Lec 12

Counting district elements in data streams

Elements arrive (a_i) from a domain $[m] = \{1..m\}$ $a_1 \ldots a_n$ n elements

Goal: Count # district items (F_0)

$$F_k = \sum_{i \in A} f_i^k$$ where f_i is the frequency of an item if A is the district set of items.

Result: with a small amount of memory we will approximate F_0 with high probability.

confidence parameter

$$\Pr[\,|F_0 - \tilde{F}_0| \leq \underset{\text{error parameter}}{\varepsilon}.F_0\,] \geq 1-\delta.$$

$M = m^3.$

Choose a random hash function $h : [m] \longrightarrow [M]$

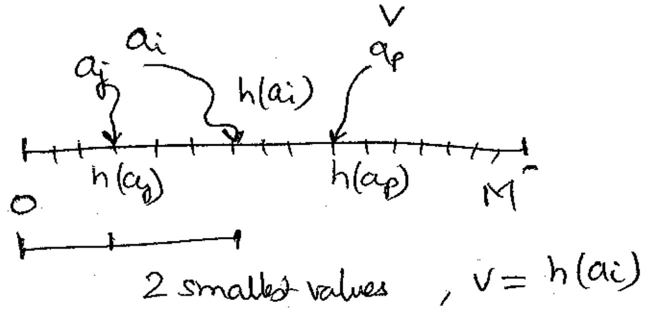
Note: ensures that the probability of a collision is very small $(\leq \frac{1}{m})$

Basic Idea: Let $t = \frac{c}{\varepsilon^2}$ $c \equiv$ constant. (TBD)

Apply $h(a_i)$, and maintain ~~min~~

$V \equiv$ max over the set of t smallest values in $\{h(a_i)\}$

Output $\tilde{F}_0 = \frac{tM}{V}$.



$$\tilde{F}_0 = \frac{t.M}{V}$$

Suppose $\tilde{F_0} > (1+\epsilon) F_0$. Let $b_1, b_2 \ldots b_{F_0}$ be a listing of the F_0 distinct values

$\Rightarrow$ $h(b_1), h(b_2) \ldots h(b_{F_0})$

contains $\geq t$ elements smaller than V.

$$\frac{tM}{V} > (1+\epsilon) F_0 \Rightarrow V < \frac{tM}{F_0(1+\epsilon)}$$

What is $\Pr\left[h(b_i) < \frac{tM}{F_0(1+\epsilon)} \right]$?

If $h(b_i)$ is uniformly distributed, then it is $< \frac{t}{F_0(1+\epsilon)}$

We have F_0 (pairwise indep.) events happening. Each event has prob $p \cancel{=} < \frac{t}{F_0(1+\epsilon)}$. What is the chance that $\geq t$ happen ?

Let $X_i = 1$ iff $h(b_i) < \frac{tM}{F_0(1+\epsilon)}$.

$= 0$ otherwise

$E[X_i] \leq \frac{t}{F_0(1+\epsilon)}$ $\qquad E\left[\sum_{i=1}^{F_0} X_i\right] < \frac{t}{1+\epsilon}$

Let $Y = \sum_{i=1}^{F_0} X_i$ $\qquad E[Y] < \frac{t}{1+\epsilon}$

$Var[Y] = \sum_{i=1}^{F_0} Var[X_i] \leq \frac{t}{1+\epsilon}$. [THIS NEEDS "PAIRWISE INDEPENDENCE"]

$\Pr(X \wedge Y) = \Pr(X) \cdot \Pr(Y)$

$E[XY] = E[X] \cdot E[Y]$

NOT TRUE FOR DEPENDENT VARIABLES.

Note that $Var[X_i]$ is actually $p(1-p)$.

Use Chebyshev's Bound.

$$\Pr\left[\,|Y - E[Y]| \geq a\right) \leq \frac{Var[Y]}{a^2}$$

$$\Rightarrow \quad \Pr\left[\,\left|Y - \frac{t}{1+\varepsilon}\right| \geq a\right] \leq \frac{\frac{t}{1+\varepsilon}}{a^2}$$

$$\Pr[Y \geq t] \leq \Pr\left[\,\left|Y - \frac{t}{1+\varepsilon}\right| \geq \frac{t \cdot \varepsilon}{1+\varepsilon}\right] \leq \frac{\frac{\cancel{t}}{\cancel{1+\varepsilon}}}{\frac{t^{\cancel{2}}\varepsilon^2}{(1+\varepsilon)^{\cancel{2}}}} = \frac{(1+\varepsilon)}{t\varepsilon^2}$$

since $t = \frac{c}{\varepsilon^2}$ we get $\frac{1+\varepsilon}{c}$.

Similar proof when $\tilde{F}_0 < (1+\varepsilon) F_0$.

Defn $\quad Var[X] = E\left[(X - \mu_X)^2\right] = E[X^2] - (\mu_X)^2 \qquad \mu_X = E[X]$

(easy proof)

FACT $\quad E[X.Y] = E[X]. E[Y]$ (if X & Y are indep)

FACT $\quad Var[X+Y] = Var[X] + Var[Y]$ if X & Y are independent.

$$= E\left[(X+Y - E[X+Y])^2\right]$$

$$= E\left[((X-\mu_X) + (Y-\mu_Y))^2\right] = E\left[(X-\mu_X)^2\right] + E\left[(Y-\mu_Y)^2\right] + 2E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

$$= Var[X] + Var[Y] + 2\underbrace{\left[E[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y]\right]}$$

$$= 2\left[\mu_X\mu_Y - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y\right]$$

$$= 0$$