Lecture 19

$$R^{f_i} \overset{A}{\frown} S^{g_{ij}} \overset{B}{\frown} T^{h_j}$$

Problem: $|R \bowtie_A S \bowtie_B T| = \sum_i \sum_j f_i g_{ij} h_j$

$X_R = \sum_i f_i \xi_i$

$X_S = \sum_{i,j} g_{ij} \xi_i \gamma_j$

$X_T = \sum_j h_j \gamma_j$

$X = X_R X_S X_T$ is unbiased

$$E[X^2] = \sum_{i_1}\sum_{i_2}\sum_{i_3}\sum_{i_4}\sum_{j_1}\sum_{j_2}\sum_{j_3}\sum_{j_4} f_{i_1} f_{i_2}\, g_{i_3 j_3}\, g_{i_4 j_4}\, h_{j_1} h_{j_2}$$

$$\underbrace{E[\xi_{i_1}\xi_{i_2}\xi_{i_3}\xi_{i_4}]}\cdot E[\gamma_{j_1}\gamma_{j_2}\gamma_{j_3}\gamma_{j_4}]$$

$\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3} - 2\,\delta_{i_1 i_2}\delta_{i_2 i_3}\delta_{i_3 i_4}$

$$= \sum_{i_1}\sum_{i_3}\sum_{j_1}\sum_{j_3} f_{i_1}^2\, g_{i_3 j_3}^2\, h_{j_1}^2 + \dots$$

$$= \sum_{i_1} f_{i_1}^2 \sum_i \sum_j g_{ij}^2 \sum_j h_j^2 - \dots - E^2[X]$$

$$\frac{SJ(R)\,SJ(S)}{SJ(M)} \le Var(X) \le 3^m\, SJ(R)\, SJ(S)\, SJ(M)$$

m – # of join constraints.

$R \bowtie_A S$

$f_i \qquad g_i$

$X_R = \sum_i f_i \xi_i$

$X_S = \sum_i g_i \xi_i$

$X = X_R X_S$

$Var(X) \simeq SJ(R)\, SJ(S)$

$\sum_i f_i^2 \sum_i g_i^2$

Suppose we can build $\tilde{X}$

$E[\tilde{X}] = \sum_i f_i g_i$

$Var(\tilde{X}) < Var(X)$

$I = I_1, I_2$

$I_1 \cap I_2 = \emptyset$

$I_1 \cup I_2 = I$

Decomposition on partitioning

$$\sum_{i \in I} f_i g_i = \sum_{i \in I_1} f_i g_i + \sum_{i \in I_2} f_i g_i$$

$$\begin{tikzcd}[row sep=small] & I_1 \\ I \arrow[ur] \arrow[dr] & \\ & I_2 \end{tikzcd}$$

$X_{R,1} = \sum_{i \in I_1} f_i \xi_i$

$X_{S,1} = \sum_{i \in I_1} g_i \xi_i$

$-\ 1,$

γ

$X_1 = X_{R,1} X_{S,1}$

$X_2 = X_{R,2} X_{S,2}$

2 times more memory

Claim $X = X_1 + X_2$ is unbiased.

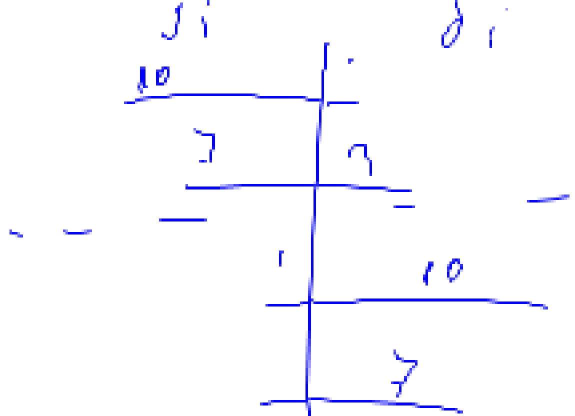
$E(\bar{X}) = E(X_1) + E(X_2) = \sum_{i \in I_1} f_i g_i + \sum_{i \in I_2} f_i g_i$

$= \sum_i f_i g_i$

$Var(\bar{X}) = Var(X_1) + Var(X_2)$

$\approx SJ(R_1)SJ(S_1) - SJ(R_2)SJ(S_2)$

$SJ(R)SJ(S)$



Original AMS	Sketch Partitioning
R: 100+49+1+1 = 151	R_1: 100+49 ≈ 150
S: 1+5+100+49 = 155	S_1: 1+9 = 10 → 1500
SJ(R)·SJ(S) ≈ 23500	R_2: 1+1 = 2
	S_2: 100+49 ≈ 150 → 300

$R \to R_1, R_2$

$S \to S_1, S_2$

$Var(X) \approx 1800$

Suppose partitioning of I is random

$SS(P_1) \approx \frac{1}{2} \ldots$

$SS(P_2)$

$SS(P_1) \approx \ldots SS$

$SS(R) = \sum_i r_i^2 = \underbrace{\sum_{i \in I_1} r_i^2}_{SS(R_1)} + \underbrace{\sum_{i \in I_2} r_i^2}_{SS(R_2)}$

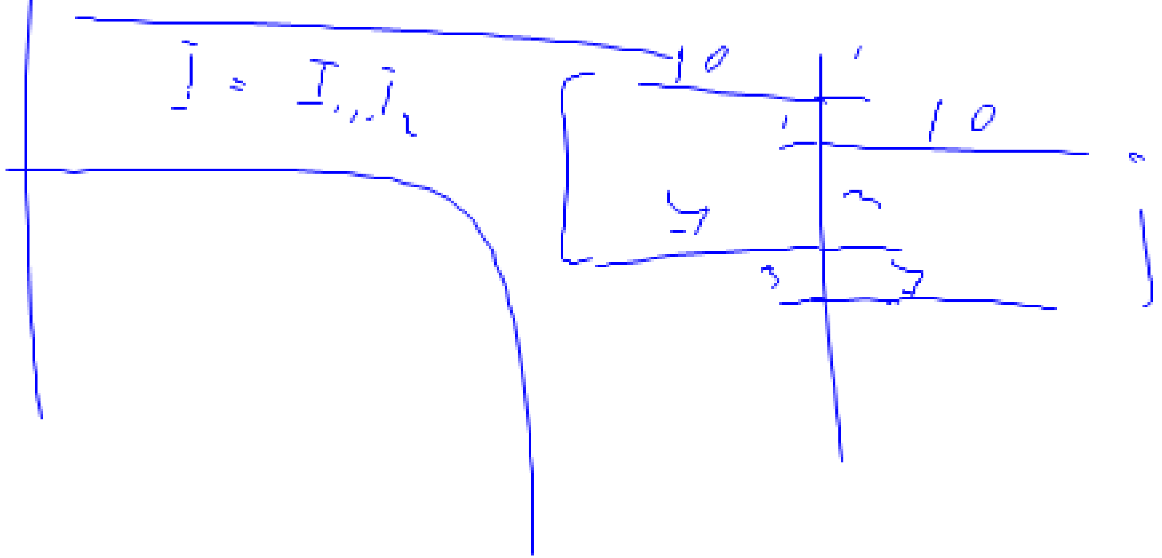
$\operatorname{Var}(\bar{X}) = \frac{1}{2} SS(R) \frac{1}{2} SS(S) + \frac{1}{2} SS(R) \frac{1}{2} SS(S)$

$= \frac{1}{2} SS(R) SS(S)$

If we expect [no] difference in size

$\operatorname{Var}(\bar{X}) \approx \operatorname{Var}(X)$

$I = I_1, I_2$



$\hat{X} = X_1 + X_2$

$m_1 \quad X_1$

$m_2 \quad X_2$

$m_1 \quad m_2 \quad \to X$

$\hat{X} = Avg_{m_1}(X_1) + Avg_{m_2}(X_2)$

$V_q(X_1) \quad V_q(X_2)$

$Var(\hat{X}) \sim \left(\sqrt{V_q(X_1)} - \sqrt{V_q(X_2)}\right)^2$

Find I_1, I_2 partition I s.t.

$\sqrt{\sum_{i \in I_1} f_i^2 \sum_{i \in I_1} s_i^2} + \sqrt{\sum_{i \in I_2} f_i^2 \sum_{i \in I_2} s_i^2}$ is minimized.

- minimize sol:

Consider all partitionings $2^{|I|}$ $|I|$

- $|I| \, \ell_2 \, |I|$ using Bernoulli

$|B| \, g(B)$