Streaming Algorithms

part 2

Last time

Distinct elements

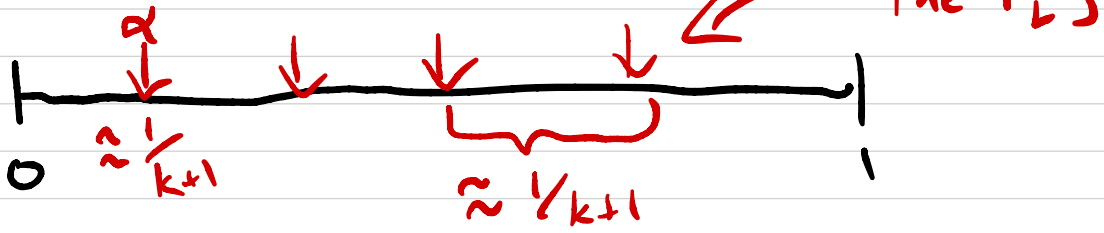
Input: A stream $s_1, \ldots, s_n \in \{1, \ldots, N\}$

Goal: Estimate number of distinct elements in stream

Algorithm

- Pick a random hash function $h: \{1, \ldots, N\} \to [0,1]$
- Compute minimum of $h(s_1), \ldots, h(s_n)$
  $=$ minimum of $r_1, \ldots, r_k = \alpha$
  (assuming k distinct elements)
- Output $1/\alpha$

Intuition:



Todo: How to construct h?

Problems with random $h: \{1,\ldots,N\} \to [0,1]$

1. Computers can't store arbitrary real numbers

Soln: Pick $h: \{1,\ldots,N\} \to \{1,\ldots,R\}$, R is large
So $h(i)/R \simeq$ random number in $[0,1]$

2. If $h: \{1,\ldots,N\} \to \{1,\ldots,R\}$ is uniformly random
needs $N \log R$ bits to store

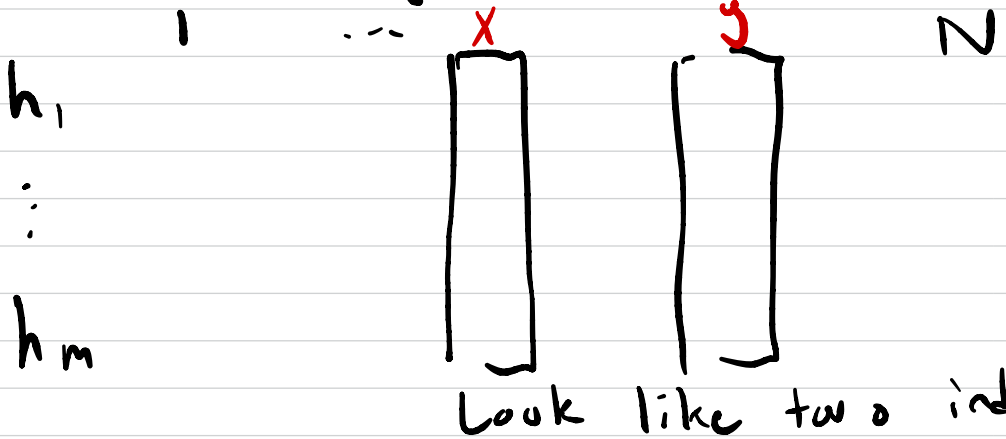
Soln: Make h "pseudorandom"
A **hash family** is a set $\mathcal{H} = \{h_1,\ldots,h_m\}$
Write $h \sim \mathcal{H}$ to mean random h_i

1. $h \sim \mathcal{H}$ looks "somewhat random"
2. m is small $\to$ $\log(m)$ bits to store

A hash family $\mathcal{H} = \{h_1, \ldots, h_m : \{1, \ldots, N\} \to \{1, \ldots, R\}\}$

is pairwise independent if

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



Look like two independent draws from $\{1, \ldots, R\}$

Implies: $\Pr_{h \sim \mathcal{H}}[h(x) = i] = \frac{1}{R}$

Example

Let p be a prime

For each $a, b \in \mathbb{Z}_p = \{0, 1, \ldots, p-1\}$

let $h_{a,b} : \mathbb{Z}_p \to \mathbb{Z}_p$

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

Then $\mathcal{H} = \{h_{a,b}\}_{a,b \in \mathbb{Z}_p}$ is pairwise independent

Pf: Let $x \neq y$ and i, j (all in $\mathbb{Z}_p$)

Goal: $\Pr_{a,b}[ax+b = i \text{ and } ay+b = j] = \frac{1}{p^2}$

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

Goal: $\Pr_{a,b}[b = i \text{ and } a+b = j] = \frac{1}{p^2}$

But $(b, a+b)$ 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)

- Pick a pairwise independent hash function
  $h: \{1, \ldots, N\} \to [0, 1]$
- Compute $\alpha =$ smallest of $h(s_1), \ldots, h(s_n)$
  $=$ t-th smallest of $r_1, \ldots, r_k$
  (assuming k distinct elems)
  (should be $\approx 1/k \cdot t$)
- Outputs $1/\alpha \cdot t$

Algorithm susceptible to outliers
one abnormally small r_i can ruin output

Idea: use t-th smallest r_i

Alg should store t smallest r_i's
and corresponding s_j's

Analysis: Suppose $k = \#$ of distinct elements

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

$$= \Pr\left[\underbrace{\left(\#\, i : r_i \leq \frac{t}{2k}\right)}_{C} \geq t\right]$$

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

Then $C = C_1 + \cdots + C_k$

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

$$= \sum_i \Pr\left[r_i \leq \frac{t}{2k}\right]$$

$$= \sum_i \frac{t}{2k} = k \cdot \frac{t}{2k} = \frac{t}{2}$$

Recall: $\mathrm{Var}[X] = E[X^2] - E[X]^2$ $(\leq E[X^2])$

Fact: If $X_1, \ldots, X_n$ are independent, then

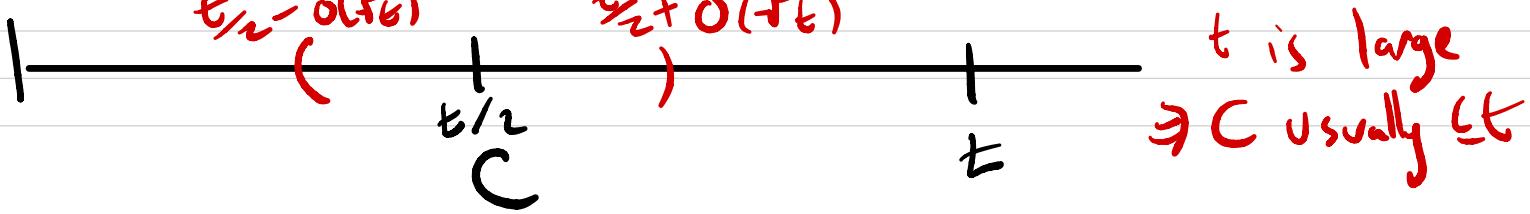
$$\mathrm{Var}[X_1 + \ldots + X_n] = \mathrm{Var}[X_1] + \ldots + \mathrm{Var}[X_n]$$

Also holds if $X_1, \ldots, X_n$ are pairwise independent

$$\mathrm{Var}[C] = \mathrm{Var}\left[\sum_{i=1}^{k} C_i\right] = \sum_{i=1}^{k} \mathrm{Var}[C_i]$$

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

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



t is large
$\Rightarrow$ C usually $\approx t$

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

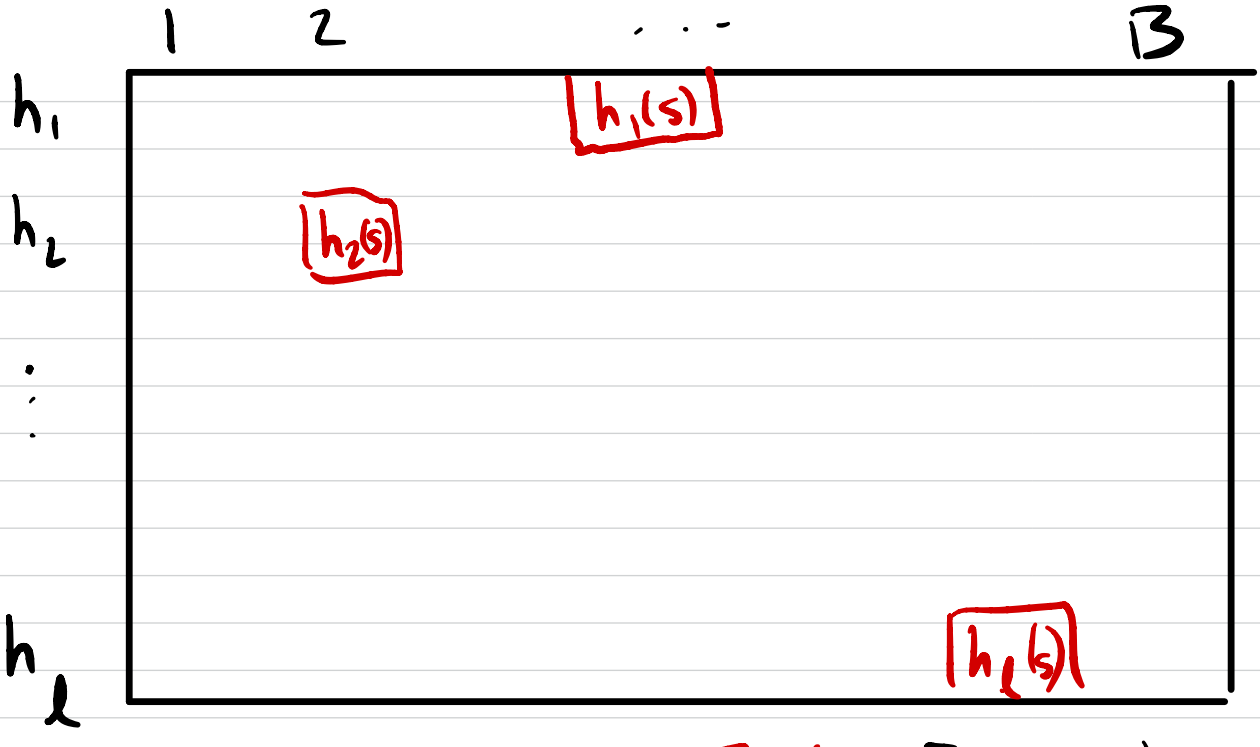
i.e. a subset $L \subseteq \{1, \ldots, N\}$ s.t.

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 $\ell \times B$ array M to all zeros
- Pick ℓ pairwise independent hash functions
  $h_1, \ldots, h_\ell : \{1, \ldots, N\} \to \{1, \ldots, B\}$
- While stream is not empty
  - Read s, next stream element
  - For $i = 1 \ldots \ell$
    $M[i, h_i(s)]$ ++
  - If min of these vals is $\geq \frac{n}{10}$, add s to L
- Return L



→ stream element s

Fact: For each symbol a

$$M[i, h_i(a)] \geq f_a$$

Fix an element a

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

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

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

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

- $X \geq 0$
- $E[X] \leq \frac{n}{B}$

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

So $\Pr[X \geq 2 \cdot \frac{n}{B}] \leq 1/2$

$\Pr[M[i, h_i(a)] \geq f_a + 2 \cdot \frac{n}{B}] \leq \frac{1}{2}$

Then $\Pr[\forall_i\ M(i, h_i(a)] \geq f_a + 2 \cdot \frac{n}{B}) \leq \frac{1}{2^\ell}$

$\geq f_a + \frac{n}{20}) \leq \frac{1}{n^2}$

for $B = 40$, $\ell = 2\log(n)$

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 $\Pr[\text{one gets into } L] \leq 1/n$ (union bound)