{aligned}$$

- Given an optimal solution $(v_i)_{i \in [n]}$ to the above program, we choose natural numbers q, $(p_i)_{i \in [n]}$ such that $v_i = \frac{p_i}{q}$
- We construct input factors $\psi_S \supseteq \times_{i \in S}[2^{p_i}]$ of sizes N_S^q

Lower Bound in Case of Input Factors with Different Sizes

Given a join $\Phi(\mathbf{x}) = \bigotimes_{S \in \mathcal{E}} \psi_S(\mathbf{x}_S)$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and input factor sizes N_S for $S \in \mathcal{E}$, the dual linear program extends to

$$\begin{aligned} \text{maximise} \quad & \textstyle\sum_{i \in [n]} v_i & \\ \text{subject to} \quad & \textstyle\sum_{i \in S} v_i \leq \log N_S & \forall S \in \mathcal{E}, \\ & v_i \geq 0 & \forall i \in [n] \end{aligned}$$

- Given an optimal solution $(v_i)_{i \in [n]}$ to the above program, we choose natural numbers q, $(p_i)_{i \in [n]}$ such that $v_i = \frac{p_i}{q}$
- We construct input factors $\psi_S \supseteq \times_{i \in S} [2^{p_i}]$ of sizes N_S^q
- Let $(w_S)_{S \in \mathcal{E}}$ be an optimal solution to the linear program computing $\rho^*(\mathcal{H})$ with minimisation objective $\prod_{S \in \mathcal{E}} (N_S^q)^{w_S}$

Lower Bound in Case of Input Factors with Different Sizes

Given a join $\Phi(\mathbf{x}) = \bigotimes_{S\in\mathcal{E}} \psi_S(\mathbf{x}_S)$ with hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and input factor sizes N_S for $S \in \mathcal{E}$, the dual linear program extends to

$$
\begin{array}{lll}
\text{maximise} & \sum_{i\in[n]} v_i & \\
\text{subject to} & \sum_{i\in S} v_i \leq \log N_S & \forall S \in \mathcal{E}, \\
 & v_i \geq 0 & \forall i \in [n]
\end{array}
$$

- Given an optimal solution $(v_i)_{i\in[n]}$ to the above program, we choose natural numbers q, $(p_i)_{i\in[n]}$ such that $v_i = \frac{p_i}{q}$
- We construct input factors $\psi_S \supseteq \times_{i\in S}[2^{p_i}]$ of sizes N_S^q
- Let $(w_S)_{S\in\mathcal{E}}$ be an optimal solution to the linear program computing $\rho^*(\mathcal{H})$ with minimisation objective $\prod_{S\in\mathcal{E}}(N_S^q)^{w_s}$
- We can show $|\Phi| \geq \prod_{S\in\mathcal{E}}(N_S^q)^{w_s}$