Streaming Algorithms

Suppose a stream of data

1 3 2 5 5 8 7 3 ...

Q: What properties can we store in low memory/space?

Addition, max, min : easy in $O(1)$ space (or $O(\log n)$ bits)

Median? much harder!

Distinct Elements?

Heavy Hitters: elements that have appeared $\geq 1\%$ of the time?

Frequency metrics are hard.

Streams as Vectors (of frequencies):

Universe U, each element is in U. Operations "add $u \in U$" "delete $u \in U$"

Maintain frequency vector $(f_1^t, \ldots, f_{|U|}^t) = \vec{f}$

$f_i^t \geq 0$: current # of i at time t

Maintaining $\|\vec{f}\|_1$ is easy. $\|x\|_2$? Biased towards "heavy hitters"

$\|x\|_0$? # distinct elements

Estimating $\|\vec{f}\|_2$ [Alon-Matias-Szegedy '00]

Algo: Initialization:

Algo: Initialization:

- Pick a random <u>hash function</u> $h: U \to \{-1, +1\}$
- Maintain integer C (counter), initially 0.

On each update "add i": $C \leftarrow C + h(i)$

"del i": $C \leftarrow C - h(i)$

Hash function $U \to R$ is <u>k-universal</u> (k-wise independent) if

$\forall$ k elements $i_1 .. i_k$ and k values $a_1 .. a_k$,

$$\Pr[h(i_1) = a_1, \ldots, h(i_k) = a_k] = \frac{1}{|R|^k}.$$ ← "any k behave independently"

Fact: if k-universal, then k'-universal $\forall k' < k$.

Suppose h is 4-universal.

① h is 1-universal, so $\Pr_h(h(i) = 1) = \Pr_h(h(i) = -1) = \frac{1}{2}$.

② For all i, j, k, l:

- $\mathbb{E}[h(i) \cdot h(j)] = \mathbb{E}[h(i)] \cdot \mathbb{E}[h(j)]$
- $\mathbb{E}[h(i) \cdot h(j) \cdot h(k) \cdot h(l)] = \mathbb{E}[h(i)] \cdot \mathbb{E}[h(j)] \cdots \mathbb{E}[h(l)]$

Claim: $\mathbb{E}[C^2] = \|\vec{f}\|_2^2$.

Proof: $\mathbb{E}[C^2] = \mathbb{E}\left[\sum_{i,j} (h(i) x_i \cdot h(j) x_j)\right]$

$$= \sum_{i,j} x_i x_j \, \mathbb{E}[h(i) \cdot h(j)]$$

$$= \sum_{i,j} x_i x_j \, \mathbb{1}(h(i) = h(j))$$

$$= \sum_{i,j} x_i x_j \, \mathbb{1}(h(i)=h(j))$$

$$= \sum_i f_i^2 = \|\vec{f}\|_2^2.$$

Claim: $\mathrm{Var}(C^2) =$

Proof: $\mathbb{E}[(C^2)^2] = \mathbb{E}\left[\sum_{ijkl} h(i)h(j)h(k)h(l)\, x_i x_j x_k x_l\right]$

$$= \sum_i x_i^4 \mathbb{E}[h(i)^4] + 6\sum_{i<j} x_i^2 x_j^2 \mathbb{E}[h(i)^2 h(j)^2] + \text{other terms } 0!$$

$$= \sum_i x_i^4 + 6\sum_{i<j} x_i^2 x_j^2$$

So $\mathrm{Var}[C^2] = \mathbb{E}[(C^2)^2] - \mathbb{E}[C^2]^2$

$$= \sum_i x_i^4 + 6\sum_{i<j} x_i^2 x_j^2 - \left(\sum_i x_i^2\right)^2$$

$$= 4\sum_{i<j} x_i^2 x_j^2 \le 2\,\mathbb{E}[C^2]^2.$$

By Chebyshev: $\Pr\left[|C^2 - \mathbb{E}[C^2]| > \varepsilon\,\mathbb{E}[C^2]\right] \le \dfrac{\mathrm{Var}(C^2)}{(\varepsilon\,\mathbb{E}[C^2])^2} \le \dfrac{2}{\varepsilon^2}.$

only good if $\varepsilon \ll 1$...

Reduce Variance by Repetition

If estimator of mean μ, variance σ^2, then average of k indep. copies: mean μ, var σ^2/k.

mean of k indep. copies: $\Pr \le \dfrac{\mathrm{Var}(C^2)}{(\varepsilon\,\mathbb{E}[C^2])^2} \le \dfrac{2k}{\varepsilon k^2}$

Set $k = \dfrac{2}{\varepsilon^2 \delta}$, prob δ of error.

Matrix View and Dimension Reduction

If $\vec{f}$ is the frequency vector, then estimator C is simply

$$\begin{bmatrix} h(e_1) & h(e_2) & \cdots & h(e_{|U|}) \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{|U|} \end{bmatrix}$$

We take average of $k = \frac{2}{\varepsilon^2 \delta}$ estimators, which is like computing

$$\frac{1}{k} \begin{bmatrix} \text{---} & h_1 & \text{---} \\ \text{---} & h_2 & \text{---} \\ & \vdots & \\ \text{---} & h_k & \text{---} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{|U|} \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_k \end{bmatrix} \leftarrow \vec{c}$$

If $h_1 \ldots h_k$ were truly random (not just 2-wise indep), then $\begin{bmatrix} \text{---} & h_1 & \text{---} \\ & \vdots & \\ \text{---} & h_k & \text{---} \end{bmatrix}$ is truly random ± 1 matrix. JL $\Rightarrow \|f_2\|_2^2 \in (1 \pm \varepsilon) \|\vec{c}\|_2^2$

J-L bound gives $k = O(\varepsilon^{-2} \log \frac{1}{\delta})$ for prob $\geq 1 - \delta$. We only get $\frac{2}{\varepsilon^2 \delta}$.

Reason is we use Chebyshev instead of Chernoff, since we only have 4-wise indep.

Estimating $\|\vec{f}\|_0$: <u>addition-only</u>

Idea: map each $u \in U$ to random real $\in [0,1]$ and take minimum: more distinct $u \in U \Rightarrow$ min should be smaller.

In fact, s random $\in [0,1]$: $\mathbb{E}[\text{smallest}] = \frac{1}{s+1}$.

Algo: Initialization:

- Pick random hash function $h: U \to \{0, \frac{1}{M}, \frac{2}{M}, \dots, \frac{M-1}{M}, 1\}$

On each update:

- Let L_t be s^{th} smallest <u>distinct</u> hash value seen so far
- output estimate $D_t = \frac{s}{L_t}$.

Thm: If H is 2-universal, and M large enough, then

① $\Pr[D_t > (1+\delta) d] \le O\left(\frac{1}{\delta^2 s}\right)$

② $\Pr[D_t < (1-\delta) d] \le O\left(\frac{1}{\delta^2 s}\right)$

Proof of ②: $D_t < (1-\delta)d \iff L_t > \frac{s}{(1-\delta)d}$

$\iff$ $< s$ elements hashed to value $\le \frac{s}{(1-\delta)d}$

(compared to expected $\approx \frac{s}{1-\delta}$)

$$X_i = \begin{cases} 1 & \text{if } h(e_i) \le \frac{s}{(1-\delta)d} \\ 0 & \text{otherwise} \end{cases}, \quad \Pr[X_i = 1] \approx \frac{s}{(1-\delta)d}$$

$X = \sum X_i = \#$ elements hashed to $\le \frac{s}{(1-\delta)d}$

$\mathbb{E}X \approx \frac{s}{1-\delta}$, $\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2 \le \mathbb{E}X^2 = \mathbb{E}X \approx \frac{s}{1-\delta}$

$$\begin{aligned}
\Pr[X < s] &= \Pr[X < (1-\delta)\mathbb{E}X] \\
&\le \Pr[|X - \mathbb{E}X| \ge \delta \mathbb{E}X] \\
&\le \frac{\operatorname{Var} X}{\delta^2 (\mathbb{E}X)^2} \\
&= \frac{1}{\delta^2 \mathbb{E}X} = \frac{1-\delta}{\delta^2 s}
\end{aligned}$$

$$= \frac{1}{\delta^2 \mathbb{E}X} = \frac{1-\delta}{\delta^2 s}$$

k-wise independent hash functions from polynomials

Case $U = R = \mathbb{F}$, finite field $\mathbb{F}$

Sampling h:

① Sample random coefficients $c_0, c_1, \dots, c_{k-1} \in \mathbb{F}$

② Hash function is $h(x) = c_0 + c_1 x + c_2 x^2 + \dots + c_{k-1} x^{k-1}$.

To show: $\Pr[h(i_1) = a_1, h(i_2) = a_2, \dots, h(i_k) = a_k] = \frac{1}{|\mathbb{F}|^k}$.

$\iff$ there is a unique solution $c_0, c_1, \dots, c_{k-1} \in \mathbb{F}$ s.t.

$$c_0 + i_1 c_1 + i_1^2 c_2 + \dots + i_1^{k-1} c_{k-1} = a_1$$
$$c_0 + i_2 c_1 + i_2^2 c_2 + \dots + i_2^{k-1} c_{k-1} = a_2$$
$$\vdots$$
$$c_0 + i_k c_1 + i_k^2 c_2 + \dots + i_k^{k-1} c_{k-1} = a_k$$

k equations, k unknown variables.

If system is independent, then unique solution.

Matrix form:

$$\underbrace{\begin{bmatrix} 1 & i_1 & i_1^2 & i_1^3 & \dots & i_1^{k-1} \\ 1 & i_2 & i_2^2 & i_2^3 & \dots & i_2^{k-1} \\ & & \vdots & & & \\ 1 & i_k & i_k^2 & i_k^3 & \dots & i_k^{k-1} \end{bmatrix}}_{A} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{k-1} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

$$\begin{bmatrix} 1 & i_k & i_k^2 & i_k^3 & \cdots & i_k^{k-1} \end{bmatrix} \begin{bmatrix} \ \end{bmatrix}$$

Claim: the Vandermonde matrix is independent.

Proof: suffices to show columns are indep.

Suppose $b_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + b_1 \begin{bmatrix} i_1 \\ i_2 \\ \vdots \\ i_k \end{bmatrix} + \cdots + b_{k-1} \begin{bmatrix} i_1^{k-1} \\ i_2^{k-1} \\ \vdots \\ i_k^{k-1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

Then $i_1, i_2, \ldots, i_k$ are roots of the polynomial

$$b_0 + b_1 x + \cdots + b_{k-1} x^{k-1}$$

But a nonzero poly of degree $\leq k-1$ has $\leq k-1$ roots.

So zero polynomial $\Rightarrow$ $b_0 = b_1 = \cdots = b_{k-1} = 0$.