CS 368

Canvas

CS368.stanford.edu

Piazza, Gradescope

3 homeworks (60%)

Optional programming assignment, sub. for 1 HW

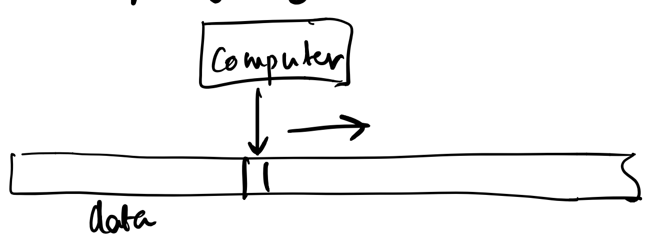
Project (35%)

Piazza participation (5%)

$\boxed{\text{Data Stream Model}}$

- Data does not fit in memory
- polynomial time not good enough!

Google query log.



<u>Warmup:</u> Counter that counts from 1 to n

$\log n$ bits

exact or deterministic $\longrightarrow$ $\log n$ bits needed!

K items : return $[K(1-\varepsilon), K(1+\varepsilon)]$

Allow failure probability δ (say $\delta < 10^{-6}$)

$O(\log \log n)$ bits suffice!

$\hookrightarrow O_{\varepsilon,\delta}$

Instead of k, keep track of $\log k$

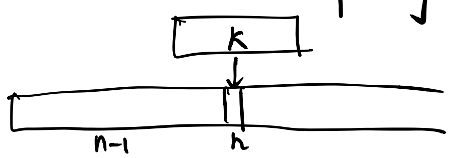
$$i \xrightarrow{\frac{1}{2^i}} i+1 \xrightarrow{\frac{1}{2^{i+1}}} i+2$$

Count Distinct Queries

(1) Hash Table : too large

(2) Random Sampling

Reservoir sampling



$\frac{k}{n}$

Suppose we pick random sample of k elements out of n

$$\underbrace{0\,0\,0\,0 \ldots 0}_{n}$$

$$\underbrace{0\,0\,0\,0\,0}_{n-\frac{n}{k}}\underbrace{1\,2 \ldots \tfrac{n}{k}}_{\frac{n}{k}}$$

$$\left(1-\frac{1}{k}\right) \qquad \frac{1}{k}$$

$$\left(1-\frac{1}{k}\right)^k \approx \frac{1}{e}$$

$$\frac{n}{k} \longrightarrow \frac{n}{100k}$$

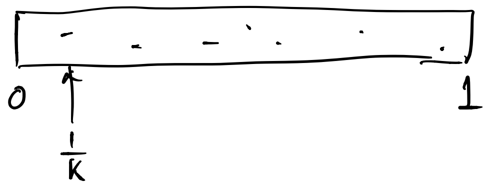
$$\left(1-\frac{1}{100k}\right)^k \approx \frac{1}{e^{1/100}} \approx 1-\frac{1}{100}$$

[Flajolet, Martin '85]

$h : U \longrightarrow [0,1]$

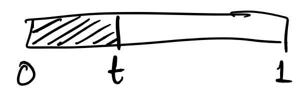
Assume $h(x) \in_R [0,1]$

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



<u>Claim</u> : If K distinct eles $x_1 \ldots x_K$

$$E\left[\min_{i=1..K} h(x_i)\right] = \frac{1}{K+1}$$



$$Y = \min(h(x_1) \ldots h(x_K))$$

$$\Pr[h(x_i) \le t] = t$$

$$\Pr[Y > t] = (1-t)^K$$

$$\Pr[Y \le t] = 1-(1-t)^K \qquad \text{CDF}$$

$$\Pr[Y \in [t, t+dt]] = K(1-t)^{K-1} dt \qquad \text{PDF}$$

$$E[Y] = \int_0^1 t \cdot K(1-t)^{K-1} dt \qquad\qquad E[Y] = \int_0^1 \Pr[Y>t]\, dt$$

$$= K \int_0^1 (1-(1-t))\,(1-t)^{K-1} dt$$

$$= K\left(\int_0^1 (1-t)^{K-1} dt - \int_0^1 (1-t)^K dt\right)$$

$$= K\left(\frac{1}{K} - \frac{1}{K+1}\right) = \frac{1}{K+1}$$

Is Y close to $E[Y]$?

$$Var[Y] = E[Y^2] - E[Y]^2$$

$$E[Y^2] = \int_0^1 t^2\, K(1-t)^{K-1} dt$$

$$= K \int_0^1 (1-(1-t))^2 (1-t)^{K-1} dt$$

$$= K\left[\int_0^1 (1-t)^{K-1} dt - 2\int_0^1 (1-t)^K + \int_0^1 (1-t)^{K+1} dt\right]$$

$$= k\left[\frac{1}{k} - \frac{2}{k+1} + \frac{1}{k+2}\right]$$

$$= k\left[\left(\frac{1}{k} - \frac{1}{k+1}\right) - \left(\frac{1}{k+1} - \frac{1}{k+2}\right)\right]$$

$$= k\left[\frac{1}{k(k+1)} - \frac{1}{(k+1)(k+2)}\right] = \frac{2}{(k+1)(k+2)} \leq 2\,E[Y]^2$$

$$\mathrm{Var}[Y] = E[Y^2] - E[Y]^2 \leq E[Y]^2$$

Chebyshev's inequality:

$$\Pr[|Y - E[Y]| > \varepsilon\, E[Y]] \leq \frac{\mathrm{Var}[Y]}{(\varepsilon . E[Y])^2} \leq \frac{1}{\varepsilon^2}$$

Take mean of multiple iid copies.

$Y_1 \ldots Y_t$ be independent copies of Y

$$Z = \frac{Y_1 + \cdots + Y_t}{t}$$

$$E[Z] = E[Y]$$

$$\mathrm{Var}[Z] = \frac{1}{t^2}\sum_1^t \mathrm{Var}[Y_i] = \frac{\mathrm{Var}[Y]}{t}$$

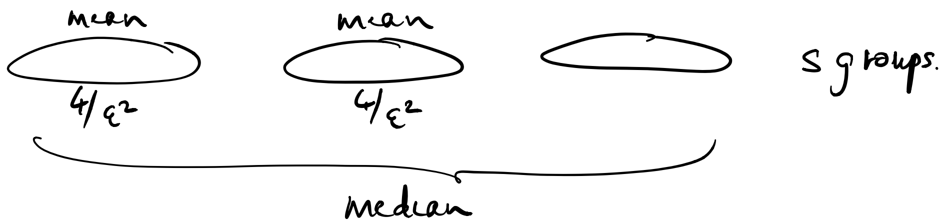
$$\mathrm{Var}[Z] \leq \frac{(E[Z])^2}{t}$$

$$\Pr[|Z - E[Z]| \geq \varepsilon . E[Z]) \leq \frac{\mathrm{Var}(Z)}{(\varepsilon . E[Z])^2} \leq \frac{1}{\varepsilon^2 t}$$

$t = \frac{4}{\varepsilon^2}$, failure probability $\leq \frac{1}{4}$

Median of Means

1. $Z_1 \ldots Z_s$ independent copies of Z
2. return median $(Z_1 \ldots Z_s)$



median bad $\Rightarrow$ at least half of $z_1 \dots z_s$ bad.

A_i indicator random variable $= \begin{cases} 1 & \text{if } z_i \text{ bad} \\ 0 & \text{otherwise} \end{cases}$

$E[A_i] \leq 1/4 \qquad A = \sum_{i=1}^{s} A_i$

$E[A] \leq s/4$

$$\begin{aligned} \Pr[\text{median}(z_1 \dots z_s) \text{ bad}] &\leq \Pr(\text{at least } \tfrac{s}{2} \text{ of } z_1 \dots z_s \text{ bad}) \\ &= \Pr[A \geq s/2] \\ &\leq e^{-(2\ln 2 - 1)s/4} \\ &\leq e^{-s/11} \end{aligned}$$

Set $s = 11 \ln(\frac{1}{\delta})$ to get failure prob. $< \delta$

<u>Chernoff Bound</u>: X sum of independent 0-1 random var.
$\mu = E[X]$

$$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$