AMS Sampling, Estimating Frequency moments, F_2 Estimation

Lecture 07

September 15, 2020

Frequency Moments

- Stream consists of $e_1, e_2, \ldots, e_m$ where each e_i is an integer in $[n]$. We know n in advance (or an upper bound)
- Given a stream let f_i denote the frequency of i or number of times i is seen in the stream
- Consider vector $\mathbf{f} = (f_1, f_2, \ldots, f_n)$
- For $k \geq 0$ the k'th frequency moment $F_k = \sum_i f_i^k$. We can also consider the ℓ_k norm of $\mathbf{f}$ which is $(F_k)^{1/k}$.

Example: $n = 5$ and stream is $4, 2, 4, 1, 1, 1, 4, 5$

Problem: Estimate F_k from stream using small memory

A more general estimation problem

- Stream consists of $e_1, e_2, \ldots, e_m$ where each e_i is an integer in $[n]$. We know n in advance (or an upper bound)
- Given a stream let f_i denote the frequency of i or number of times i is seen in the stream
- Consider vector $\mathbf{f} = (f_1, f_2, \ldots, f_n)$
- Define a function $g(\sigma)$ of stream σ to be $\sum_{i=1}^{\overset{n}{\cancel{m}}} g_i(f_i)$ where $g_i : \mathbb{R} \rightarrow \mathbb{R}$ is a real-valued function such that $g_i(0) = 0$.

$$g(\sigma) = \sum_{i=1}^{5} f_i^2 + \sum_{i=6}^{n} f_i^3 \qquad F_k(\sigma) = \sum_{i=1}^{n} f_i^k$$

$$h(x) = x^k \qquad g(\sigma) = \sum_{\underline{\underline{i=1}}}^{n} h(f_i)$$

A more general estimation problem

- Stream consists of $e_1, e_2, \ldots, e_m$ where each e_i is an integer in $[n]$. We know n in advance (or an upper bound)
- Given a stream let f_i denote the frequency of i or number of times i is seen in the stream
- Consider vector $\mathbf{f} = (f_1, f_2, \ldots, f_n)$
- Define a function $g(\sigma)$ of stream σ to be $\sum_{i=1}^{m} g_i(f_i)$ where $g_i : \mathbb{R} \rightarrow \mathbb{R}$ is a real-valued function such that $g_i(0) = 0$.

Examples:

- Frequency moments F_k where for each i, $g_i(f_i) = h(f_i)$ where $h(x) = x^k$
- Entropy of stream: $g(\sigma) = \sum_i f_i \log(f_i)$ (assume $0 \log 0 = 0$)

Part I

AMS Sampling

AMS Sampling

An unbiased statistical estimator for $g(\sigma)$

- Sample e_J uniformly at random from stream of length m
- Suppose $e_J = i$ where $i \in [n]$
- Let $R = |\{j \mid J \leq j \leq m, e_j = e_J = i\}|$
- Output $(g_i(R) - g_i(R-1)) \cdot m$

$n = [10]$

$m = 13$

1, 3, 1, 10, 5, 1, 10, 5, 5, 2, 1, 3, 3.

$J = 1$ $\quad J = 4 \quad J = 7$

$13 \cdot (g_{10}(2) - g_{10}(1))$.

$13 \cdot (g_{10}(1) - g_{10}(0))$

$13 \cdot (g_1(4) - g_1(3))$.

AMS Sampling

An unbiased statistical estimator for $g(\sigma)$

- Sample e_J uniformly at random from stream of length m
- Suppose $e_J = i$ where $i \in [n]$
- Let $R = |\{j \mid J \leq j \leq m, e_j = e_J = i\}|$
- Output $(g_i(R) - g_i(R-1)) \cdot m$

Can be implemented in streaming setting with reservoir sampling.

Streaming Implementation

```
AMSEstimate:
    $s \leftarrow$ null
    $m \leftarrow 0$
    $R \leftarrow 0$
    While (stream is not done)
        $m \leftarrow m + 1$
        $a_m$ is current item
        Toss a biased coin that is heads with probability $1/m$
        If (coin turns up heads)
            $s \leftarrow a_m$
            $R \leftarrow 1$
        Else If ($a_m == s$)
            $R \leftarrow R + 1$
    endWhile
    Output $(g_s(R) - g_s(R-1)) \cdot m$
```

Expectation of output

Let Y be the output of the algorithm.

Lemma

$E[Y] = g(\sigma) = \sum_{i \in [n]} g_i(f_i)$.

$$E[Y] = \sum_{i=1}^{n} \Pr[e_J = i] \cdot E[Y \mid e_J = i]$$

$i \quad i \quad i$

$f_i = ?$

$$= \sum_{i=1}^{n} \frac{f_i}{m} \cdot E[Y \mid e_J = i]$$

$$= \sum_{i=1}^{n} \cancel{\frac{f_i}{m}} \cancel{m} \sum_{l=1}^{f_i} \cancel{\frac{1}{f_i}} \left(g_i(l) - g_i(l-1)\right)$$

$$= \sum_{i=1}^{n} \left(g_i(f_i) - g_i(0)\right) = \underline{\underline{\sum_{i=1}^{n} g_i(f_i)}}$$

$$= g_i(\sigma).$$

Expectation of output

Let Y be the output of the algorithm.

Lemma

$E[Y] = g(\sigma) = \sum_{i \in [n]} g_i(f_i)$.

$\mathbf{Pr}[e_J = i] = f_i/m$ since e_J is chosen uniformly from stream.

Expectation of output

Let Y be the output of the algorithm.

Lemma

$E[Y] = g(\sigma) = \sum_{i \in [n]} g_i(f_i)$.

$\Pr[e_J = i] = f_i/m$ since e_J is chosen uniformly from stream.

$$\begin{aligned}
\mathbf{E}[Y] &= \sum_{i \in [n]} \Pr[a_J = i]\, \mathbf{E}[Y | a_J = i] \\
&= \sum_{i \in [n]} \frac{f_i}{m} \mathbf{E}[Y | a_J = i] \\
&= \sum_{i \in [n]} \frac{f_i}{m} \sum_{\ell=1}^{f_i} m \frac{1}{f_i} \left(g_i(\ell) - g_i(\ell - 1)\right) \\
&= \sum g_i(f_i).
\end{aligned}$$

Application to estimating frequency moments

Suppose $g(\sigma) = F_k$ for some $k > 1$. That is $g_i(x) = x^k$ for each i. What is $Var(Y)$?

Application to estimating frequency moments

Suppose $g(\sigma) = F_k$ for some $k > 1$. That is $g_i(x) = x^k$ for each i. What is $Var(Y)$?

Lemma

When $g(x) = x^k$ and $k \geq 1$, $Var[Y] \leq kF_1F_{2k-1} \leq kn^{1-\frac{1}{k}}F_k^2$.

$$E[Y] = \underline{\underline{F_k}}$$

$$Var[Y] \leq \underbrace{k\, n^{1-\frac{1}{k}}} \cdot \overline{\overline{\underline{\underline{F_k}}^2}}$$

$$\underline{\underline{\sqrt{n}}}\ F_2^{\,2}$$

Application to estimating frequency moments

Suppose $g(\sigma) = F_k$ for some $k > 1$. That is $g_i(x) = x^k$ for each i. What is $Var(Y)$?

Lemma

When $g(x) = x^k$ and $k \geq 1$, $Var[Y] \leq kF_1F_{2k-1} \leq kn^{1-\frac{1}{k}}F_k^2$.

$\mathbf{E}[Y] = F_k$ and $Var(Y) \leq \boxed{kn^{1-\frac{1}{k}}}F_k^2$. Hence, if we want to use averaging and Cheybyshev we need to average $h = \Omega(\frac{1}{\epsilon^2}kn^{1-\frac{1}{k}})$ parallel runs and space to get a $(1 \pm \epsilon)$ estimate to F_k with constant probability.

ϵF_k $\quad$ $\frac{1}{\epsilon^2}kn^{1-\frac{1}{k}}$ $\quad$ $\epsilon^2 F_k^2$

Variance calculation

$$g_i(x) = \underline{\underline{x}}^k$$

$$
\begin{aligned}
Var[Y] &\leq \underline{\underline{\mathbf{E}\left[Y^2\right]}} \\
&\leq \underline{\sum_{i \in [n]} \Pr[\underline{\underline{a_J = i}}] \sum_{\ell=1}^{f_i} \frac{m^2}{f_i} \left(\ell^k - (\ell-1)^k\right)^2} \\
&\leq \sum_{i \in [n]} \frac{f_i}{m} \sum_{\ell=1}^{f_i} \frac{m^2}{f_i} (\underline{\underline{\ell^k - (\ell-1)^k}})(\ell^k - (\ell-1)^k) \\
&\leq \underline{\underline{m}} \sum_{i \in [n]} \sum_{\ell=1}^{f_i} \underline{\underline{k\ell^{k-1}}}(\ell^k - (\ell-1)^k) \quad \text{using } x^k - (x-1)^k \leq kx^{k-1} \\
&\leq km \sum_{i \in [n]} f_i^{k-1} f_i^k \\
&\leq kmF_{2k-1} = kF_1 F_{2k-1}.
\end{aligned}
$$

$$\boxed{km \sum_{i=1}^{n} f_i^{2k-1}}$$

$$\left(\sum f_i^k\right)^2$$

Variance calculation

Claim: For $k \geq 1$, $\underline{F_1 F_{2k-1}} \leq \underline{n^{1-1/k}(F_k)^2}$.

m^k $\quad$ m^{2k}

$$\boxed{F_k = \sum_{i=1}^{n} f_i^k.} \qquad \boxed{km \sum_{i=1}^{n} f_i^{2k-1}.}$$

$\sum f_i = m$

- ✓ $f_i = m$ $\quad$ $f_i = 0$ $\quad$ $i > 2$
- $\boxed{f_i = \frac{m}{n}}$ $\forall i$

$$F_k = n \cdot \left(\frac{m}{n}\right)^k$$

$$km \cdot n \cdot \left(\frac{m}{n}\right)^{2k-1} = n^{1-\frac{1}{k}} \cdot F_k^2$$

Variance calculation

Claim: For $k \geq 1$, $F_1 F_{2k-1} \leq \underline{\underline{n^{1-1/k}}}(F_k)^2$.

The function $g(x) = x^k$ is convex for $k \geq 1$.
Implies $\sum_i x_i/n \leq ((\sum_i x_i^k)/n)^{1/k}$.

$$\begin{aligned}
F_1 F_{2k-1} &= (\sum_i f_i)(\sum_i f_i^{2k-1}) \leq (\sum_i f_i)(F_\infty)^{k-1}(\sum_i f_i^k) \\
&\leq (\sum_i f_i)(\sum_i f_i^k)^{\frac{k-1}{k}}(\sum_i f_i^k) \\
&\leq n^{1-1/k}(\sum_i f_i^k)^{1/k}(\sum_i f_i^k)^{\frac{k-1}{k}}(\sum_i f_i^k) \\
&= n^{1-1/k}(F_k)^2
\end{aligned}$$

Worst case is when $f_i = m/n$ for each $i \in [n]$.

Frequency moment estimation

AMS-Estimator shows that F_k can be estimated in $O(n^{1-1/k})$ space.

Question: Can one do better?

Frequency moment estimation

AMS-Estimator shows that F_k can be estimated in $\underline{\underline{O(n^{1-1/k})}}$ space.

Question: Can one do better?

- For F_2 and $1 \leq k \leq 2$ one can do $O(polylog(n))$ space!
- For $k > 2$ space complexity is $\tilde{O}(n^{1-\underline{\underline{2}}/k})$ which is known to be essentially tight.

polylog

Thus a phase transition at $\underline{\underline{k = 2}}$.

$\tilde{O}(\ ) - \underline{\underline{\log(\ )}}$

Part II

F_2 Estimation

Estimating F_2

- Stream consists of $e_1, e_2, \ldots, e_m$ where each e_i is an integer in $[n]$. We know n in advance (or an upper bound)
- Given a stream let f_i denote the frequency of i or number of times i is seen in the stream
- Consider vector $\mathbf{f} = (f_1, f_2, \ldots, f_n)$

Question: Estimate $F_2 = \sum_{i=1}^{m} f_i^2$ in small space.

Using generic AMS sampling scheme we can do this in $O(\underline{\underline{\sqrt{n}\log}}\, n)$ space. Can we do it better?

AMS Scheme for F_2

```
AMS-$F_2$-Estimate:
    Let $h : [n] \to \{-1, 1\}$ be chosen from
        a $4$-wise independent hash family $\mathcal{H}$.
    $z \leftarrow 0$
    While (stream is not empty) do
        $a_j$ is current item
        $z \leftarrow z + h(a_j)$
    endWhile
    Output $z^2$
```

$h(1) = -1$

$h(2) = 1$

$h(3) = -1$

$h(4)$.

$h(5) = +1$

$h(10) = -1$

2, 5, 1, 10, 3, 1, 1, 2, 5, 5, 5

$z=0$ 1 2 1 0 -1 -2 -3 -2 -1 0 1

$\boxed{1^2 = 1}$.

$F_2 = 3^2 + 2^2 + 1^2 + 4^2 + 1^2 = 31$

AMS Scheme for F_2

```
**AMS-$F_2$-Estimate**:
    Let $h : [n] \to \{-1, 1\}$ be chosen from
        a $4$-wise independent hash family $\mathcal{H}$.
    $z \leftarrow 0$
    While (stream is not empty) do
        $a_j$ is current item
        $z \leftarrow z + h(a_j)$
    endWhile
    Output $z^2$
```

```
**AMS-$F_2$-Estimate**:
    Let $Y_1, Y_2, \ldots, Y_n$ be $\{-1, +1\}$ random variable that are
        $4$-wise independent
    $z \leftarrow 0$
    While (stream is not empty) do
        $a_j$ is current item
        $z \leftarrow z + Y_{a_j}$
    endWhile
    Output $z^2$
```

Analysis

$Z = \sum_{i=1}^{n} f_i Y_i$ and output is Z^2

$$E[Z^2] = E\left[\left(\sum_{i=1}^{n} f_i Y_i\right)^2\right] = E\left[\sum_{i=1}^{n} f_i^2 + 2\sum_{1 \le i < j \le n} f_i f_j Y_i Y_j\right]$$

$$= \underbrace{\sum_{i=1}^{n} f_i^2}_{F_2} + 2\sum_{1 \le i < j \le n} f_i f_j \underbrace{E[Y_i Y_j]}_{= E[Y_i]E[Y_j] = 0}$$

$$= \cancel{2}\, \boxed{F_2}$$

Analysis

$Z = \sum_{i=1}^{n} f_i Y_i$ and output is Z^2

- $\mathbf{E}[Y_i] = 0$ and $Var(Y_i) = \mathbf{E}\left[Y_i^2\right] = 1$
- For $i \neq j$, since Y_i and Y_j are pairwise-independent $\mathbf{E}[Y_i Y_j] = 0$.

Analysis

$Z = \sum_{i=1}^{n} f_i Y_i$ and output is Z^2

- $\mathbf{E}[Y_i] = 0$ and $Var(Y_i) = \mathbf{E}\left[Y_i^2\right] = 1$
- For $i \neq j$, since Y_i and Y_j are pairwise-independent $\mathbf{E}[Y_i Y_j] = 0$.

$$Z^2 = \sum_i f_i^2 Y_i^2 + 2 \sum_{i \neq j} f_i f_j Y_i Y_j$$

and hence

$$\mathbf{E}\left[Z^2\right] = \sum_i f_i^2 = F_2.$$

Variance

What is $Var(\underset{=}{Z}^2)$?

Variance

What is $Var(Z^2)$?

$$E[Z^4] = \sum_{i\in[n]} \sum_{j\in[n]} \sum_{k\in[n]} \sum_{\ell\in[n]} f_i f_j f_k f_\ell E[\underline{\underline{Y_i Y_j Y_k Y_\ell}}].$$

Variance

What is $Var(Z^2)$?

$Y_1^2 \; Y_2^2$

$Y_1 \; Y_2 \; Y_3 \; Y_4 \qquad \underline{Y_1} \, \underline{Y_3}^2 Y_4$

$$E[Z^4] = \sum_{i\in[n]} \sum_{j\in[n]} \sum_{k\in[n]} \sum_{\ell\in[n]} f_i f_j f_k f_\ell E[\underline{\underline{Y_i Y_j Y_k Y_\ell}}].$$

4-wise independence implies $\mathbf{E}[Y_i Y_j Y_k Y_\ell] = 0$ if there is a number among i, j, k, ℓ that occurs only once. Otherwise 1.

$i=j=k=l \qquad Y_i^4 = 1$

$i=j \quad k=l \qquad Y_i^2 \, Y_j^2 = \textcircled{1}$

Variance

What is $Var(Z^2)$?

$$E[Z^4] = \sum_{i \in [n]} \sum_{j \in [n]} \sum_{k \in [n]} \sum_{\ell \in [n]} f_i f_j f_k f_\ell E[Y_i Y_j Y_k Y_\ell].$$

4-wise independence implies $\mathbf{E}[Y_i Y_j Y_k Y_\ell] = 0$ if there is a number among i, j, k, ℓ that occurs only once. Otherwise 1.

$$\begin{aligned} \underline{\underline{E[Z^4]}} &= \sum_{i \in [n]} \sum_{j \in [n]} \sum_{k \in [n]} \sum_{\ell \in [n]} f_i f_j f_k f_\ell E[Y_i Y_j Y_k Y_\ell] \\ F_4 \rightarrow\ &= \boxed{\sum_{i \in [n]} f_i^4} + 6 \sum_{i=1}^{n} \sum_{\underline{\underline{j=i+1}}}^{n} f_i^2 f_j^2 . \end{aligned}$$

Variance

$$
\begin{aligned}
Var(Z^2) &= \underline{\mathbf{E}}\left[Z^4\right] - (\underline{\mathbf{E}}\left[\overset{\downarrow}{Z^2}\right])^2 \\
&= \underline{F_4} - \underline{F_2^2} + \underline{6}\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_i^2 f_j^2 \\
&= \underline{F_4} - (F_4 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_i^2 f_j^2) + 6\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_i^2 f_j^2 \\
&= 4\sum_{i=1}^{n}\sum_{j=i+1}^{n} f_i^2 f_j^2 \\
&\leq \underline{2F_2^2}.
\end{aligned}
$$

$$E[Z^2] = \boxed{F_2}$$

$$Var[Z^2] \leq \underline{2}\,\boxed{F_2^{\,2}}$$

Averaging and median trick again

Output is Z^2: and $\mathbf{E}\left[Z^2\right] = F_2$ and $Var(Z^4) \leq 2F_2^2$

- Reduce variance by averaging $8/\epsilon^2$ independent estimates. Let Y be the averaged estimator.
- Apply Chebyshev to average estimator. $\Pr[|Y - F_2| \geq \epsilon F_2] \leq 1/4$.
- Reduce error probability to δ by independently doing $O(\log(1/\delta))$ estimators above.
- Total space $\underline{\underline{O(\log(1/\delta)\frac{1}{\epsilon^2}\log n)}}$

Geometric Interpretation

Observation: The estimation algorithm works even when f_i's can be negative. What does this mean?

Geometric Interpretation

Observation: The estimation algorithm works even when f_i's can be negative. What does this mean?

$F_1 = \sum \underline{\underline{f_i}}$ $\qquad F_2 = \sum_{i=1}^{n} f_i^2 = \sum_{i=1}^{n} x_i^2$

Richer model:

- Want to estimate a function of a vector $x \in \mathbb{R}^n$ which is initially assume to be the all $\mathbf{0}$'s vector. (previously we were thinking of the frequency vector f)
- Each element e_j of a stream is a tuple (i_j, Δ_j) where $i_j \in [n]$ and $\Delta_i \in \mathbb{R}$ is a real-value: this updates x_{i_j} to $x_{i_j} + \Delta_j$. (Δ_j can be positive or negative)

$$\left(0, \underset{\uparrow}{\overset{+1}{0}}, 0, \underset{\uparrow}{\overset{+10\ -5}{0}}, \dots, 0\right).$$

$$(i, \Delta)$$

$$(i, \underline{\underline{10}})$$

$$(i, -5)$$

Algorithm revisited

```
AMS-$\ell_2$-Estimate:
    Let $Y_1, Y_2, \ldots, Y_n$ be $\{-1, +1\}$ random variable that are
        $4$-wise independent
    $z \leftarrow 0$.
    While (stream is not empty) do
        $a_j = (i_j, \Delta_j)$ is current update
        $z \leftarrow z + \underline{\Delta_j} Y_{i_j}$
    endWhile
    Output $z^2$
```

$(2, 5)$, $(10, 1.9)$, $(1, -5.2)$, $(3, 100)$, $(1, -6)$

$\rightarrow \boxed{5.3}$

$\bar{x} = (x_1, x_2, \ldots, x_n)$

$\bar{f} = (f_1, f_2, \ldots, f_n)$

$\underline{\sum x_i^2}$

Algorithm revisited

```
AMS-$\ell_2$-Estimate:
    Let $Y_1, Y_2, \ldots, Y_n$ be $\{-1, +1\}$ random variable that are
        $4$-wise independent
    $z \leftarrow 0$
    While (stream is not empty) do
        $a_j = (i_j, \Delta_j)$ is current update
        $z \leftarrow z + \Delta_j Y_{i_j}$
    endWhile
    Output $z^2$
```

Claim: Output estimates $\underline{\underline{||x||_2^2}}$ where x is the vector at end of stream of updates.

Analysis

$Z = \sum_{i=1}^{n} x_i Y_i$ and output is Z^2

Analysis

$Z = \sum_{i=1}^{n} x_i Y_i$ and output is Z^2

- $\mathbf{E}[Y_i] = 0$ and $Var(Y_i) = \mathbf{E}\left[Y_i^2\right] = 1$
- For $i \neq j$, since Y_i and Y_j are pairwise-independent $\mathbf{E}[Y_i Y_j] = 0$.

$$Z^2 = \sum_i x_i^2 Y_i^2 + 2 \sum_{i \neq j} x_i x_j Y_i Y_j$$

and hence

$$\mathbf{E}\left[Z^2\right] = \sum_i x_i^2 = ||x||_2^2.$$

Analysis

$Z = \sum_{i=1}^{n} x_i Y_i$ and output is Z^2

- $\mathbf{E}[Y_i] = 0$ and $Var(Y_i) = \mathbf{E}\left[Y_i^2\right] = 1$
- For $i \neq j$, since Y_i and Y_j are pairwise-independent $\mathbf{E}[Y_i Y_j] = 0$.

$$Z^2 = \sum_i x_i^2 Y_i^2 + 2 \sum_{i \neq j} x_i x_j Y_i Y_j$$

and hence

$$\mathbf{E}\left[Z^2\right] = \sum_i x_i^2 = ||x||_2^2.$$

And as before one can show that $Var(Z^2) \leq 2(\mathbf{E}\left[Z^2\right])^2$.

Introduction to (Linear) Sketching

A *sketch* of a stream σ is a summary data structure $C(\sigma)$ (ideally of small space) such that the sketch of the composition $\sigma_1 \cdot \sigma_2$ of two streams σ_1 and σ_1 can be computed from $C(\sigma_1)$ and $C(\sigma_2)$. The output of the algorithm is some function of the sketch.

Introduction to (Linear) Sketching

A *sketch* of a stream σ is a summary data structure $C(\sigma)$ (ideally of small space) such that the sketch of the composition $\sigma_1 \cdot \sigma_2$ of two streams σ_1 and σ_1 can be computed from $C(\sigma_1)$ and $C(\sigma_2)$. The output of the algorithm is some function of the sketch.

What is the summary of algorithm for F_2 estimation? Is it a sketch?

h σ_1 h σ_2 h $\sigma_1 \cdot \sigma_2$

$\underline{\underline{Z_1}}$ $\underline{\underline{Z_2}}$ $\underline{\underline{Z = Z_1 + Z_2}}$

Z_1^2 Z_2^2 $\boxed{Z^2}$

Introduction to (Linear) Sketching

A *sketch* of a stream σ is a summary data structure $C(\sigma)$ (ideally of small space) such that the sketch of the composition $\sigma_1 \cdot \sigma_2$ of two streams σ_1 and σ_1 can be computed from $C(\sigma_1)$ and $C(\sigma_2)$. The output of the algorithm is some function of the sketch.

What is the summary of algorithm for F_2 estimation? Is it a sketch?

A sketch is a *linear* sketch if $C(\sigma_1 \cdot \sigma_2) = \underline{\underline{C(\sigma_1)}} + \underline{\underline{C(\sigma_2)}}$.

Introduction to (Linear) Sketching

A *sketch* of a stream σ is a summary data structure $C(\sigma)$ (ideally of small space) such that the sketch of the composition $\sigma_1 \cdot \sigma_2$ of two streams σ_1 and σ_1 can be computed from $C(\sigma_1)$ and $C(\sigma_2)$. The output of the algorithm is some function of the sketch.

What is the summary of algorithm for F_2 estimation? Is it a sketch?

A sketch is a *linear* sketch if $C(\sigma_1 \cdot \sigma_2) = C(\sigma_1) + C(\sigma_2)$.

Is the sketch for F_2 estimation a linear sketch?

$$\begin{array}{c|ccccc} & Y_1 & Y_2 & \cdots & Y_n & \\ h_1 & -1 & +1 & +1 & -1 & \cdots \\ h_2 & & & & & \\ \vdots & & & & & \\ h_{10} & & & & & \end{array} \begin{bmatrix} - \\ - \\ - \\ - \\ - \\ - \end{bmatrix} \bar{x} = \bar{z}$$

$$\underline{\underline{M}}\,\bar{x} = \bar{z}$$

F_2 Estimation as Linear Sketching

Recall that we take average of independent estimators and take median to reduce error. Can we view all this as a sketch?

```
AMS-$\ell_2$-Sketch:
    $\ell = c\log(1/\delta)/\epsilon^2$
    Let $M$ be a $\ell \times n$ matrix with entries in $\{-1, 1\}$ s.t
        (i) rows are independent and
        (ii) in each row entries are $4$-wise independent
    $z$ is a $\ell \times 1$ vector initialized to $0$
    While (stream is not empty) do
        $a_j = (i_j, \Delta_j)$ is current update
        $z \leftarrow z + \Delta_j M e_{i_j}$
    endWhile
    Output vector $z$ as sketch.
```

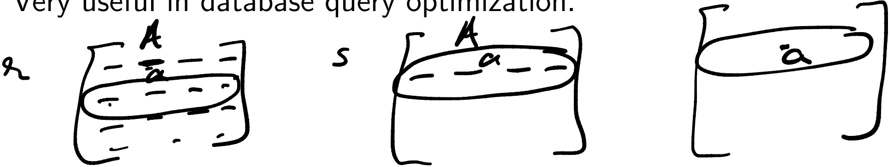
M is compactly represented via ℓ hash functions, one per row, independently chosen from 4-wise independent hash familty.

An Application to Join Size Estimation

In Databases an important operation is the “join” operation

- A relation/table r of arity k consists of tuples of size k where each tuple element is from some given type. Example: (netid, uin, last name, first name, dob, address) in a student data base
- Given two relations r and s and a common attribute a one often needs to compute their join $r \bowtie s$ over some common attribute that they share
- $r \bowtie s$ can have size quadratic in size of r and s

Question: Estimate size of $r \bowtie s$ without computing it explicitly. Very useful in database query optimization.



An Application to Join Size Estimation

In Databases an important operation is the “join” operation

- A relation/table r of arity k consists of tuples of size k where each tuple element is from some given type. Example: (netid, uin, last name, first name, dob, address) in a student data base
- Given two relations r and s and a common attribute a one often needs to compute their join $r \bowtie s$ over some common attribute that they share
- $r \bowtie s$ can have size quadratic in size of r and s

Question: Estimate size of $r \bowtie s$ without computing it explicitly. Very useful in database query optimization.
Estimating $r \bowtie r$ over an attribute a is same as F_2 estimation. Why?

Sketching: a shift in perspective

- Sketching ideas have many powerful applications in theory and practice
- In particular linear sketches are powerful. Allows one to handle negative entries and deletions. Surprisingly linear sketches are feasible in several settings.
- Connected to dimension reduction (JL Lemma), subspace embeddings and other important topics