Streaming Algorithms

- Streaming Model of Computation
- Streaming Algorithms and DFA
- Stream: Motivation and Applications
- Synopsis: Sliding Window, Histogram, Wavelets
- Sampling from Stream: Reservoir Sampling
- Linear Sketch
- Count-Min Sketch
- AMS Sketch

Imdad ullah Khan

AMS Sketch

AMS sketch to estimate second frequency moment of a stream

- The AMS Sketch (Alon, Mathias, Szegedy, 1996)
- A sketch to estimate F_2 (paper has other algorithms for higher moments)

$$\mathcal{S} = < a_1, a_2, a_3, \ldots, a_m > \qquad a_i \in [n]$$

f_i: frequency of i in $\mathcal{S}$ $\quad \mathbf{F} = (f_1, f_2, \ldots, f_n)$

$$F_2 = \sum_{i=1}^{n} f_i^2$$ ▷ second frequency moment

Easy to compute if we store F ▷ $O(n)$ intgers

Can store $f_1 + f_2 + \ldots + f_n$ ▷ $O(1)$ integers

Also easy $(f_1 + f_2 + \ldots + f_n)^2$

AMS sketch

AMS sketch to estimate second frequency moment of a stream

$F_2 := \sum_{i=1}^{n} f_i^2$

Can store $f_1 + f_2 + \ldots + f_n$ ▷ $O(1)$ integers

$(f_1 + f_2 + \ldots + f_n)^2$ can be computed by the following algorithm

Algorithm : Compute square of sum of frequencies ($\mathcal{S}$)

$X \leftarrow 0$ ▷ sketch consists of 1 integer

On input a_i

$X \leftarrow X + 1$

return X^2

$X^2 = (f_1 + f_2 + \ldots + f_n)^2$

▷ Square of sum of frequencies, we want sum of squares of frequencies

AMS sketch

AMS sketch to estimate second frequency moment of a stream

$$F_2 = \sum_{i=1}^{n} f_i^2 = \underline{f_1^2 + f_2^2 + \ldots + f_n^2}$$ ▷ We want this

$$(f_1 + f_2 + \ldots + f_n)^2$$ ▷ Easy but overestimate

$$(f_1 + f_2 + f_3 + f_4)^2 = \underline{f_1^2 + f_2^2 + f_3^2 + f_4^2} + \underbrace{2(f_1 f_2 + f_1 f_3 + f_2 f_3 + f_1 f_4 + f_2 f_4 + f_3 f_4)}_{\text{error}}$$

What if we randomly add/subtract frequencies

$$(f_1 - f_2 + f_3 - f_4)^2 = \underline{f_1^2 + f_2^2 + f_3^2 + f_4^2} + \underbrace{2(-f_1 f_2 + f_1 f_3 - f_2 f_3 - f_1 f_4 + f_2 f_4 - f_3 f_4)}_{\text{error}}$$

AMS sketch

AMS sketch to estimate second frequency moment of a stream

What if we randomly add/subtract frequencies

$$(f_1 - f_2 + f_3 - f_4)^2 = \underline{f_1^2 + f_2^2 + f_3^2 + f_4^2} + \underbrace{2(-f_1 f_2 + f_1 f_3 - f_2 f_3 - f_1 f_4 + f_2 f_4 - f_3 f_4)}_{\text{error}}$$

Algorithm : AMS sketch to estimate second frequency moment of $\mathcal{S}$

Pick a random hash function $g : [n] \to \{-1, +1\}$

$X \leftarrow 0$ ▷ sketch consists of 1 integer

On input a_i

$X \leftarrow X + g(a_i)$

return X^2

$$X = f_1 g(1) + f_2 g(2) + \ldots + f_n g(n)$$

AMS sketch

AMS sketch is an unbiased estimate of second frequency moment

$X = f_1 g(1) + f_2 g(2) + \ldots + f_n g(n) \qquad X^2 = \big(f_1 g(1) + f_2 g(2) + \ldots + f_n g(n)\big)^2$

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \mathbb{E}\Big[\sum_i \big(f_i g(i)\big)^2\Big] + \mathbb{E}\Big[\sum_{i \neq j} f_i g(i) f_j g(j)\Big] \\
&= \mathbb{E}\Big[\sum_i f_i^2 g(i)^2\Big] + \mathbb{E}\Big[\sum_{i \neq j} f_i f_j g(i) g(j)\Big] \\
&= \sum_i f_i^2 \mathbb{E}\big[g(i)^2\big] + \sum_{i \neq j} f_i f_j \, \mathbb{E}\big[g(i) g(j)\big] = F_2
\end{aligned}$$

$\because \ \mathbb{E}\big[g(i)^2\big] = 1$ and $\mathbb{E}\big[g(i) g(j)\big] = 0$ for $i \neq j$

$$\mathbb{E}\left[X^2\right] = F_2$$

AMS sketch

The variance of AMS sketch estimate for F_2 is bounded

$$X^2 = \left(f_1 g(1) + f_2 g(2) + \ldots + f_n g(n) \right)^2 \qquad \mathbb{E}\left[X^2\right] = F_2$$

$$Var(X^2) \;=\; \mathbb{E}\left[X^4\right] - (\mathbb{E}\left[X^2\right])^2$$

$$\mathbb{E}\left[X^4\right] \;=\; \mathbb{E}\Big[\sum_i (f_i g(i))^4 + 6 \sum_{i \neq j} (f_i g(i)^2 f_j g(j))^2\Big] + \ldots$$

other terms: $\mathbb{E}\left[g(i)g(j)g(k)g(l)\right] = \mathbb{E}\left[g(i)^2 g(j) g(k)\right] = \mathbb{E}\left[g(i)^3 g(j)\right] = 0$

▷ 4-wise independence

$$\mathbb{E}\left[X^4\right] \;=\; \sum_i f_i^4 + 6 \sum_{i \neq j} f_i^2 f_j^2$$

$$Var(X^2) \;=\; \sum_i f_i^4 + 6 \sum_{i \neq j} f_i^2 f_j^2 - (\sum_i f_i^2)^2 \;=\; 4 \sum_{i \neq j} f_i^2 f_j^2 \;\leq\; 2F_2^2$$

Quality Specs of basic AMS Sketch

Algorithm : AMS sketch to estimate second frequency moment of $\mathcal{S}$

Pick a random hash function $g : [n] \mapsto \{-1, +1\}$

$X \leftarrow 0$ ▷ sketch consists of 1 integer

On input a_i

$X \leftarrow X + g(a_i)$

return X^2

$$\mathbb{E}\left[X^2\right] = F_2 \qquad Var\left(X^2\right) \leq 2F_2^2$$

Amplifying the probability of basic AMS Sketch

- Keep $k = {}^{8}/_{\epsilon^2} \times \log({}^{1}/_{\delta})$ estimates, $X_1, X_2, \ldots, X_k$
- Return $\bar{X}$: median of $\log({}^{1}/_{\delta})$ averages of groups of ${}^{2}/_{\epsilon^2}$ estimates

Algorithm : AMS sketch to estimate F_2 of $\mathcal{S}$ (ϵ, δ)

Pick $k = {}^{8}/_{\epsilon^2} \times \log({}^{1}/_{\delta})$ random hash functions $g_j : [n] \rightarrow \{-1, +1\}$

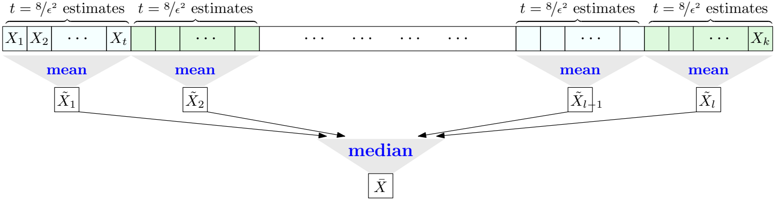
$X \leftarrow$ ZEROS(k) ▷ sketch consists of k integer

On input a_i

for $j = 1 \rightarrow k$ **do**

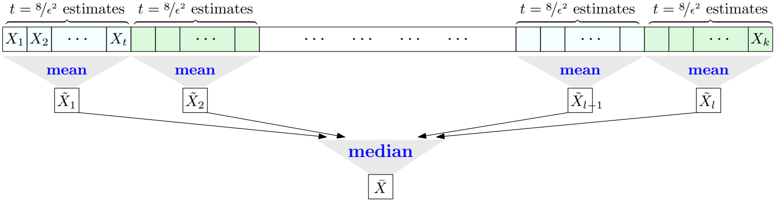
$X[j] \leftarrow X[j] + g_j(a_i)$

return $\bar{X}$: median of $\log({}^{1}/_{\delta})$ means of groups of ${}^{8}/_{\epsilon^2}$ estimates $(X[\cdot]^2)$



Amplifying the probability of basic AMS Sketch

- Keep $k = {}^{8}/_{\epsilon^2} \times \log({}^{1}/_{\delta})$ estimates, $X_1, X_2, \ldots, X_k$
- Return $\bar{X}$: median of $\log({}^{1}/_{\delta})$ averages of groups of ${}^{2}/_{\epsilon^2}$ estimates



- $\mathbb{E}\left[X_j^2\right] = F_2 \qquad Var\left(X_j^2\right) \leq 2F_2^2$
- $\mathbb{E}\left[\tilde{X}_j\right] = F_2 \qquad Var\left(\tilde{X}_j\right) \leq {}^{\epsilon^2}/_{4}F_2^2$
- $Pr\left[\,|\tilde{X}_j - F_2| \geq \epsilon F_2\right] \leq {}^{Var\left(\tilde{X}_j\right)}/_{\epsilon^2 F_2^2} = {}^{1}/_{4}$ ▷ Chebyshev Inequality
- $Pr\left[\,|\bar{X} - F_2| \geq \epsilon F_2\right] \leq \delta$

  The last inequality uses the Chernoff bound. For $\bar{X}$ to deviate this much from F_2 at least half of $\tilde{X}_j$ have to deviate more than that

Linear Transformation View of AMS Sketch

Algorithm : AMS sketch to estimate F_2 of $\mathcal{S}$

Pick k random hash functions $g : [n] \mapsto \{-1, +1\}$

$X \leftarrow \text{ZEROS}(k)$ ▷ sketch consists of 1 integer

On input a_i

for $j = 1 \to k$ **do**

  $X[j] \leftarrow X[j] + g_j(a_i)$

$\mathbf{g} =$

$g(1)$	$g(2)$	$\dots$		$g(n)$

F

f_1
f_2
$\vdots$
$\vdots$
f_n

$= X$

Algorithm : AMS sketch to estimate F_2 of $\mathcal{S}$

Pick k random hash functions $g : [n] \mapsto \{-1, +1\}$

$X \leftarrow \text{ZEROS}(k)$ ▷ sketch consists of 1 integer

On input a_i

for $j = 1 \to k$ **do**

$X[j] \leftarrow X[j] + g_j(a_i)$

$\mathbf{g} =$

$+1$	-1	$\dots$		$+1$

F

f_1
f_2
$\vdots$
$\vdots$
f_n

$= X$

Linear Transformation View of AMS Sketch

Algorithm : AMS sketch to estimate F_2 of $\mathcal{S}$

Pick k random hash functions $g : [n] \mapsto \{-1, +1\}$

$X \leftarrow \text{ZEROS}(k)$ ▷ sketch consists of 1 integer

On input a_i

for $j = 1 \rightarrow k$ **do**

$X[j] \leftarrow X[j] + g_j(a_i)$

$\mathbf{G} =$

$+1$	-1	$\dots$		$+1$
-1	-1	$\dots$		-1

F

f_1
f_2
$\vdots$
$\vdots$
f_n

X

X_1
X_2

Linear Transformation View of AMS Sketch

Algorithm : AMS sketch to estimate F_2 of $\mathcal{S}$

Pick k random hash functions $g : [n] \mapsto \{-1, +1\}$

$X \leftarrow \text{ZEROS}(k)$ ▷ sketch consists of 1 integer

On input a_i

for $j = 1 \to k$ **do**

$X[j] \leftarrow X[j] + g_j(a_i)$

G =

$+1$	-1	$\dots$		$+1$
-1	-1	$\dots$		-1
$\vdots$		$\dots$		$\vdots$
-1	$+1$	$\dots$		-1

F

f_1
f_2
$\vdots$
$\vdots$
f_n

X

X_1
X_2
$\vdots$
X_k

Linear Transformation View of AMS Sketch

$$\mathbf{G} = \begin{array}{|c|c|c|c|c|} \hline +1 & -1 & \dots & & +1 \\ \hline -1 & -1 & \dots & & -1 \\ \hline \vdots & & \dots & & \vdots \\ \hline -1 & +1 & \dots & & -1 \\ \hline \end{array}$$

$$\mathbf{F} = \begin{array}{|c|} \hline f_1 \\ \hline f_2 \\ \hline \vdots \\ \hline \vdots \\ \hline f_n \\ \hline \end{array} \qquad \mathbf{X} = \begin{array}{|c|} \hline X_1 \\ \hline X_2 \\ \hline \vdots \\ \hline X_k \\ \hline \end{array}$$

$$\bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i^2 \qquad \Pr\left[|\bar{X} - F_2| > \epsilon F_2\right] \leq \delta$$

With probability at leat $1 - \delta$

$$(1-\epsilon)\sum_{i=1}^{n} f_i^2 < \frac{1}{k}\sum_{i=1}^{k} X_i^2 < (1+\epsilon)\sum_{i=1}^{n} f_i^2$$

$$\sqrt{(1-\epsilon)}\|F\|_2 < \frac{1}{\sqrt{k}}\|X\|_2 < \sqrt{(1+\epsilon)}\|F\|_2$$

AMS Sketch as a dimensionality reduction algorithm

$$\mathbf{G} = \begin{array}{|c|c|c|c|c|} \hline +1 & -1 & \dots & & +1 \\ \hline -1 & -1 & \dots & & -1 \\ \hline \vdots & & \dots & & \vdots \\ \hline -1 & +1 & \dots & & -1 \\ \hline \end{array}$$

$$\mathbf{F} = \begin{array}{|c|} \hline f_1 \\ \hline f_2 \\ \hline \vdots \\ \hline \vdots \\ \hline f_n \\ \hline \end{array} \qquad \mathbf{X} = \begin{array}{|c|} \hline X_1 \\ \hline X_2 \\ \hline \vdots \\ \hline X_k \\ \hline \end{array}$$

$$\sqrt{(1-\epsilon)}\|F\|_2 \; < \; \frac{1}{\sqrt{k}}\|X\|_2 \; < \; \sqrt{(1+\epsilon)}\|F\|_2$$

G is a random linear transformation reduces the dimension of F while preserving its ℓ_2-norm

Since G is linear it is easy to see that given $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$

$$\text{w.h.p} \qquad \|\frac{1}{\sqrt{k}}\mathbf{Gu}\|_2 - \|\frac{1}{\sqrt{k}}\mathbf{Gv}\|_2 \; \sim \; \|\mathbf{u} - \mathbf{v}\|_2$$