Today’s problem: Distinct Elements

- Given a stream of integers \( < x_1, \ldots, x_m > \) where \( x_i \in [U] := \{1, 2, \ldots, u\} \), output the number \( n \) of distinct elements seen.

- **Example**: There exist 5 distinct elements in the stream \( < 3, 3, 1986, 1, 6, 12, 1, 12, 6, 1, 3 > \), i.e., \( n = 5 \).

- The number of distinct elements of a stream is also known as its \( (F_0) \) moment\(^1\).

**Claim**: To solve the distinct elements problem \( (F_0) \) exactly we need at least \( \min(\{m \log u, u\}) \) space.

\(^1\)We will follow Jelani Nelson’s exposition (FM, FM+, FM++).
\[
\mathbb{E} \left[ \min(X_1, \ldots, X_n) \right] = \frac{1}{n+1}
\]

- \(X_i \in U[0, 1]\) for \(i \in [n]\)
- \(Z = \min(X_1, \ldots, X_n)\)

\[
\mathbb{E}[Z] = \int_0^1 \Pr[Z > t]dt = \int_0^1 \Pr[X_1 > t]^n
\]

\[
= \int_0^1 (1 - t)^n dt = \frac{1}{n+1}.
\]

A slick proof follows ...
\[ \mathbb{E} \left[ \min(X_1, \ldots, X_n) \right] = \frac{1}{n+1} \]

- \( X_{n+1} \in U[0, 1] \)

- What is \( \Pr \left[ X_{n+1} < \min(X_1, \ldots, X_n) \right] \) equal to?

1. By symmetry to \( \frac{1}{n+1} \)

2. On the other hand by definition of uniform distribution, it is equal to \( \mathbb{E} \left[ \min(X_1, \ldots, X_n) \right] \).

QED
Suppose that we have access to a random hash function $h : [u] \rightarrow [0, 1]$.

**FM method** (Flajolet-Martin)

- We initialize $X \leftarrow +\infty$.
- When $x_i$ arrives, we use $h$ to hash it to $h(x)$
- If $h(x) < X$ we set $X \leftarrow h(x)$
- At the end of the stream, $X = \min_{x \in \text{stream}} h(x)$
- Output $1/X - 1$
• To store $h$ we need $\Omega(u)$ space
• Floating-Point Arithmetic
**Idea:** Average together multiple estimates from the idealized algorithm FM.

1. Instantiate \( q = \frac{1}{\epsilon^2 \eta} \) FMs independently.

2. Let \( X_i \) come from \( \text{FM}_i \).

3. Output \( 1/Z - 1 \), where \( Z = \frac{1}{q} \sum_i X_i \).

To analyze \( \text{FM}^+ \) we need to upper bound the variance of each \( X_i \), and apply Chebyshev’s inequality.
\[ \mathbb{E}[X^2] = \int_0^1 \text{Pr}[X^2 > t] dt = \int_0^1 (\text{Pr}[X_1^2 > t])^n dt = \ldots \]
\[ = \frac{2}{(n + 1)(n + 2)}. \]

Therefore, the variance \( \text{Var}[X] \) is equal to

\[ \text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{n}{(n + 1)^2(n + 2)} < \frac{1}{(n + 1)^2} \]
Theorem

\[ \Pr \left[ \left| Z - \frac{1}{n+1} \right| > \frac{\epsilon}{n+1} \right] < \eta. \]

Proof.

We apply Chebyshev’s inequality

\[ P\left( \left| Z - \frac{1}{n+1} \right| > \frac{\epsilon}{n+1} \right) < \frac{(n+1)^2}{\epsilon^2} \frac{1}{q(n+1)^2} = \eta \]

Notice that we care about the concentration of \( \frac{1}{Z} \), not \( Z \).
**Theorem**

\[ \Pr \left[ \left| \left( \frac{1}{2} - 1 \right) - n \right| > O(\epsilon)n \right] < \eta \]

**Proof sketch:** We use Taylor expansion as follows:

\[
\frac{1}{(1 \pm \epsilon)^{\frac{1}{n+1}}} - 1 = (1 \pm O(\epsilon))(n + 1) - 1 = (1 \pm O(\epsilon))n \pm O(\epsilon)
\]
Median Boosting Trick: FM++

We use FM+ as a blackbox. We take multiple estimates of it, and we take the median.

1. Instantiate \( s = \lceil 36 \ln(2/\delta) \rceil \) independent copies of FM+ with \( \eta = 1/3 \).

2. Output the median \( \hat{n} \) of \( \{1/Z_j - 1\}_{j=1}^{s} \) where \( Z_j \) is from the \( j \)th copy of FM+.
**Theorem:** \( \Pr[|\hat{n} - n| > \epsilon n] < \delta. \)

**Proof:**

Let

\[
Y_j = \begin{cases} 
1 & \text{if } |(1/Z_j - 1) - n| > \epsilon n \\
0 & \text{else}
\end{cases}
\]

using Chernoff

\[
\Pr \left[ \sum Y_j > s/2 \right] = \Pr \left[ \sum Y_j - s/3 > s/6 \right] = \Pr \left[ \sum Y_j - \mathbb{E} \sum Y_j > \frac{1}{2} \mathbb{E} \sum Y_j \right] < e^{-\frac{(\frac{1}{2})^2 s/3}{3}} < \delta
\]
Implementation

- We show a constant approximation in $O(\lg u)$ bits, our estimate $\tilde{n}$ satisfies

$$\frac{n}{C} \leq \tilde{n} \leq Cn.$$

Algorithm

1. Pick $h$ from 2-wise family $[u] \rightarrow [u]$, for $u$ a power of 2 (round up if necessary)

2. Maintain $X = \max_{x \in \text{stream}} \text{lsb}(h(x))$ where $\text{lsb}$ is the least significant bit of a number

3. Output $2^X$
2-wise independent family

**Reminder from Lecture 4:** We can construct a 2-wise independent family as follows.

- $p$ is prime
- $a \neq 0$, $b$ chosen uar from $[p]$
- The hash of $x$ is

\[ h(x) = ax + b \mod p, \]
Implementation Analysis

- For fixed $j$, let $Z_j$ be the number of $i$ in stream with $\text{lsb}(h(i)) = j$.
- Define
  
  $Y_x = \begin{cases} 
  1 & \text{lsb}(h(x)) = j \\
  0 & \text{else}
  \end{cases}$

- $Z_j = \sum_{x \in \text{stream}} Y_x$
- $\mathbb{E}[Z_j] = 2^{-(j+1)}$ (why?)
- $\mathbb{V}[Z_j] = \sum_{x \in \text{stream}} \mathbb{V}[Y_x] < \frac{n}{2^{j+1}}$ (pairwise independence $\Rightarrow$ no covariance)
Implementation Analysis

- Let $Z_j$ be the number of $i$ with $\text{lsb}(h(i)) > j$.
- For $j^* = \lceil \log n - 5 \rceil$, we have

  $$16 \leq \mathbb{E}Z_{j^*} \leq 32$$

  $$P(Z_{j^*} = 0) \leq P(|Z_{j^*} - \mathbb{E}Z_{j^*}| \geq 16) < 1/5$$

  by Chebyshev.

- For $j = \lceil \log n + 5 \rceil$

  $$\mathbb{E}Z_{j} \leq 1/16$$

  $$P(Z_{j} \geq 1) < 1/16$$

  by Markov.
Readings

- Lecture 2, CS 229r: Algorithms for Big Data (Harvard, Jelani Nelson)

Additional readings

- “An Optimal Algorithm for the Distinct Elements Problem” by Kane, Nelson, Woodruff
  a Uses $O\left(\frac{1}{\epsilon^2} + \log n\right)$ bits space complexity,
  b and provides update, and query times $O(1)$