Frequency Moments

f_i: # times i occurs in stream

$F_t = \sum f_i^t$

F_0 = # distinct elements $\quad 0^0 = 0$

F_1 = size of stream

F_2 = measure of skewness

[Alon, Matias, Szegedy '96]

$h: \mathcal{U} \to \{\pm 1\}$

$$Y = \sum_i h(x_i)$$

Estimator: Y^2

$$X_i = h(x_i) \qquad Y = \sum_i f_i X_i$$

$$\mathbb{E}[Y^2] = \mathbb{E}\left[\left(\sum_i f_i X_i\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{i,j} f_i f_j X_i X_j\right]$$

$$= \sum_i f_i^2 + \mathbb{E}\left[\sum_{i \neq j} f_i f_j X_i X_j\right]$$

$$i \neq j \quad \mathbb{E}[X_i X_j] = 0$$

$$\mathbb{E}[Y^2] = \sum f_i^2$$

$$\mathbb{E}[Y^4] = \mathbb{E}\left[\left(\sum_i f_i X_i\right)^4\right]$$

$$= \mathbb{E}\left[\sum_{i_1 i_2 i_3 i_4} f_{i_1} f_{i_2} f_{i_3} f_{i_4} X_{i_1} X_{i_2} X_{i_3} X_{i_4}\right]$$

$$= \sum_i f_i^4 + 6 \sum_{i<j} f_i^2 f_j^2$$

$$\left(\sum f_i X_i\right)\left(\sum f_i X_i\right)\left(\sum f_i X_i\right)\left(\sum f_i X_i\right)$$

$$\binom{4}{2}$$

$$\mathrm{Var}[Y^2] = \mathbb{E}[Y^4] - (\mathbb{E}[Y^2])^2 \qquad \mathbb{E}[Y^2] = \sum_i f_i^2$$

$$= \Big(\sum_i f_i^2 + 6 \sum_{i<j} f_i^2 f_j^2\Big) - \Big(\sum_i f_i^4 + 2 \sum_{i<j} f_i^2 f_j^2\Big)$$

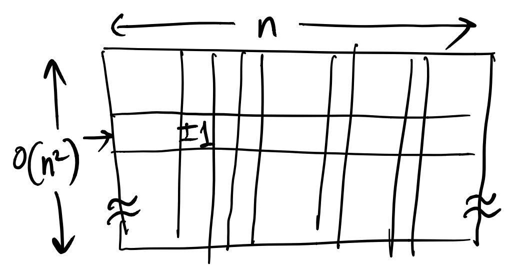
$$= 4 \sum_{i<j} f_i^2 f_j^2 \leq 2 F_2^2 = 2(\mathbb{E}[Y^2])^2$$

$O(\frac{1}{\varepsilon^2})$ copies and take mean get $(1 \pm \varepsilon)$ approx to F_2
w. prob. $\frac{2}{3}$

$O(\log(\frac{1}{\delta}))$ groups and compute median of means
$\rightarrow$ prob. of success $1-\delta$

$(1+\varepsilon)$ approximation to F_2 with prob. $(1-\delta)$
$O(\frac{1}{\varepsilon^2} \log(\frac{1}{\delta}))$ space

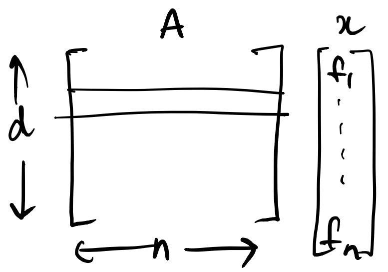
$\mathbb{E}[X_i^2] = 1 \quad \mathbb{E}[X_i X_j] = 0$
$\mathbb{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}] = 0$



Any 4 columns $i_1\ i_2\ i_3\ i_4$
$\{\pm 1\}^4 \quad +1, -1, -1, +1$
Orthogonal arrays of strength 4
parity check matrices of BCH codes

seed $\equiv$ one of $O(n^2)$ rows $\quad O(\log n)$
Any entry of matrix can be computed in $O(\log n)$ space & $O(\log n)$ time

$Y_1, Y_2 \ldots Y_d \qquad d = O(\frac{1}{\varepsilon^2} \log(\frac{1}{\delta}))$
$x \in \mathbb{R}^n \qquad x = (f_1\ f_2 \ldots \ldots f_n)$



Algorithm stores Ax

$x + \Delta x$ $\quad \Delta x = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$

$Ax + A\Delta x$

linear sketch $\quad Ax + Ay = A(x+y)$

turnstile model: input is x

updates to co-ordinates of x (need not be $+1$)

Linear sketches can be used in turnstile model.

$$\underbrace{A(x+\Delta x)}_{\text{new sketch}} = \underbrace{Ax}_{\text{old sketch}} + \underbrace{A\Delta x}_{\text{update}}$$

Johnson-Lindenstrauss Lemma

<u>Lemma</u>: $G \subset (\mathbb{R}^d, \ell_2)$ be a set of n points.

For any $0 < \varepsilon < \frac{1}{2}$ $\quad K = O(\log n/\varepsilon^2)$

$\exists$ mapping $f : \mathbb{R}^d \longrightarrow \mathbb{R}^K$

such that for all $v_i, v_j \in G$

$$(1-\varepsilon)\,\|v_i - v_j\| \le \|f(v_i) - f(v_j)\|_2 \le (1+\varepsilon)\,\|v_i - v_j\|_2$$

linear mapping: $f(v) = \dfrac{Mv}{\sqrt{K}}$

$M \in \mathbb{R}^{K \times d}$ $\quad M_{ij} = \mathcal{N}(0,1)$ $\quad$ data-oblivious

Use tail bound for chi-squared distributions.

<u>Lemma</u>: Let $Z_1 \ldots Z_k$ be i.i.d. unit normal random variables.

Let $Y = \sum_i Z_i^2$

Then $Pr[(1-\varepsilon)^2 k \le Y \le (1+\varepsilon)^2 k] \geq 1 - 2e^{-c\varepsilon^2 k}$

for some suitable constant c

[Pf idea] $Pr\left[\frac{Y}{k} \geq 1+\varepsilon\right] \le Pr\left[e^{tY} \geq e^{tk(1+\varepsilon)}\right]$

$$\le \frac{\mathbb{E}[e^{tY}]}{e^{tk(1+\varepsilon)}}$$

$$= \prod_i \frac{\mathbb{E}[e^{tZ_i^2}]}{e^{t(1+\varepsilon)}}$$

(Crux): Bounding $\mathbb{E}[e^{tZ_i^2}]$ where Z_i is unit normal.

$f(v) = \frac{Mv}{\sqrt{k}}$ $\qquad$ M is $k \times d$ matrix

$$(Mv)_i = M_i^T v = \sum_{j=1}^{d} M_{ij} v_j = \Big(\sum_{j=1}^{d} v_j^2\Big)^{1/2} Y = Y$$

unit vector $v \in \mathbb{R}^d$ $\qquad$ unit normal

$$\alpha_1 X_1 + \alpha_2 X_2 = (\alpha_1^2 + \alpha_2^2)^{1/2} X$$

$$Pr\left[k(1-\varepsilon)^2 \le \sum_{i=1}^{k} (Mv)_i^2 \le k(1+\varepsilon)^2\right] \geq 1 - 2e^{-c\varepsilon^2 k}$$

$$Pr\left[(1-\varepsilon) \le \left\|\frac{Mv}{\sqrt{k}}\right\|_2 \le (1+\varepsilon)\right] \geq 1 - 2e^{-c\varepsilon^2 k}$$

$v \qquad \tilde{v} = \frac{v}{\|v\|_2}$

$$Pr\left[(1-\varepsilon) \le \left\|\frac{M\tilde{v}}{\sqrt{k}}\right\|_2 \le (1+\varepsilon)\right] \geq 1 - 2e^{-c\varepsilon^2 k}$$

$$\Pr\left[(1-\varepsilon)\|v\|_2 \le \left\|\frac{Mv}{\sqrt{k}}\right\|_2 \le (1+\varepsilon)\|v\|_2\right] \ge 1-2e^{-c\varepsilon^2 k}$$

$v = v_i - v_j$ $\quad\binom{n}{2}$ such pairs.

$$\Pr\Big[(1-\varepsilon)\|v_i-v_j\|_2 \le \underbrace{\left\|\frac{M(v_i-v_j)}{\sqrt{k}}\right\|_2}_{\|f(v_i)-f(v_j)\|_2} \le (1+\varepsilon)\|v_i-v_j\|_2\Big] \ge 1-2e^{-c\varepsilon^2 k}$$

$f(v_i) = \frac{Mv_i}{\sqrt{k}}$ $\quad \|f(v_i)-f(v_j)\|_2 = \left\|\frac{M(v_i-v_j)}{\sqrt{k}}\right\|_2$

Choosing $k = \frac{c' \log(n)}{\varepsilon^2}$

Failure prob : $2e^{-cc'\log n} < \frac{1}{n^3}$

Overall failure prob. $\le \frac{1}{n}$ (union bound)

[Alon '03] $\Omega\left(\frac{\log n}{\varepsilon^2 \log(\frac{1}{\varepsilon})}\right)$ dimensions

[Larsen, Nelson '17] $\Omega\left(\frac{\log n}{\varepsilon^2}\right)$ dimensions

Single pair : $O\left(\frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ dimensions

ℓ_2 is special

ℓ_p $\quad p=\infty$ dimension red^n not possible.

roughly, to preserve distances to factor α need $n^{1/\alpha}$ dimensions

$p=1$ dimension red^n not possible.

to achieve factor α, need $n^{\frac{1}{\alpha^2}}$ dimensions [Brinkman, C, '03]

$\frac{n}{\alpha}$ upper bound