Frequency moments:

$a_1, \ldots, a_T$ stream of elts from universe $U = \{0, 1, \ldots, p\}$

f_j^t # of times $j \in U$ appears in $\{a_1, \ldots, a_t\}$

$f_j \triangleq f_j^T$

F_k = k^{th} frequency moment of stream $= \sum_{j \in U} f_j^k$

Algorithm for approximating F_2 [Alon, Mathias, Szegedy]

- devise randomized unbiased estimator of F_2

  that can be computed on the fly

- repeat for accuracy

$h: U \to \{\pm 1\}$ equally likely

$Z := 0$

for $i := 1$ to T

$Z := Z + h(a_i)$ [++ if $h(a_i) = 1$ -- if $h(a_i) = -1$]

Return $X = Z^2$

Lemma

$E(X) = F_2$

Proof

$$Z = \sum_{j \in U} h(j) f_j$$

$$E(Z^2) = E\left[\left(\sum_{j \in U} h(j) f_j\right)^2\right] = E\left[\sum_{j \in U} \underbrace{\boxed{h(j)^2}}_{=1} f_j^2 + \sum_{j \in U, k \in U} h(j) h(k) f_j f_k\right]$$

$E(h(j)h(k)) = 0$

$$= F_2$$

Note: so far only used pairwise independence

To show that Z^2 is close to F_2, need to compute variance (and then can use trick of repetition to reduce variance)

$\boxed{\text{Lemma:}}$ $\mathrm{Var}(X) \le 3F_2^2$

$\Rightarrow$ repeating $t = \frac{2}{\varepsilon^2\delta}$ times and averaging

$$\Rightarrow \mathrm{Var}(\bar{X}) \le \frac{\mathrm{Var}(X)}{t} \le \frac{\varepsilon^2\delta\,\mathrm{Var}(X)}{2} \le \varepsilon^2\delta F_2^2$$

$$\Rightarrow \Pr\left(|\bar{X} - F_2| > \varepsilon F_2\right) \le \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2 F_2^2} \le \delta$$

Space: $O\left(\log m + \frac{\log T}{\varepsilon^2\delta}\right)$

[Actually can be improved to $O\left(\log m + \frac{\log\log T}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$

Guarantee

$\tilde{F}_2 (= \bar{X}) \in [F_2 \pm \varepsilon F_2]$ with prob $\ge 1-\delta$

Proof of lemma:

$$E(Z^4) = E\left(\left(\sum_{j\in U} h(j) f_j\right)^4\right)$$

$$= \sum_{j_1}\sum_{j_2}\sum_{j_3}\sum_{j_4} E\left[h(j_1)h(j_2)h(j_3)h(j_4)\right] f_{j_1} f_{j_2} f_{j_3} f_{j_4}$$

$E\left[h(j_1)h(j_2)h(j_3)h(j_4)\right] = 0$ if $\exists$ elt that appears only once

suppose j_1 appears once

$$E\left[h(j_1)h(j_2)h(j_3)h(j_4) \mid h(j_2)h(j_3)h(j_4) = v\right] = 1\cdot v + (-1)\cdot v = 0$$

$\Rightarrow$ only relevant terms are those where all terms appear an even # of times

$$= \sum_{j\in U} E\left[h(j)^4\right] f_j^4 + \binom{4}{2} \sum_{j_1 \neq j_2} E\left[h(j_1)^2\right] E\left[h(j_2)^2\right] f_{j_1}^2 f_{j_2}^2$$

$$= \sum_j f_j^4 + 6 \sum_{j_1\neq j_2} f_{j_1}^2 f_{j_2}^2 \qquad \lesssim 3 F_2^2$$

since $F_2^2 = \left(\sum_j f_j^2\right)\left(\sum_j f_j^2\right) = \sum_j f_j^4 + 2 \sum_{j_1\neq j_2} f_{j_1}^2 f_{j_2}^2$

4-wise independence sufficient!

Summing up:

sublinear space	heavy hitters,	F_0 (# distinct elts),	F_2
randomization		necessary	necessary
approx	necessary	necessary	necessary