AMS Sampling, Estimating Frequency moments, $F_2$ Estimation

Lecture 07
September 15, 2020
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 $f = (f_1, f_2, \ldots, f_n)$.

For $k \geq 0$ the $k$'th frequency moment $F_k = \sum_i f_i^k$. We can also consider the $\ell_k$ norm of $f$ 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 $f = (f_1, f_2, \ldots, f_n)$.
- Define a function $g(\sigma)$ of stream $\sigma$ to be $\sum_{i=1}^{n} g_i(f_i)$ where $g_i : \mathbb{R} \rightarrow \mathbb{R}$ is a real-valued function such that $g_i(0) = 0$.

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

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

$$F_k(\sigma) = \sum_{i=1}^{n} f_i^k$$
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 $f = (f_1, f_2, \ldots, f_n)$.
- Define a function $g(\sigma)$ of stream $\sigma$ 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 \( (g_i(R) - g_i(R - 1)) \cdot m \)

\[ 1, 3, 1, 10, 5, 1, 10, 5, 5, 2, 1, 3, 3. \]

\[ J = 1 \]

\[ \uparrow \]

\[ J = 4 \]

\[ \uparrow \]

\[ J = 7 \]

\[ 12 \cdot (g_{10}(2) - g_{10}(1)) \]

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

\[ 13 \cdot (g_i(u) - g_i(13)) \]
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 $R = |\{j | J \leq j \leq m, e_j = e_J = i\}|$
- Output $(g_i(R) - g_i(R - 1)) \cdot m$

Can be implemented in streaming setting with reservoir sampling.
AMSEstimate:

\[ s \leftarrow \text{null} \]
\[ m \leftarrow 0 \]
\[ R \leftarrow 0 \]

While (stream is not done)

\[ m \leftarrow m + 1 \]
\[ a_m \text{ is current item} \]
\[ \text{Toss a biased coin that is heads with probability } \frac{1}{m} \]
\[ \text{If (coin turns up heads)} \]
\[ s \leftarrow a_m \]
\[ R \leftarrow 1 \]
\[ \text{Else If } (a_m == s) \]
\[ R \leftarrow R + 1 \]

endWhile

Output \((g_s(R) - g_s(R - 1)) \cdot m\)
Expected output

Let $Y$ be the output of the algorithm.

**Lemma**

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

$$E\left[Y\right] = \sum_{i=1}^{n} \Pr\left[e_j = i\right] \cdot E\left[Y | e_j = i\right]$$

$$= \sum_{i=1}^{n} \frac{f_i}{m} \cdot E\left[Y | e_j = i\right]$$

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

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

$$= g_i(\sigma).$$
Expectation of output

Let $Y$ be the output of the algorithm.

**Lemma**

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

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

Let \( Y \) be the output of the algorithm.

**Lemma**

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

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

\[
E[Y] = \sum_{i \in [n]} \Pr[a_J = i] E[Y|a_J = i]
\]

\[
= \sum_{i \in [n]} \frac{f_i}{m} E[Y|a_J = i]
\]

\[
= \sum_{i \in [n]} \frac{f_i}{m} \sum_{\ell=1}^{f_i} m \frac{1}{f_i} (g_i(\ell) - g_i(\ell - 1))
\]

\[
= \sum g_i(f_i).
\]
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 $\text{Var}(Y)$?
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 \( \text{Var}(Y) \)?

**Lemma**

When \( g(x) = x^k \) and \( k \geq 1 \), \( \text{Var}[Y] \leq k F_1 F_{2k-1} \leq kn^{1 - \frac{1}{k}} F_k^2 \).

\[
\mathbb{E}[Y] = F_k = F_k
\]

\[
\text{Var}[Y] \leq kn^{1 - \frac{1}{k}} F_k^2 = \sqrt{n} F_k^2
\]
Suppose $g(\sigma) = F_k$ for some $k > 1$. That is $g_i(x) = x^k$ for each $i$. What is $\text{Var}(Y)$?

**Lemma**

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

$E[Y] = F_k$ and $\text{Var}(Y) \leq kn^{1-\frac{1}{k}} F_k^2$. Hence, if we want to use averaging and Chebyshev 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.
Variance calculation

\[ \text{Var}[Y] \leq \mathbb{E}[Y^2] \]

\[ \leq \sum_{i \in [n]} \text{Pr}[a_j = i] \sum_{\ell=1}^{f_i} \frac{m^2}{f_i} \left( \ell^k - (\ell - 1)^k \right)^2 \]

\[ \leq \sum_{i \in [n]} \frac{f_i}{m} \sum_{\ell=1}^{f_i} \frac{m^2}{f_i} \left( \ell^k - (\ell - 1)^k \right) \left( \ell^k - (\ell - 1)^k \right) \]

\[ \leq m \sum_{i \in [n]} \sum_{\ell=1}^{f_i} k \ell^{k-1} \left( \ell^k - (\ell - 1)^k \right) \]

\[ \leq km \sum_{i \in [n]} f_i^{k-1} f_i^k \]

\[ \leq km F_{2k-1} = kF_1 F_{2k-1}. \]
Claim: For $k \geq 1$, $F_1 F_{2k-1} \leq n^{1-1/k}(F_k)^2$.

\[
F_k = \sum_{i=1}^{n} f_i^k
\]

$\sum f_i = m$

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

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

\[
k m \geq f_i^{2k-1}
\]

$\left\{ \begin{array}{c}
  f_i = m \\
  f_i = 0 & i > 2
\end{array} \right.$

\[
\forall i
\]

\[
k m \cdot n \cdot \left(\frac{m}{n}\right)^{2k-1} = n^{1-\frac{1}{k}} F_k^2
\]
Claim: For $k \geq 1$, $F_1 F_{2k-1} \leq n^{1-1/k} (F_k)^2$.

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

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

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

Chandra (UIUC)
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 $O(n^{1-1/k})$ space.

**Question:** Can one do better?

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

Thus a phase transition at $k = 2$. 

$$O(1) \quad \approx \quad \text{polylog}$$
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 $f = (f_1, f_2, \ldots, f_n)$.

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

Using generic AMS sampling scheme we can do this in $O(\sqrt{n \log n})$ space. Can we do it better?
AMS Scheme for $F_2$

AMS-$F_2$-Estimate:

Let $h : [n] \rightarrow \{-1, 1\}$ 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 $z^2$

\[ z = 0 \quad 1 \quad 2 \quad 1 \quad 0 \quad -1 \quad -2 \quad -3 \quad -2 \quad -1 \quad 0 \quad 1 \]

$F_2 = 3^2 + 2^2 + 1^2 + 4^2 + 1^2 = 31$
AMS Scheme for $F_2$

**AMS-$F_2$-Estimate:**

Let $h: [n] \rightarrow \{-1, 1\}$ be chosen from a 4-wise independent hash family $\mathcal{H}$.

1. $z \leftarrow 0$
2. While (stream is not empty) do
   - $a_j$ is current item
   - $z \leftarrow z + h(a_j)$
3. endWhile
4. Output $z^2$

**AMS-$F_2$-Estimate:**

Let $Y_1, Y_2, \ldots, Y_n$ be $\{-1, +1\}$ random variables that are 4-wise independent.

1. $z \leftarrow 0$
2. While (stream is not empty) do
   - $a_j$ is current item
   - $z \leftarrow z + Y_{a_j}$
3. endWhile
4. Output $z^2$
Analysis

\[ Z = \sum_{i=1}^{n} f_i Y_i \] and output is \( Z^2 \)

\[
\mathbb{E}[Z^2] = \mathbb{E} \left[ \left( \sum_{i=1}^{n} f_i Y_i \right)^2 \right] = \mathbb{E} \left[ \sum_{i=1}^{n} f_i^2 + 2 \sum_{i=1}^{n} f_i f_j \mathbb{E}[Y_i Y_j] \right] \\
= \sum_{i=1}^{n} f_i^2 + 2 \sum_{i=1}^{n} f_i f_j \mathbb{E}[Y_i Y_j] \\
= \sum_{i=1}^{n} f_i^2 + 2 \sum_{i=1}^{n} f_i f_j \mathbb{E}[Y_i Y_j]
\]
Analysis

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

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

\[ Z = \sum_{i=1}^{n} f_i Y_i \] and output is \( Z^2 \)

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

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

and hence

\[ E[Z^2] = \sum_{i} f_i^2 = F_2. \]
Variance

What is $\text{Var}(Z^2)$?
Variance

What is \( \text{Var}(Z^2) \)?

\[
E[Z^4] = \sum_{i \in [n]} \sum_{j \in [n]} \sum_{k \in [n]} \sum_{\ell \in [n]} f_if_jf_kf_\ell E[Y_iY_jY_kY_\ell].
\]
Variance

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

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

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

$$\begin{align*}
i &= j = k = \ell & Y_i^4 &= 1 \\
i = j & k = 1 & Y_i^2 Y_j^2 &= 0
\end{align*}$$
Variance

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

$$E[Z^4] = \sum \sum \sum \sum f_i f_j f_k f_\ell E[Y_i Y_j Y_k Y_\ell].$$

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

$$E[Z^4] = \sum \sum \sum \sum f_i f_j f_k f_\ell E[Y_i Y_j Y_k Y_\ell]$$

$$= \sum f_i^4 + 6 \sum_{i=1}^n \sum_{j=i+1}^n f_i^2 f_j^2.$$
\[
\text{Var}(Z^2) = \mathbb{E}[Z^4] - (\mathbb{E}[Z^2])^2
\]
\[
= F_4 - F_2^2 + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_i^2 f_j^2
\]
\[
= F_4 - (F_4 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_i^2 f_j^2) + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_i^2 f_j^2
\]
\[
= 4 \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_i^2 f_j^2
\]
\[
\leq 2F_2^2.
\]
Averaging and median trick again

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

- 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 $\delta$ by independently doing $O(\log(1/\delta))$ estimators above.
- Total space $O(\log(1/\delta) \frac{1}{\epsilon^2} \log n)$
Geometric Interpretation

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

Richer model: Want to estimate a function of a vector $x \in \mathbb{R}^n$ which is initially assume to be the all 0’s vector. (previously we were thinking of the frequency vector $f$)

Each element $e_{ij}$ of a stream is a tuple $(i_j, j)$ where $i_j \in [n]$ and $i \in \mathbb{R}$ is a real-value: this updates $x_{ij}$ to $x_{ij} + j$. ($j$ can be positive or negative)
Observation: The estimation algorithm works even when $f_i$’s can be negative. What does this mean?

$$f_\perp = \sum f_i$$  
$$f_\perp = \sum_{i=1}^{n} f_i^2 = \sum_{i=1}^{n} x_i^2$$

Richer model:

- Want to estimate a function of a vector $x \in \mathbb{R}^n$ which is initially assume to be the all 0’s vector. (previously we were thinking of the frequency vector $f$)

- Each element $e_j$ of a stream is a tuple $(i_j, \Delta_j)$ where $i_j \in [n]$ and $\Delta_i \in \mathbb{R}$ is a real-value: this updates $x_{i_j}$ to $x_{i_j} + \Delta_j$. ($\Delta_j$ can be positive or negative)

$$\begin{pmatrix} +1 & +10 & -5 \\ 0, 0, 0, 0, \vdots, 0 \end{pmatrix} \quad \begin{pmatrix} (i, \Delta) \\ (i, 10) \\ (i, 5) \end{pmatrix}$$
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
\[ a_j = (i_j, \Delta_j) \text{ is current update} \]
\[ z \leftarrow z + \Delta_j Y_{i_j} \]
e ndWhile
Output \(z^2\)

\[
(2, 0), (10, 19), (1, -5.2), (\frac{d}{3}, 100), (1, -6) \]

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

\[
\sum x_i^2
\]

\[
f = (f_1, f_2, \ldots, f_n)
\]
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 $||x||_2^2$ 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 \( Z^2 \)

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

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

and hence

\[ \mathbb{E}[Z^2] = \sum_i x_i^2 = \|x\|_2^2. \]
Analysis

\[ Z = \sum_{i=1}^{n} x_i Y_i \] and output is \( Z^2 \)

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

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

and hence

\[
E[Z^2] = \sum_{i} x_i^2 = \|x\|_2^2.
\]

And as before one can show that \( \text{Var}(Z^2) \leq 2(E[Z^2])^2 \).
Introduction to (Linear) Sketching

A sketch of a stream $\sigma$ 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 $\sigma_1$ and $\sigma_1$ can be computed from $C(\sigma_1)$ and $C(\sigma_2)$. The output of the algorithm is some function of the sketch.
A *sketch* of a stream $\sigma$ 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 $\sigma_1$ and $\sigma_1$ 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?

\[
\begin{align*}
\mathcal{h} & \quad \mathcal{h} \\
\sigma_1 & \quad \sigma_2 \\
Z_1 & \quad Z_2 \\
Z_1^2 & \quad Z_2^2 \\
\underbrace{Z = Z_1 + Z_2} & \quad Z^2
\end{align*}
\]
Introduction to (Linear) Sketching

A *sketch* of a stream $\sigma$ 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 $\sigma_1$ and $\sigma_2$ 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 $C(\sigma_1 \cdot \sigma_2) = C(\sigma_1) + C(\sigma_2)$. 
Introduction to (Linear) Sketching

A sketch of a stream $\sigma$ 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 $\sigma_1$ and $\sigma_1$ 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?

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
V_n
\end{bmatrix}
\begin{bmatrix}
h_1 \\
h_2 \\
\vdots \\
h_m
\end{bmatrix}
= 
\begin{bmatrix}
\bar{x} \\
\bar{y}
\end{bmatrix}
= \bar{z}
\]

\[M \bar{x} = \bar{z}\]
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

$z \leftarrow z + \Delta_j M e_i$

endWhile

Output vector $z$ as sketch.

$M$ is compactly represented via $\ell$ hash functions, one per row, independently chosen from 4-wise independent hash family.
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.
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.