Streaming Algorithms

part 2
Last time

Distinct elements

Input: A stream $s_1, ..., s_n \in \{\emptyset, ..., N\}$

Goal: Estimate number of distinct elements in stream
Algorithm

- Pick a random hash function $h : \{1, \ldots, N\} \rightarrow [0, 1]$
- Compute minimum of $h(s_1), \ldots, h(s_n)$
  
  \[ \text{minimum of } r_1, \ldots, r_n = a \]

- Output $1/a$

Intuition: $\frac{1}{a} \approx k + 1$

Todo: How to construct $h$?
Problems with random $h : \mathbb{E}_1 \ldots, N^3 \to [0,1]$

1. Computers can't store arbitrary real numbers
   \[ \text{Soln: Pick } h : \mathbb{E}_1 \ldots, N^3 \to \mathbb{E}_1, \ldots, R^3, \text{ } R \text{ is large} \]
   \[ \text{So } h(i)/R \sim \text{ random number in } [0,1] \]

2. If $h : \mathbb{E}_1 \ldots, N^3 \to \mathbb{E}_1, \ldots, R^3$ is uniformly random
   \[ \text{needs } N \log R \text{ bits to store} \]
   \[ \text{Soln: Make } h \text{ "pseudorandom"} \]
   A hash family is a set $H = \{ h_1, \ldots, h_m \}$
   \[ \text{Write } h \sim H \text{ to mean random } h \]

1. $h \sim H$ looks "somewhat random"
2. $m$ is small $\Rightarrow \log(m) \text{ bits to store}$
A hash family \( H = \{ h_1, \ldots, h_m : \{1, \ldots, N\} \rightarrow \{1, \ldots, R3\} \} \) is pairwise independent if

- for all \( x \neq y \in \{1, \ldots, N\} \):
  \[
  \Pr \left[ h(x) = i \text{ and } h(y) = j \right] = \frac{1}{R^2}
  \]
  \( i, j \in \{1, \ldots, R3\} \) \( h \sim H \)

Look like two independent draws from \( \mathbb{E} \{1, \ldots, R3\} \)

Implies:

\[
\Pr \left[ h(x) = i \right] = \frac{1}{R}
\]

\( h \sim H \)
Example
Let $p$ be a prime
For each $a, b, c \in \mathbb{Z}_p = \{0, 1, \ldots, p-1\}$
let $h_{a, b}: \mathbb{Z}_p \to \mathbb{Z}_p$

$$h_{a, b}(x) = ax + b \pmod{p}$$

Then $H = \{ h_{a, b} \mid a, b \in \mathbb{Z}_p \}$ is pairwise independent
Let \( x \neq y \) and \( i, j \) (all in \( \mathbb{Z}_p \))

Goal: \( \Pr[ax+b=i \text{ and } ay+b=j] = \frac{1}{p^2} \)

Suppose \( x = 0 \) and \( y = 1 \) (for simplicity)

Goal: \( \Pr[b=i \text{ and } a+b=j] = \frac{1}{p^2} \)

But \((b, at+b)\) is random pair in \( \mathbb{Z}_p \)

(General case left as exercise)

\( x=0 \ y=1 \ z=2 \ f(z)=2a+b = 2(a+b)-b \)

= \( 2f(1)-f(0) \)
Algorithm (modified)

1. Pick a pairwise independent hash function
   \[ h: \mathbb{Z}_N \rightarrow [0,1] \]

2. Compute \( \alpha = \text{smallest of } h(s_1), \ldots, h(s_n) \)
   \[ = t\text{-th smallest of } r_1, \ldots, r_k \]

3. Outputs \( \frac{1}{d} \cdot \frac{1}{t} \)
   (assuming \( k \) distinct ekms)
   (should be \( \approx \frac{1}{k \cdot t} \))

Algorithm susceptible to outliers
one abnormally small \( r_i \) can ruin output

Idea: use \( t\text{-th smallest } r_i \)

Alg should store \( t \) smallest \( r_i \)'s
and corresponding \( S_j \)'s
Analysis: Suppose $k = \# \text{ of distinct elements}$

$\Pr[\text{any outputs } \geq 2k] = \Pr[\alpha \leq \frac{t}{2k}]$

Define $C_i = \begin{cases} 1 & \text{if } r_i \leq \frac{t}{2k} \\ 0 & \text{otherwise} \end{cases}$

Then $C = C_1 + \ldots + C_k$

$E[C] = E\left[ \sum_{i=1}^{k} C_i \right] = \sum_{i=1}^{k} E[C_i]$ (linearity of expectation)

$= \sum_{i=1}^{k} \Pr[ r_i \leq \frac{t}{2k}]$

$= \sum_{i=1}^{k} \frac{t}{2k} = k \cdot \frac{t}{2k} = \frac{t}{2}$
Recall: \( \text{Var}[X] = E[X^2] - E[X]^2 \) (\( \leq E[X^2] \))

Fact: If \( X_1, \ldots, X_n \) are independent, then
\[
\text{Var}(X_1 + \ldots + X_n) = \text{Var}(X_1) + \ldots + \text{Var}(X_n)
\]
Also holds if \( X_1, \ldots, X_n \) are pairwise independent

\[
\text{Var}[C] = \text{Var}\left[ \frac{X_i}{\sum_{i=1}^n C_i} \right] = \sum_{i=1}^n \text{Var}[C_i]
\]

\[
\text{Var}[C_i] \leq E[C_i^2] = E[C_i] = \frac{t}{2k}
\]

\[
\text{Var}[C] \leq k \cdot \frac{t}{2k} = \frac{t}{2} \text{ Standard Dev} \leq \sqrt{\frac{t}{2}}
\]

\[
\frac{t}{2} - O(\sqrt{tE}) \quad \frac{t}{2} + O(\sqrt{tE})
\]

\[
T \quad \exists C \text{ is large}
\]
Heavy hitters

Input: a stream $S_1, \ldots, S_n \in \{1, \ldots, N\}$

Output: each $a \in \{1, \ldots, N\}$ whose frequency

$$f_a = \# i \text{ s.t. } S_i = a$$

is large

i.e. a subset $L \subseteq \{1, \ldots, N\}$ s.t.

1. every $a$ s.t. $f_a \geq \frac{n}{10}$ is in $L$

2. no $a$ s.t. $f_a \leq \frac{n}{20}$ is in $L$
Count-Min-Sketch \((l, B)\)

- Initialize \(l \times B\) array \(M\) to all zeros
- Pick \(l\) pairwise independent hash functions \(h_i, \ldots, h_{l-1}: \{1, \ldots, N\} \rightarrow \{1, \ldots, B\}\)
- While stream is not empty
  - Read \(s\), next stream element
  - For \(i = 1 \ldots l\)
    - \(M[i, h_i(s)]++\)
  - If min of these vals is \(\geq \frac{n}{10}\), add \(s\) to \(L\)
- Return \(L\)
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>...</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>$h_1(s)$</td>
<td></td>
<td>$M$</td>
</tr>
<tr>
<td>$h_2$</td>
<td>$h_2(s)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_k$</td>
<td>$h_k(s)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fact: For each symbol $a$, $M[i, h_i(a)] > f_a$
Fix an element $a$

$$M[i, h_i(a)] = f_a + \sum_{b \neq a} f_b$$

$$\because h_i(b) = h_i(a)$$

$$E[M[j, h_j(a)]] = f_a + \sum_{b \neq a} \Pr[h_i(b) = h_i(a)] \cdot f_b$$

$$= f_a + \sum_{b \neq a} f_b \cdot \frac{1}{B} \leq f_a + \frac{n}{B}$$

So if $X = M[i, h_i(a)]$, then:

- $X \geq 0$  
- $f_a$
- $E[X] \leq \frac{n}{B}$
Markov's inequality: \( \text{Pr}[X \geq t \cdot \mathbb{E}(X)] \leq \frac{1}{t} \) (if \( X \geq 0 \))

So \( \text{Pr}[X \geq 2 \cdot \frac{n}{3}] \leq \frac{1}{2} \)

\( \text{Pr}[M(c, h;i(a)) \geq f_a + 2 \cdot \frac{n}{3}] \leq \frac{1}{2} \)

Then \( \text{Pr} \left( \bigvee_{i} M(i, h;i(a)) \geq f_a + 2 \cdot \frac{n}{3} \right) \leq \frac{1}{2}e \)

\[ \geq f_a + \frac{n}{20} \right) \leq \frac{1}{n^2} \]

for \( B=40, \ e = 2 \log(n) \)

If \( f_a \leq \frac{n}{20} \), it is included in \( L \) w/prob \( \leq \frac{1}{n^2} \)

Only \( n \) possible "bad" \( a \)'s, so \( \text{Pr}(\text{one gets into } L) \leq \frac{1}{n} \) (union bound)