Streaming

Setting: Set $[n]$ of items. Items come in a "stream" $e_1, e_2, \ldots, e_m \in [n]$. Once read, e_i is "gone."

(applications: massive real-time data, network switch)

Total allowed memory: $\text{poly}\log(n, m)$ bits (e.g., $\log n + \log^2 n \log^3 m$)

So cannot remember $e_1, \ldots, e_m$

$\forall i \in [n]$, let $f_i = (\#$ times i appeared). Then $f = (f_1, \ldots, f_n)$ is n-dimensional vector.

Goal: At the end of the stream output some function of f.

- $\|f\|_0 = |\{i : f_i > 0\}| = (\#$ of distinct items)
- $\|f\|_2^2 = \|f\|^2 = \sum_i f_i^2$ (this lecture)
- $\|f\|_\infty = \max_i f_i = (\text{frequency of most frequent item})$

Alon-Matias-Szegedy

AMS-F_2-Basic

Let $h : [n] \to \{+1, -1\}$ be a 4-wise independent random hash function.

$z = 0$.

While (stream is not empty)

  Get e_i.

  $z = z + h(e_i)$

End

Output z^2.

Example

Stream = (1, 2, 3, 2, 3, 2) (m = length of stream = 6
n = # items = 3)

Assume $h(1)=1$, $h(2)=1$, $h(3)=-1$.

i	e_i	f	Z	
1	1	(1, 0, 0)	1	($h(1)=1$)
2	2	(1, 1, 0)	2	($h(2)=1$)
3	3	(1, 1, 1)	1	($h(3)=-1$)
4	2	(1, 2, 1)	2	($h(2)=1$)
5	3	(1, 2, 2)	1	($h(3)=-1$)
6	2	(1, 3, 2)	2	($h(2)=1$)

Output $Z^2=4$.

If $h(1)=h(2)=h(3)=1$, Z would be 6, so output = 36.
($\|f\|_2^2 = 1^2+3^2+2^2=14$)

For any fixed stream $e_1, \ldots, e_m$, z^2 is a random variable (since h is.)

$$\mu$$

$$\mathbb{E}[z^2] = \mathbb{E}\left[\left(\sum_{i=1}^{m} h(e_i)\right)^2\right] = \mathbb{E}\left[\sum_{i,j=1}^{m} h(e_i)h(e_j)\right] = \sum_{i,j} \mathbb{E}[h(e_i)h(e_j)]$$

$$= \sum_{\substack{i,j: \\ e_i = e_j}} 1 = \sum_{t=1}^{n} f_t^2 = \|f\|^2$$

if $e_i \neq e_j$ $h(e_i)$ and $h(e_j)$ are independent, so $\mathbb{E}[h(e_i)h(e_j)] = \mathbb{E}[h(e_i)]$.

$\mathbb{E}[h(e_j)] = 0$.

$$\mathbb{E}[z^4] = \sum_{i,j,k,l=1}^{m} \mathbb{E}[h(e_i)h(e_j)h(e_k)h(e_l)]$$

if one of e_i, e_j, e_k, e_l appears only once (say e_i), becomes $\mathbb{E}[h(e_i)] \cdot \mathbb{E}[h(e_j)h(e_k)h(e_l)] = 0$.

$$= \sum_{\substack{i,j,k,l: \\ e_i = e_j = e_k = e_l}} 1 + \sum_{\substack{i,j,k,l: \\ e_i = e_j \neq e_k = e_l}} 1 + \sum_{\substack{i,j,k,l: \\ e_i = e_k \neq e_j = e_l}} 1 + \sum_{\substack{i,j,k,l \\ e_i = e_l \neq e_j = e_k}} 1$$

$$= \sum_{t=1}^{n} f_t^4 + 6 \sum_{t=1}^{n-1} \sum_{s=t+1}^{n} f_t^2 f_s^2$$

Then $\sigma^2 = \mathbb{E}[z^4] - \mathbb{E}[z^2]^2 \leq \mathbb{E}[z^4] \leq 3\mu^2$

(because $3\mu^2 = 3\sum_{j=1}^{n} f_j^4 + 6\sum_{j<k} f_j^2 f_k^2 \geq \mathbb{E}[z^4]$)

Then by sampling $h_1, \ldots, h_d$ independently and computing z_i according to h_i guarantees that if $\bar{z}^2 = (z_1^2 + \cdots + z_d^2)/d$, $\mathrm{Var}[\bar{z}^2] \leq \mathrm{Var}[z_1]/d \leq 3\mu^2/d$, so when $d \geq \frac{24}{\varepsilon^2}$, $\mathrm{Var}[\bar{z}^2] \leq \varepsilon^2\mu^2/8$ and Chebyshev, $\Pr[|\bar{z}^2 - \|f\|^2| > \varepsilon\|f\|^2] \leq 1/4$.

Reduce error probability to $\delta > 0$, let $t = c \cdot \log(1/\delta)$ for some constant c, and compute $\bar{Z}_1^2, \ldots, \bar{Z}_t^2$ independently and output the median.

- If the median "fails" (i.e., not in $[1-\varepsilon, 1+\varepsilon] \cdot \|f\|^2$), then at least of half of $\bar{Z}_1^2, \ldots, \bar{Z}_t^2$ should "fail", but the expected fraction of failure is $\leq 1/4$.

$\therefore$ Chernoff bound will show that $\Pr[\text{median fails}] \leq e^{-c' \cdot t}$ for some constant c', so by setting $t = c \cdot \log(1/\delta)$, $e^{-c' \cdot t} \leq \delta$.

$\therefore$ Total memory: $O\left(\frac{1}{\varepsilon^2} \cdot \log\left(\frac{1}{\delta}\right) (\log n + \log m)\right)$ bits (# hash functions)

Guarantee: Output $(1 \pm \varepsilon)$-approximate estimate of $\|f\|^2$ w.p. $\geq 1 - \delta$.

Linear Sketch

What was going on in AMS?

- f: freq. vector (one entry updated each iteration)
- $d = O(1/\varepsilon^2)$, $t = O(\log(1/\delta))$ s.t.
  sampled dt hash functions $h_{ij} \in \mathcal{H}$. $(i \in [t], j \in [d])$
  and let $Z_{ij} = \langle h_{ij}, f \rangle$.
  ↖ $h_{ij} : [n] \to \{\pm 1\}$ can be viewed as n-dim. vec.
- $\forall i$, $\overline{Z_i}^2 = (Z_{i1}^2 + \cdots + Z_{id}^2)/d$ and
  final answer = median $(\overline{Z}_1, \ldots, \overline{Z}_t)$.
- Consider $(dt \times n)$-matrix H s.t.

$$H = \begin{bmatrix} — & h_{11} & — \\ & \vdots & \\ — & h_{td} & — \end{bmatrix}, \text{ and } Z = Hf \in \mathbb{R}^{dt}.$$

Z is a "sketch" of f!

* $dt \ll n$ (if ε, δ are not too small)
* $\|f\|_2^2$ can be computed only from Z.
* f can be efficiently stored / computed!

Furthermore, sketch function $Z = Hf$ is "linear";
if you have f_1 and f_2, then sketch of $(f_1 + f_2)$
$= (\text{sketch of } f_1) + (\text{sketch of } f_2)$.

k-wise Independence

Fix $N \geq m$. Assume they are powers of 2.

Lemma. $\forall k \geq 1$, $\exists$ k-wise independent $\mathcal{H}$ s.t.
- Sampling $h \in \mathcal{H}$ takes time $O(k \cdot \log N)$
- Storing h takes $O(k \cdot \log N)$ bits
- Given $i \in [N]$, $h(i)$ can be computed in time $\text{poly}(k, \log N)$.

Proof. Will use "finite field" – a finite set s.t. addition, subtraction, multiplication, division are well-defined.

Can treat it as "set of numbers" like $\mathbb{R}$ and $\mathbb{C}$, and study algebra

(e.g., $\mathbb{Z}_p = \{0, \ldots, p-1\}$ where operations are as usual mod p.

non-example: $\mathbb{Z}_6$, $\mathbb{Z}$, $\mathbb{N}$, etc.)

Theorem. For any prime p and $\ell \in \mathbb{N}$, there exists a finite field $\mathbb{F}_{p^\ell}$ with $|\mathbb{F}_{p^\ell}| = p^\ell$. (And they are the only ones.) $\square$

Fix $r : \mathbb{F}_N \to [m]$ s.t. $\forall i \in [m]$, $|r^{-1}(i)| = N/m$.

($\mathbb{F}_N \to [m]$)

Sampling k-wise independent hash function.

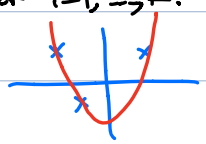
Sample $a_0, \ldots, a_{k-1} \in \mathbb{F}_N$ independently.

Let $f(x) = a_0 + a_1 x + \cdots + a_{k-1} x^{k-1}$ be a deg-$(k-1)$ polynomial over $\mathbb{F}_N$.

For $x \in \mathbb{F}_N$, $h(x) = r(f(x))$.

Remark. Enough to store $a_0, \dots, a_{k-1}$. ($k \log N$ bits)
Given $x \in \mathbb{F}_N$ and $a_0, \dots, a_{k-1}$, $f(x)$ can be computed by $O(k)$ additions / multiplications over $\mathbb{F}_N$. ($\exists$ efficiently computable r too.)

Fix (all distinct) $z_1, \dots, z_k \in \mathbb{F}_N$. $\forall y_1, \dots, y_k \in \mathbb{F}_N$, there is exactly one polynomial $f(x) = a_0 + a_1 x_1 + \dots + a_{k-1} x^{k-1}$ s.t. $f(z_i) = y_i$ for $i = 1, \dots, k$.
(k points uniquely decides deg $k-1$ poly!)
$\Rightarrow \Pr[f(z_i) = y_i \ \forall i] = 1/N^k$.



So, $\forall z_1, \dots, z_k \in \mathbb{F}_N$, $\forall w_1, \dots, w_k \in [m]$, there are exactly $(N/m)^k$ choices of $y_1, \dots, y_k \in \mathbb{F}_N$ s.t. $r(y_i) = w_i \ \forall i = 1, \dots, k$.

Then, $$\Pr[h(z_i) = w_i \ \forall i] = \sum_{\substack{y_1, \dots, y_k : \\ r(y_i) = w_i \ \forall i}} \Pr[f(z_i) = y_i \ \forall i]$$

$$= \left(N/m\right)^k \cdot 1/N^k = 1/m^k$$

□