6.854 | 18.419: Advanced Algorithms

Last Time: Consistent hashing and random trees

hash functions that evolve well

routing schemes that work with inconsistent views

Today: Distinct elements and count-min

Meta Question: what can we do if we can't store our data, but it is streaming by?

O(log n) bits of memory

Problem #1: Count the number of distinct elements in a sequence

$$x_1, x_2, x_3, \ldots$$

How many distinct words did Shakespeare use?

31,534 of which 14,376 only used once

"Using only memory equivalent to 5 lines of printed text, you can estimate with a typical accuracy of 5% and in a single pass [the number of words Shakespeare used]"

Durand and Flajolet 2003

Attempt #1: Choose a random hash function

$$h: U \rightarrow [0,1]$$

Pass through data $h(x_1), h(x_2), h(x_3), \ldots$ and compute min

Let $Y = \min(h(x_1), h(x_2), h(x_3), \ldots)$

Lemma: If there are N distinct elements in $x_1, x_2, x_3, \ldots$ then $\mathbb{E}[Y] = 1/N+1$

Pf:
$$\mathbb{E}[Y] = \int_{z=0}^{1} \Pr[Y \geq z] \, dz$$ (convince yourself in discrete case)

$$= \int_0^1 N z (1-z)^{N-1} dz$$

cdf $\Rightarrow$ pdf

$$= \int_{z=0}^{1} (1-z)^N dz = -\frac{(1-z)^{N+1}}{N+1} \Big|_{z=0}^{z=1}$$

$$= \frac{1}{N+1}$$ ☒

Alternatively $\mathbb{E}[Y] =$ "probability of choosing $N+1$ values in $[0,1]$, and last is min"

$= 1/N+1$ (by symmetry)

How do we bound Y's deviation? Markov? Chebyshev? Chernoff?

Lemma: $\mathrm{var}(Y) < \left(\frac{1}{N+1}\right)^2$

Pf:

$$\mathbb{E}[Y^2] = \int_0^1 N z^2 (1-z)^{N-1} dz = \frac{2}{(N+1)(N+2)}$$

$$\mathrm{var}(Y) = \underbrace{\mathbb{E}[Y^2]}_{\frac{2}{(N+1)(N+2)}} - \underbrace{(\mathbb{E}[Y])^2}_{\frac{1}{(N+1)^2}} = \left(\frac{1}{N+1}\right)\left(\frac{N}{(N+1)(N+2)}\right) \quad \boxtimes$$

Alternatively it is a beta distribution with $\alpha = 1, \beta = N$

$$\mathrm{var} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

But the standard deviation is comparable to expectation
(why is this bad?)

any ideas how to fix?

Attempt #2 Choose k hash functions almost Flajolet-Martin

$$h_1, h_2 \ldots h_k : U \to [0,1]$$

Find minimum $Y_1, Y_2 \ldots Y_k$ for each one

Set $\bar{Y} = \frac{1}{k}\sum_{i=1}^{k} Y_i$, output $1/\bar{Y} - 1$

Analysis: $\mathbb{E}[\bar{Y}] = 1/N+1$, $\mathrm{var}(\bar{Y}) < \frac{1}{k(N+1)^2}$

Thus by Chebyshev

$$\Pr\left[\left|\bar{Y} - \frac{1}{N+1}\right| > \frac{\varepsilon}{N+1}\right] \leq \frac{\frac{1}{k\cancel{(N+1)^2}}}{\frac{\varepsilon^2}{\cancel{(N+1)^2}}} = \frac{1}{k\varepsilon^2}$$

Then $\frac{N+1}{1+\varepsilon} \leq \frac{1}{\bar{Y}} \leq \frac{N+1}{1-\varepsilon}$, what about bit complexity of min?

(First, Misra-Gries)

Problem #2: ε-approximate frequency counts

Given a sequence $x_1, x_2, x_3 \ldots x_n$ output a list of values so that

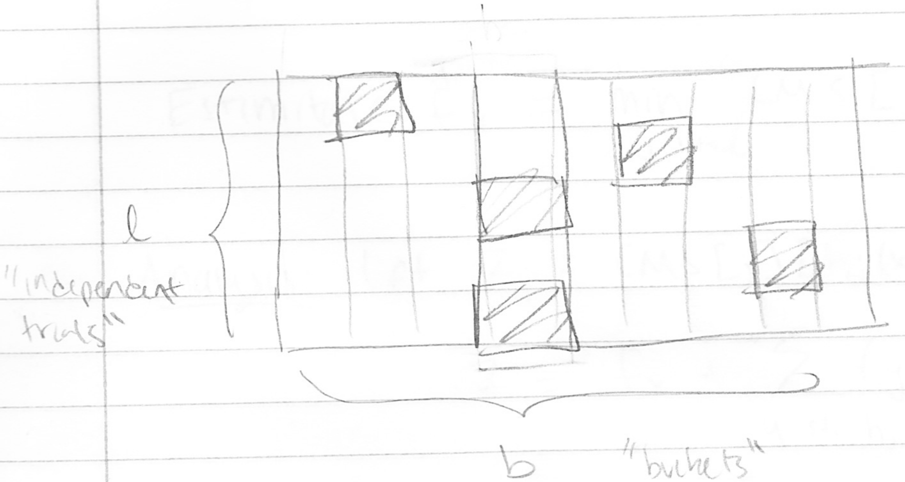
be able to answer queries $f_x - \varepsilon n \le \text{Count}(x) \le f_x + \varepsilon n$

(1) every value that occurs at least $\frac{n}{k}$ times is on list

weaker goal: answer queries on frequencies

(2) no value that occurs less than $\frac{n}{k} - \varepsilon n$ times is on list

Count-Min Sketch: Cormode, Muthukrishnan 2005



CMS

[box] = $h_j(x)$'s

Choose ℓ random hash functions $h_1, h_2, \ldots h_\ell : U \to [b]$

For all x,
  For $j = 1$ to ℓ
    Increment CMS[j][$h_j(x)$]

How do we approximate the number of times x appeared?

Let f_x = frequency of x

Claim: For all x, j

$$CMS[j][h_j(x)] \geq f_x$$

Pf: Fix j. Each time we see x, we increment

$$CMS[j][h_j(x)]$$

although we can increment it for other reasons (collisions) $\boxtimes$

Estimate: $\hat{f}_x = \min_{j=1 \text{ to } \ell} CMS[j][h_j(x)]$

Analysis: Let $Z_j = CMS[j][h_j(x)]$

$$Z_j = f_x + \sum_{y \text{ st. } h_j(y)=h_j(x)} f_y$$

2-universal hash

Hence $\mathbb{E}[Z_j] = f_x + \frac{\sum_y f_y}{b} \leq f_x + \frac{n}{b}$

markov $Z_j - f_j$ nonnegative

Set $b = \frac{2}{\varepsilon}$, then $\Pr[Z_j - f_x \geq \varepsilon n] \leq \frac{1}{2}$

Finally $\Pr\left[\min_{j=1 \text{ to } \ell} Z_j \geq f_x + \varepsilon n\right] \leq \frac{1}{2^\ell}$

$O(\log n)$

Space: $O(k \log n)$ where $\varepsilon = \frac{1}{k}$

heavy hitters

Frequent Items, Misra - Gries 1982

Given a sequence $x_1, x_2, x_3, \ldots x_n$ output a list with

(1) at most k value

(2) every value that appears at least $\frac{n}{k+1}+1$ times is on (list)

Initialize empty list

For each item

If same as some item on list, increment its counter

Else if less than k items in list, add to list set its counter to one

Else decrement all counters, delete items whose counter reaches zero

let f_x = frequency of value x

Lemma: At end, value x's counter is at least $f_x - n/_{k+1}$

either x is decremented or not added

Pf: Each elimination of x eliminates k other symbols

Hence no more than $\frac{n}{k+1}$ occurrences of x can be eliminated ☐

Space: $O(k \log n)$