Estimating Distinct Elements

$(1+\varepsilon)$ approx w. prob. $(1-\delta)$

Space $O(\frac{1}{\varepsilon^2}\log(\frac{1}{\delta}))$

$h: U \longrightarrow [0,1]$ $\quad Y = \min h(x) \quad E[Y] = \frac{1}{K+1}$

$O(\log(\frac{1}{\delta}))$ groups of $O(\frac{1}{\varepsilon^2})$ hash fns each

"median of means"

Sketching : "sketch" of data stream

Composable $\quad sk(S_1) \quad sk(S_2) \quad sk(S_1 \cup S_2)$

Assumption: hash fn is completely random

$h(x) \in_R [0,1]$

$h(x_1)\ h(x_2)\ \ldots\ h(x_k)$ independent

Pairwise Independent Hash fn

H: family of hash fn $\quad h \in_R H$

$$P_h[h(x_1) = y_1,\ h(x_2) = y_2] = P_h[h(x_1) = y_1] \cdot P_h[h(x_2) = y_2]$$

p: large prime $\quad x \in [p]$

$h_{a,b} : [p] \longrightarrow [p]$

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

$H = \{h_{a,b},\ a,b \in [p]\}$

$Y = \min\{\frac{h(x)}{p}\} \quad \frac{h(x)}{p}$

$P[Y < \frac{1}{3k}] < \frac{2}{5} \longrightarrow \in [0, \frac{1}{3k}] \quad k \times \frac{1}{3k} = \frac{1}{3}$

$P[Y > \frac{3}{k}] < \frac{1}{3} \longrightarrow$ HW constant $\nearrow O(\log(\frac{1}{\delta}))$ copies of hash f^n & take median

then this is an estimate within $[\frac{1}{3}, 3]$ of $\frac{1}{K}$

$a_i, b_i : h_i \qquad a_i, b_i \in_R [p]$

$(1+\varepsilon)$ approx using pairwise independence

[Bar Yossef et al, 2002]

Change algo: track smallest t hash elements

y_i : i^{th} smallest element $\quad E[y_i] = \frac{i}{K+1}$

Estimator: $\frac{t}{y_t} \approx K$

<u>Thm</u> $\quad t = \frac{c}{\varepsilon^2} \quad$ with prob $\geq \frac{2}{3} \quad \frac{(1-\varepsilon)t}{K} \leq y_t \leq \frac{(1+\varepsilon)t}{K}$

<u>Pf</u> (2nd ineq)

$I = [0, (1+\varepsilon)\frac{t}{K}]$

X_i : indicator for $h(x_i) \in I$

$X = \sum X_i \qquad$ # hash values in I

$\mathbb{E}[X] = \sum \mathbb{E}[X_i] = K \cdot \frac{(1+\varepsilon)t}{K} = (1+\varepsilon)t$

$$\begin{aligned} \Pr[y_t > (1+\varepsilon)\tfrac{t}{K}] &= \Pr[X < t] \\ &= \Pr[X - \mathbb{E}[X] < -\varepsilon t] \\ &\leq \Pr[|X - \mathbb{E}[X]| > \varepsilon t] \end{aligned}$$

Chebyshev: $\Pr[|X - \mathbb{E}[X]| > \varepsilon t] \leq \frac{Var(X)}{\varepsilon^2 t^2}$

$p = \Pr[X_i = 1] = \frac{(1+\varepsilon)t}{K}$

$\mathbb{E}[X_i] = p \quad Var(X_i) = p(1-p)$

$\mathbb{E}[X] = kp \quad \underbrace{Var(X) = kp(1-p)}_{\text{uses pairwise independence}} \leq \mathbb{E}[X] = (1+\varepsilon)t$

$$\Pr[|X - \mathbb{E}[X]| > \varepsilon t] \leq \frac{(1+\varepsilon)t}{\varepsilon^2 t^2} = \frac{(1+\varepsilon)}{c} \leq \frac{1}{6}$$

by suitable choice of c

$\blacksquare$

$O(\log(\frac{1}{\delta}))$ copies

Store $t = O(\frac{1}{\varepsilon^2})$ hash values

$(1+\varepsilon)$ approximation w. prob. $1-\delta$

Practical algo: HyperLogLog [Flajolet et al 2007]

Assume: hash f^n is completely random

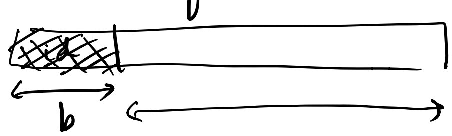
estimate cardinalities beyond 10^9

w. accuracy 2% using ~ 1.5 Kbytes

Stochastic averaging: [Flajolet, Martin]

maintain m random variables $m = 2^b$

break up stream into m substreams by using first b bits of hash value



$0^{t-1}1 \qquad \approx 2^t$

Each substream tracks max pos^n of leading 1

$M(i)$: pos^n of leading 1

estimate $2^{m(i)}$

Harmonic mean of these estimates

$\sum_i 2^{-m(i)}$ $\qquad \frac{K}{m}$

$$2^{-m(i)} \approx \frac{m}{K}$$

$$\sum_i 2^{-m(i)} \approx \frac{m^2}{K}$$

$$\text{Estimator} = \frac{m^2 \alpha_m}{\sum_i 2^{-m(i)}}$$

$$E[\text{Estimator}] \approx K$$

$$\frac{\sqrt{\mathrm{Var}[\text{Estimator}]}}{K} \approx \frac{1.04}{\sqrt{m}}$$

Lower Bounds for Streaming Algorithms

Any deterministic algorithm that gives 1.4 approximation to # distinct elements must use $\Omega(n)$ memory.

If we have inputs $I_1 \ldots I_N$ for which algo must have distinct states, then $\Omega(\log N)$ space.

$I_i \quad I_j \quad \exists$ input I'

$I_i \cup I'$ vs. $I_j \cup I'$ have very different answers.

$\{S_i\}_{i=1}^{N}$ subsets of $[n]$

$$\forall i \ |S_i| = \frac{n}{10}$$

$$\forall i \neq j \quad |S_i \cap S_j| \leq \frac{n}{20}$$

$S_i \quad S_j \qquad |S_i \cup S_i| = \frac{n}{10}$

$$|S_j \cup S_i| \geq \frac{3}{2} \cdot \frac{n}{10}$$

N different streams

$N = 2^{cn}$ many subsets. (Probabilistic method!)

$\log N = \Omega(n)$ lower bound.

Next Time: Frequency Moments.

f_i: # of times that i appears

$$F_t = \sum_i f_i^t$$

$$\boxed{F_2 = \sum_i f_i^2}$$

[Alon, Matias, Szegedy 1996]

Streaming Model

$$F_1 = \sum_i f_i$$