DISTINCT ELEMENTS

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

GOAL: Estimate # of distinct elements in the stream.

$\{0, 1/2^k, 2/2^k, 3/2^k \ldots 1\}$

1) Pick a (pairwise independent) random hash function $h: \{1 .. N\} \to [0, 1]$

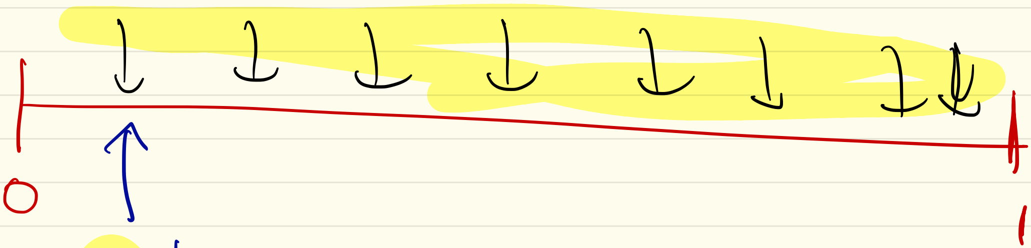
2) Compute α = minimum of $\{h(s_1) \ldots h(s_n)\}$

implement in small space

3) Output $\underbrace{1/\alpha}_{k+1} - 1$

$\{s_1 \ldots s_n\} \rightarrow$ K distinct elements in stream.

$\{h(s_1) \ldots h(s_n)\} \rightarrow$ K distinct real #s in $[0,1]$



$\approx 1/K$

$$\min\{h(s_1), h(s_2) \ldots h(s_n)\}$$
$$\approx 1/K$$

Thm:

$$\mathbb{E}_h \left[\min \{ h(s_1) \dots h(s_n) \} \right] = \frac{1}{K+1}$$

where K = # of distinct elements in $\{s_1 \dots s_n\}$

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

random!!

Proof:

Lemma: If $r_1 \dots r_K \in [0,1]$ uniformly random

the $\mathbb{E}[\min(r_1, \dots r_K)] = 1/_{K+1}$

Truly random hash function $h : \{1..N\} \to [0,1]$

Let $\underbrace{s_1 \ldots s_k}_{s_{i_1}, s_{i_2} \ldots s_{i_k}}$ be distinct elements in the stream.

1) $$\Pr_h\left[\min(h(s_1), \ldots h(s_k)) \leq \frac{1}{4k}\right] \leq \frac{1}{4}.$$

$\Updownarrow$

estimate $\geq 4k$

Proof: $$= \Pr_h\left[\left(h(s_1) < \frac{1}{4k}\right) \vee \left(h(s_2) < \frac{1}{4k}\right) \ldots \vee \left(h(s_k) < \frac{1}{4k}\right)\right]$$

$$\leq \sum_{i=1}^{k} \Pr\left[h(s_i) < \frac{1}{4k}\right] \quad \text{(union bound)}$$

$$= \sum_{i=1}^{k} \left(\frac{1}{4k}\right) = \frac{1}{4}$$

2) $\Pr\left[\min\left(h(s_1), \dots, h(s_n)\right) > 4/k\right] \leq e^{-4}$

$\Updownarrow$

estimate < $k/4$

truly random function.

$$= \Pr\left[\left(h(s_1) > 4/k\right) \wedge \left(h(s_2) > 4/k\right) \dots \wedge \left(h(s_k) > 4/k\right)\right]$$

(independence)

$$= \prod_{i=1}^{k} \Pr\left[h(s_i) > 4/k\right]$$

(truly random hash function)

$$= \left(1 - 4/k\right)^{k}$$

$$\approx e^{-4}$$

hash family

$$\mathcal{H} = \{ h_1 \ldots h_M : \{1..N\} \to [R] \}$$

$\mathcal{H}$ = set of all possible functions

To remember a $h \in \mathcal{H}$ need $\log |\mathcal{H}|$ bits.

$\mathcal{H}$ must be small yet somehow random

Pairwise Independent Hash family

A hash family $\mathcal{H}$ is pairwise independent if

$\forall\, x, y \quad x \neq y \in$ Domain $\{1 .. N\}$

$\forall\, \alpha, \beta \in$ Range $\{1 .. R\}$

$$\Pr_{h \sim \mathcal{H}}\Big[(h(x) = \alpha) \wedge (h(y) = \beta)\Big] = \Pr \text{ under a completely random function}$$

$$= \frac{1}{R} \cdot \frac{1}{R} = \frac{1}{R^2}$$

Example:

p - prime

$$\mathbb{Z}_p = \{0, 1 \dots p-1\}$$

integers modulo p

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

$a, b \in \mathbb{Z}_p$

$$|\mathcal{H}| = p^2$$

$h_{a,b}(1)$

$h_{a,b}(2)$

$\Downarrow$

a, b