Frequency Moments

$F_K = \sum_{i=1}^{n} f_i^K$ $\quad f_i = \#$ times i appears

F_2 using $O(\frac{1}{\epsilon^2}\log(\frac{1}{\delta}))$

$\boxed{F_K \quad K > 2}$

lower bound of $\Omega(n^{1-\frac{2}{K}})$ space

Today: $\tilde{O}(n^{1-\frac{1}{K}})$ space [AMS '96]

Stream S of length m

$S = a_1, a_2 \cdots a_m \in [n]$

Choose random element a_p $\quad p \in [m]$

Suppose $a_p = \ell \in [n]$

$$R = |\{q \mid q \geq p,\ a_q = \ell\}|$$

$$Y = m\,(g(R) - g(R-1)) \qquad g(R) = R^K$$

Implementation:

maintain (x, R)

while $t = 0, 1 \ldots$ do

$t = t + 1$

$$(x, R) = \begin{cases} (a_t, 1) & \text{with prob. } 1/t \quad \text{(sampling)} \\ (x, R) & \text{if } S_t \neq x \\ (x, R+1) & \text{if } S_t = x \end{cases}$$

end

Lemma: For all T the prob. that the last sampling operation happened at time T is $\frac{1}{m}$

Proof: Prob. is product of

(1) Prob. that we sampled at time T: $\frac{1}{T}$

(2) Prob. that we did not sample at $t > T$ $\quad (1 - \frac{1}{t}) = \frac{t-1}{t}$

$$\text{Prob.} = \frac{1}{T}\left(\frac{T}{T+1}\right)\left(\frac{T+1}{T+2}\right)\cdots\left(\frac{m-1}{m}\right) = \frac{1}{m}$$

$$Y = m\,(g(R) - g(R-1)) \qquad g(R) = R^k$$

$$E[Y] = \sum_{i=1}^{n} \frac{f_i}{m} \frac{1}{f_i} \sum_{j=1}^{f_i} m\,(g(f_i - j + 1) - g(f_i - j))$$

$$= \sum_{i=1}^{n} (g(f_i) - g(f_i - 1)) + (g(f_i - 1) - g(f_i - 2)) \dots (g(1) - g(0))$$

$$= \sum_{i=1}^{n} g(f_i) = \sum_{i=1}^{n} f_i^k$$

Computing $Var[Y]$

__Lemma:__ $F_\infty \le F_k^{1/k}$

$$F_\infty = \max_i f_i = \left(\max_i f_i^k\right)^{1/k} \le \left(\sum_i f_i^k\right)^{1/k} = F_k^{1/k}$$

__Lemma:__ $$\frac{F_1}{n} = \underbrace{\frac{\sum f_i}{n} \le \left(\frac{\sum f_i^k}{n}\right)^{1/k}} = \left(\frac{F_k}{n}\right)^{1/k}$$

Hence $F_1 \le n^{1-\frac{1}{k}} F_k^{1/k}$

__Proof:__ x^k is convex

$$\left(\sum_{i=1}^{n} \frac{1}{n} f_i\right)^k \le \sum_{i=1}^{n} \frac{1}{n} f_i^k$$

$$\frac{\sum f_i}{n} \le \left(\frac{\sum f_i^k}{n}\right)^{1/k}$$

__Claim__ $g(x) = x^k$

$E[Y] = \sum_i g(f_i)$, $Var[Y] \le k F_1 F_{2k-1} \le k.n^{1-\frac{1}{k}} (F_k)^2$

__Proof:__ $Var[Y] \lesssim E[Y^2]$

$$= \sum_{i=1}^{n} \sum_{j=1}^{f_i} \frac{1}{m} \cdot (m\,(g(j) - g(j-1)))^2$$

$$= m \sum_{i=1}^{n} \sum_{j=1}^{f_i} \underbrace{[j^k - (j-1)^k]}_{\le k j^{k-1}} [j^k - (j-1)^k]$$

$$\mathrm{Var}[Y] \leq m \sum_{i=1}^{n} \sum_{j=1}^{f_i} k\, j^{k-1}\left[j^k - (j-1)^k\right]$$

$\frac{d}{dx} x^k = k x^{k-1}$

$j^k - (j-1)^k \leq k \cdot j^{k-1}$

$$= k m \sum_{i=1}^{n} \sum_{j=1}^{f_i} j^{2k-1} - \underbrace{j^{k-1}(j-1)^k}_{\leq -(j-1)^{2k-1}}$$

$$\leq k m \sum_{i=1}^{n} f_i^{2k-1}$$

$\boxed{F_1 = m}$

$$= k F_1 F_{2k-1}$$

$$F_1 F_{2k-1} = F_1 \left(\sum_{i=1}^{n} f_i^{2k-1}\right)$$

$f_i^{2k-1} = f_i^{k-1} \cdot f_i^k$

$\leq F_\infty^{k-1} \cdot f_i^k$

$$\leq F_1 \left(F_\infty^{k-1} \sum_{i=1}^{n} f_i^k\right)$$

$$\leq F_1 F_k^{\frac{k-1}{k}} \cdot F_k$$

using $F_\infty \leq F_k^{1/k}$

$F_1 \leq n^{1-\frac{1}{k}} F_k^{1/k}$

$$\leq n^{1-\frac{1}{k}} F_k^2$$

$\mathrm{Var}[Y] = k \cdot n^{1-\frac{1}{k}} F_k^2$

$E[Y] = F_k$

Take mean of $O\left(\frac{k \cdot n^{1-\frac{1}{k}}}{\varepsilon^2}\right)$ copies of Y

Take median of $O(\log(\frac{1}{\delta}))$ such means to get failure prob. $< \delta$

Estimating F_k for $k \leq 2$

linear sketch for F_k

F_1 sketch will estimate $\sum_{i=1}^{n} |f_i - g_i|$

(JL-lemma) $X_1, X_2 \sim N(0,1)$

$a_1 X_1 + a_2 X_2 \sim \sqrt{a^2+b^2}\, X \quad X \sim N(0,1)$

$a = (a_1 \dots a_n)$

$$\sum_{i=1}^{n} a_i X_i \sim \|a\|_2 X \qquad X \sim N(0,1)$$

<u>**Def^n**</u> (p-stable distrib^n) D is p-stable if given independent $X_1, X_2 \sim D$, for any $a, b \in \mathbb{R}$

$$aX_1 + bX_2 \sim \underbrace{(|a|^p + |b|^p)^{1/p}}_{(a,b)_p} \cdot X \qquad X \sim D$$

$$a_1X_1 + a_2X_2 + \dots + a_nX_n \sim (a)_p \cdot X \qquad X \sim D$$

<u>Claim</u>:
1. [Zol 86] p-stable distrib^n exist for $p \in (0, 2]$
2. Gaussian is 2-stable
3. Cauchy distrib^n $p(x) = \frac{1}{\pi} \frac{1}{x^2+1}$ is 1-stable
4. $p \in (1,2)$ D_p has finite mean & infinite variance

Distrib^n of absolute value → $p \in (0,1]$ D_p has infinite mean & infinite variance.

5. Easy way to sample from D_p

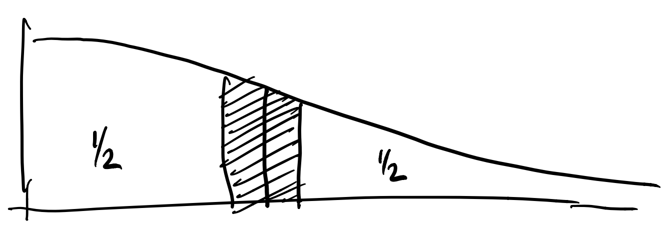
(Idea:)

(f_i)

$$\sum X_i f_i = \underbrace{(\sum f_i^p)^{1/p}}_{\text{estimate}} \cdot X$$

p-stable distrib^n p-stable distrib^n

[Indyk '00]



median

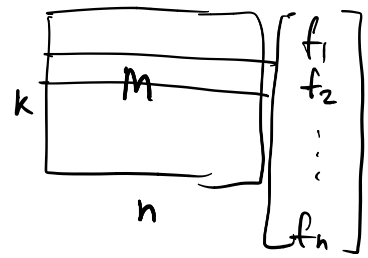
F_p for $0 < p \leq 2$

$f \leftarrow (f_1 \dots f_n)$

$k \leftarrow \theta\left(\frac{1}{\varepsilon^2}\log\left(\frac{f}{\varepsilon}\right)\right)$

M: $k \times n$ matrix where $M_{ij} \sim D_p$

$y \leftarrow Mx$ (streaming fashion)



$\sum f_i X_i$

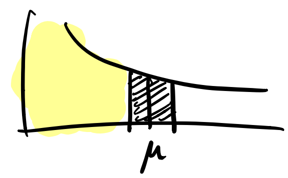
return $Y \leftarrow \left[\dfrac{\text{median}(|y_1|, |y_2| \dots |y_k|)}{\text{median}(|D_p|)}\right]$

Lemma: $\mu = \text{median}\, |D_p|$

$Pr(y \leq (1-\varepsilon)\mu) \leq \frac{\delta}{2}$ $\qquad Pr(y \geq (1+\varepsilon)\mu) \leq \frac{\delta}{2}$

$F(t)$: pdf of $|D_p|$

$\alpha = \min_{t \in [\mu(1-\varepsilon), \mu(1+\varepsilon)]} F(t)$



$\int_0^\mu F(t)\,dt = \frac{1}{2}$

$F(t)$ is continuous on $[(1-\varepsilon)\mu, (1+\varepsilon)\mu]$

$Pr[|y_i| \leq \mu(1-\varepsilon)] = \frac{1}{2} - \int_{\mu(1-\varepsilon)}^{\mu} F(t)\,dt \leq \frac{1}{2} - \alpha\mu\varepsilon = \frac{1}{2} - \gamma$

Let $\gamma = \alpha\mu\varepsilon$

$L = \#\, |y_i|$ in range $[0, \mu(1-\varepsilon)]$

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

for median to be low, $L > \frac{k}{2}$

$1 + \beta = \frac{1}{1 - 2\gamma}$

$$\Pr(y \leq (1-\varepsilon)\mu) = \Pr\left(L > \frac{K}{2}\right) = \Pr\left(L > \frac{1}{1-2\gamma} E[L]\right)$$

$$= \Pr\left(L > (1+\beta)\, E[L]\right)$$

$$\Pr(y \leq (1-\varepsilon)\mu) \leq e^{-\frac{\beta^2 E[L]}{3}}$$

$$-\frac{\beta^2 E[L]}{3} \leq -\frac{\gamma^2 E[L]}{3} \leq -\frac{K}{2}(1-2\alpha\varepsilon\mu)\,\alpha^2\varepsilon^2\mu^2$$

$$\Pr(\qquad) \leq e^{-c\,\varepsilon^2 K} \leq \frac{\delta}{2}$$

by setting $K = O\left(\frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$