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?

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]."—Duran 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 $\mathbb{E}[Y] = \frac{1}{N+1}$

**P.F.:**

\[
\mathbb{E}[Y] = \int \frac{\Pr[Y \geq z]}{N+1} \, dz
\]

\[
= \int_0^{\frac{1}{N+1}} (1-z)^N \, dz = \left[ -\frac{(1-z)^{N+1}}{N+1} \right]_0^{\frac{1}{N+1}} = \frac{1}{N+1}
\]

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

= \frac{1}{N+1} \quad \text{(by symmetry)}$


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

**P.F.**

\[
\mathbb{E}[Y^2] = \int N^2 z^2 (1-z)^{N-1} \, dz = \frac{2}{(N+1)(N+2)}
\]
\[
\text{var}(Y) = \# [Y^2] - (\mathbb{E}[Y])^2
\]
\[
= \frac{2}{(N+1)(N+2)} - \left(\frac{1}{N+1}\right)^2 = \left(\frac{1}{N+1}\right) \left(\frac{N}{(N+1)(N+2)}\right)
\]
Alternatively, it is a beta distribution with \(\alpha = 1, \beta = N\)

\[
\text{var} = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}
\]

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

Any ideas how to fix?

Attempt #2: Choose 2 hash functions

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

Find minimum \(Y_1, Y_2, \ldots, Y_k\) for each one

Set \(\overline{Y} = \frac{1}{k} \sum_{i=1}^{k} Y_i\), output \(1/\overline{Y} - 1\)

Analysis: \(\mathbb{E}[\overline{Y}] = \frac{1}{N+1}\), \(\text{var}(\overline{Y}) < \frac{1}{k(N+1)^2}\)

Thus by Chebyshev
\[
\Pr \left[ |\overline{Y} - \frac{1}{N+1}| > \frac{\varepsilon}{N+1} \right] \leq \frac{\text{var}(\overline{Y})}{\varepsilon^2} = \frac{1}{k(N+1)^2} < \frac{1}{k3^2}
\]

Then
\[
\frac{N+1}{k+1} \leq \frac{1}{\overline{Y}} \leq \frac{N+1}{k-1} 
\]

what about bit complexity of min?
Problem #2: ε-approximate frequency counts

Given a sequence \( x_1, x_2, x_3, \ldots, x_n \) adapt a list of values so that:

1. Every value that occurs at least \( \frac{n}{12} \) times is on list
2. No value that occurs less than \( \frac{n}{12} - \varepsilon n \) times is on list

Count-Min Sketch: Cormode, Muthukrishnan 2005

Choose \( l \) random hash functions \( h_1, h_2, \ldots, h_l : U \to [b] \)

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 $f_x = \text{frequency of } x$

Claim: For all $x, j$

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

Proof: Fix $j$. Each time we see $x$, we increment

$$\text{CMS}[j][h_j(x)]$$

although we can increment it for other reasons (collisions)

Estimate: $f_x = \min_{j: \text{time}} \text{CMS}[j][h_j(x)]$

Analysis: Let $z_j = \text{CMS}[j][h_j(x)]$

$$z_j = f_x + \sum_{y \text{ s.t. } h_j(y) = h_j(x)} f_y$$

Hence $E[z_j] = f_x + \frac{\sum f_y}{b} \leq f_x + \frac{n}{b}$

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

Finally, $Pr[\min_{j: \text{time}} z_j \geq f_x + 3\varepsilon n] \leq \frac{1}{2e^{2}}$

$$\text{space: } O(k log n) \text{ where } \varepsilon = \frac{1}{k}$$
Frequent Items, Misra-Gries 1982

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

1. at most $k$ values
2. every value that appears at least $\frac{n}{k+1}$ times

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 $f_x = \frac{n}{k+1}$

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: $O(k \log n)$