

Streaming Algorithms

part 2

Last time

Distinct elements

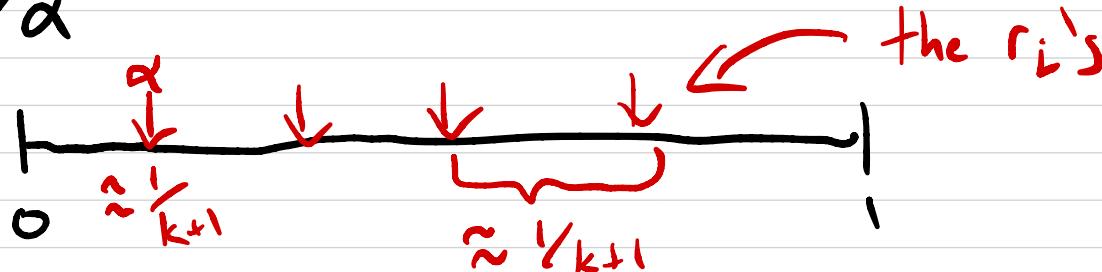
Input: A stream $s_1, \dots, s_n \in \{1, \dots, N\}$

Goal: Estimate number of distinct
elements in stream

Algorithm

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

Intuition:



To do: How to construct h^2 .

Problems with random $h: \{1, \dots, N\} \rightarrow [0, 1]$

1. Computers can't store arbitrary real numbers

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

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

Solⁿ: Make h "pseudorandom"

A **hash family** is a set $\mathcal{H} = \{h_1, \dots, h_m\}$

Write $h \sim \mathcal{H}$ to mean random h_i

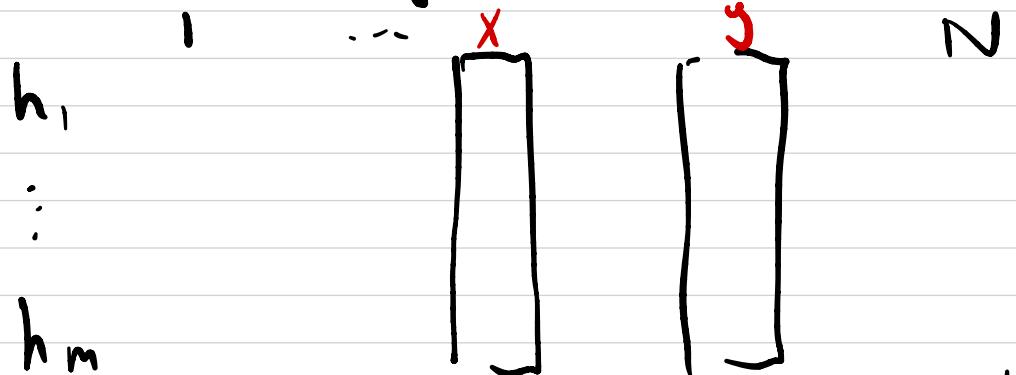
1. $h \sim \mathcal{H}$ looks "somewhat random"

2. m is small $\rightarrow \log(m)$ bits to store

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

is pairwise independent if

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



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

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

Example

Let p be a prime

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

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

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

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

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

(General case left as exercise)

$$\begin{aligned} (x=0 \quad y=1 \quad z=2 \quad f(z) = 2a + b = 2(a+b) - b \\ = 2f(1) - f(0)) \end{aligned}$$

Algorithm (modified)

- Pick a pairwise independent hash function

$$h: \{1, \dots, N\} \rightarrow [0, 1]$$

- Compute $\alpha = t\text{-th smallest of } h(s_1), \dots, h(s_n)$

$$= t\text{-th smallest of } r_1, \dots, r_k$$

- Outputs $1/d \cdot t$ (assuming k distinct elems)

$$\text{(should be } \approx \frac{1}{k} \cdot t \text{)}$$

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

$$\text{Then } C = C_1 + \dots + C_k$$

$$\mathbb{E}[C] = \mathbb{E}\left[\sum_{i=1}^k C_i\right] = \sum_{i=1}^k \mathbb{E}[C_i] \quad (\text{linearity of expectation})$$
$$= \sum_i \Pr[r_i \leq \frac{t}{2k}]$$
$$= \sum_i \frac{t}{2k} = k \cdot \frac{t}{2k} = \frac{t}{2}$$

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

Fact: If X_1, \dots, X_n are independent, then

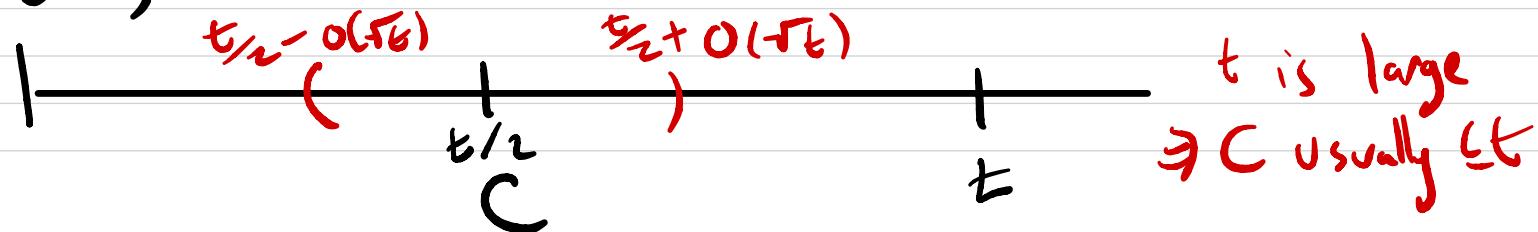
$$\text{Var}[X_1 + \dots + X_n] = \text{Var}[X_1] + \dots + \text{Var}[X_n]$$

Also holds if X_1, \dots, X_n are pairwise independent

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

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

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



Heavy hitters

Input: a stream $s_1, \dots, s_n \in \{1, \dots, N\}$

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

$$f_a = \#\{i \text{ s.t. } s_i = a\} \text{ is large}$$

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

1. every a s.t. $f_a > \frac{n}{10}$ 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
 $h_1, \dots, h_l : \{1, \dots, N\} \rightarrow \{1, \dots, B\}$
- While stream is not empty
 - Read s , next stream element
 - For $i = 1 \dots l$
 $M[i, h_i(s)]++$
 - If min of these vals is $\geq \frac{n}{10}$, add s to L
- Return L



Fact: For each symbol a
 \rightarrow stream element s

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

Fix an element a

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

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

$$\begin{aligned} E_{h_i} M[i, h_i(a)] &= f_a + \sum_{b \neq a} \Pr[h_i(b) = h_i(a)] \cdot f_b \\ &= f_a + \sum_{b \neq a} f_b \cdot \frac{1}{B} \leq f_a + \frac{n}{B} \end{aligned}$$

So if $X = M[i, h_0(a)]$, then:

$$-f_a$$

$$\begin{aligned} &\bullet X \geq \delta \\ &\bullet E[X] \leq \frac{n}{B} \end{aligned}$$

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

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

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

$$\text{Then } \Pr[\bigcup_i M(i, h_i(a)) \geq f_a + 2 \cdot \frac{n}{B}] \leq \frac{1}{2} \ell$$
$$\geq f_a + \frac{\Delta_{20}}{20} \leq \frac{1}{n^2}$$

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

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)