Plan:

Frequency Moments $F_K$ $K \in (0,2]$

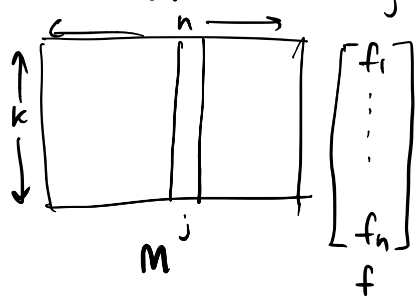
Lower Bound $\ell_2$ dimension reduction

Heavy Hitters

$F_K$ $\quad K \in (0,2]$

p-stable Distribution $D_p$

$K \times n$ matrix $M$ $\quad M_{ij} \sim D_p$ $\quad K = O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta})$



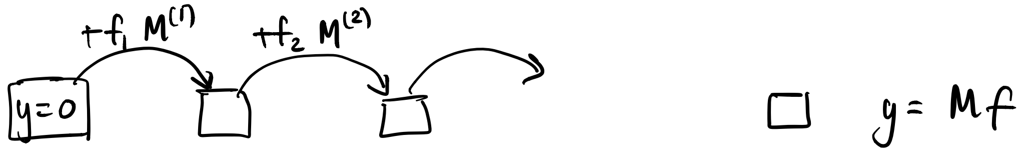

$y = Mf$

$y_i = \sum_{j=1}^{n} M_{ij} f_j$

$y_i \sim (\sum |f_j|^p)^{1/p} D_p$

$y := y + M^{(j)}$ when $j$ appears

Space $= \underbrace{O(\frac{1}{\varepsilon^2} \log(\frac{f}{\delta}) \log n)}_{S} +$ space for random matrix




n steps

S random bits in each step $\quad R = n$ steps

$S = O(\frac{1}{\varepsilon^2} \log(\frac{f}{\delta}) \log n)$

$U_t$: uniform random string in $\{0,1\}^t$

Nisan's pseudorandom generator:

$\exists\, h: \{0,1\}^{S \log R} \longrightarrow \{0,1\}^{SR}$

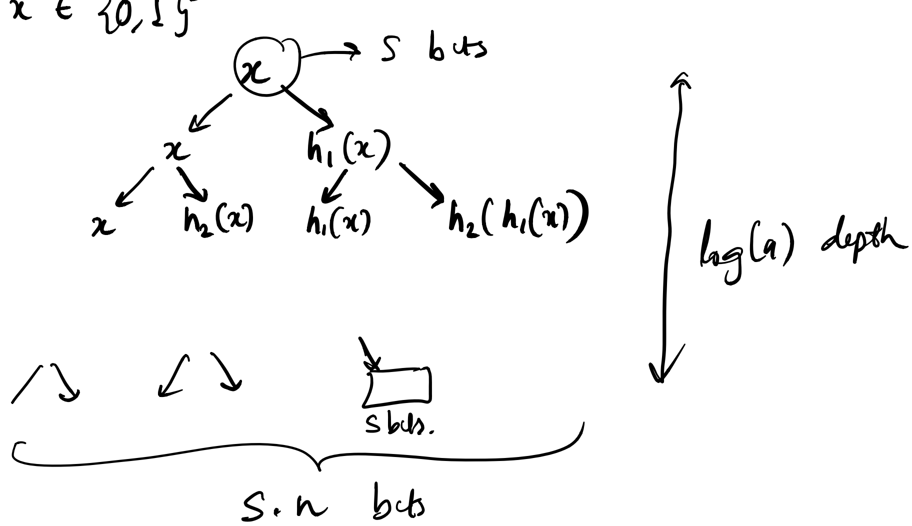
$|Pr(f(U_{SR}) = 1) - Pr(f(h(U_{S \log R})) = 1)| \leq 2^{-O(S)}$

$f$: result of executing space $S$ decision algorithm for $R$ steps with $S$ random bits per step

$|\text{Seed}| = S \log R = O\left(\frac{1}{\varepsilon^2} \log \frac{1}{\delta} \log^2 n\right)$

$h_1 \; \ldots \; h_{\log n}$ pairwise independent hash fns.
$h_i : [2^S] \longrightarrow [2^S]$

choose $x \in \{0,1\}^S$




$F_1 \qquad f_i \qquad g_i \qquad \sum |f_i - g_i|$
$Mf \qquad Mg \qquad M(f-g) \longrightarrow$ estimate $\|f-g\|_1$

lower bound for dimension redn in $l_1$ $\quad n^{1/\alpha}$ dimensions
$(1+\varepsilon)$ approxn in space $O\left(\frac{1}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ with prob $1-\delta$
$O\left(\frac{1}{\varepsilon^2} \log n\right)$

Dimn redn: treat sketch as mapped onto $l_1^{\text{small dimension}}$
compute $l_1$ norm of sketch

# Lower Bound $l_2$ dimension reduction

Thm: [Alon '00] Let $v_1 \ldots v_{n+1} \in \mathbb{R}^d$ $\quad \frac{1}{\sqrt{n}} \leq \varepsilon \leq \frac{1}{3}$

$$1 \leq \|v_i - v_j\| \leq 1+\varepsilon \quad \forall\ i \neq j \in [n+1]$$

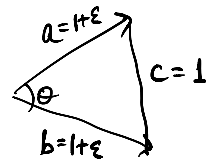
Then subspace spanned by $v_1 \ldots v_{n+1}$ has dimension

$$d = \Omega\left(\frac{\log n}{\varepsilon^2 \log(1/\varepsilon)}\right)$$

$n$ dimensional simplex cannot be embedded with distortion $1+\varepsilon$ in fewer than $\Omega(\log n)$ dimensions

Assume $v_{n+1} = 0$. So $1 \leq \|v_i\| \leq 1+\varepsilon$

Set $v_i' = \dfrac{v_i}{\|v_i\|}$




$$\langle v_i', v_j' \rangle = \cos \angle(v_i, v_j)$$
$$\leq \frac{1}{2} + \varepsilon + \frac{\varepsilon^2}{2}$$

$$c^2 = a^2 + b^2 - 2ab\cos\theta$$

$$\left|\langle v_i', v_j' \rangle - \frac{1}{2}\right| = O(\varepsilon)$$

$$\cos\theta = \frac{a^2+b^2-c^2}{2ab} = \frac{(1+\varepsilon)^2 + (1+\varepsilon)^2 - 1}{2(1+\varepsilon)^2}$$

Define $n \times n$ matrix $B$ $\quad B_{ij} = \langle v_i', v_j' \rangle$

$$\begin{bmatrix} 1 & & & & \frac{1}{2}+O(\varepsilon) \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

If $\varepsilon = 0$, $B$ has rank $n$

$\Rightarrow$ $v_i$'s span subspace of dim. $n$

No distortion $\Rightarrow$ dim. $\geq n$

$d = \text{rank}(B)$ $\quad$ Define $C = 2B - J$, $J = ee^T$ (all ones)

$$|\text{rank}(C) - \text{rank}(B)| \leq 1$$

$$?? \leq \text{rank}(C) \leq d+1$$

$$C = \begin{bmatrix} 1 & & & O(\varepsilon) \\ & 1 & & \\ & & 1 & \\ O(\varepsilon) & & & 1 \end{bmatrix}$$

**Lemma** Consider symmetric matrix $C$ $C_{ii} = 1$

$$|C_{ij}| \leq \frac{1}{\sqrt{n}} \quad i \neq j$$

then $\text{rank}(C) \geq \frac{n}{2}$

**Proof:** $C$ symmetric $\Rightarrow$ all eigenvalues real

$d = \text{rank}(C)$ $\lambda_1 \dots \lambda_d \in \mathbb{R}$ non-zero eigenvalues

$$\text{Tr}(C) = \sum_{i \in [n]} C_{ii} = n = \sum_{i \in [d]} \lambda_i$$

Non-zero eigenvalues of $C^2 = C^T C$ are $\lambda_1^2 \dots \lambda_d^2$

$$\text{Tr}(C^2) = \sum_{i \in [d]} \lambda_i^2$$

$$\text{Tr}(C^2) = \sum_i \sum_j C_{ij}^2 \leq n + n(n-1)\frac{1}{n} = 2n - 1 < 2n$$

$$\frac{2n}{d} > \frac{\sum \lambda_i^2}{d} \geq \left(\frac{\sum \lambda_i}{d}\right)^2 = \left(\frac{n}{d}\right)^2$$

$$\Rightarrow d > \frac{n}{2}$$

**Lemma:** Suppose $n \times n$ matrix $A$ has rank $d$

then $F = (A_{ij}^k)$ has rank at most $\binom{d+k-1}{d-1}$

**Proof** Let $v_1 \dots v_d \in \mathbb{R}^n$ be basis for row space of $A$

$i^{th}$ row $A_i = \sum_{l \in [d]} \lambda_l v_l$ for some coeff $\lambda_l$

$$A_{ij} = \sum_{l \in [d]} \lambda_l v_{l,j}$$

$$F_{ij} = A_{ij}^k = \left(\sum_{l \in [d]} \lambda_l v_{l,j}\right)^k$$

$$= \sum_{k_1 + \dots k_d = k} \underbrace{\binom{k}{k_1 \dots k_d}\left(\prod_{l \in [d]} \lambda_l^{k_l}\right)}_{\text{coefficient}} \underbrace{\left(\prod_{l \in [d]} v_{l,j}^{k_l}\right)}_{j^{th} \text{ coord of basis}}$$

row-space of $F$ spanned by

$$(W_{k_1 \ldots k_d})_j = \prod_{\ell \in [d]} v_{\ell,j}^{k_\ell}$$

one vector for every choice of $k_1 \ldots k_d$ $\sum k_\ell = K$

$\#$ basis vectors $= \#$ partitions $= \binom{K+d-1}{d-1} = \binom{K+d-1}{K}$

Proof of Thm

$?? \leq \operatorname{rank}(C) \leq d+1 \qquad |C_{ij}| \leq O(\varepsilon)$

$K$ integer $\quad \varepsilon^K \leq \frac{1}{\sqrt{n}}$

Consider $F = \left(C_{ij}^K\right)_{1 \leq i,j \leq n}$

$$|F_{ij}| \leq \frac{1}{\sqrt{n}}$$

$\operatorname{rank}(F) \leq \binom{K+d}{d}$ (lemma 2)

$\operatorname{rank}(F) \geq \frac{n}{2}$

$$\frac{n}{2} \leq \binom{K+d}{d} = \binom{K+d}{K} = \frac{(K+d)!}{d!} \cdot \frac{1}{K!} \leq (K+d)^K \left(\frac{e}{K}\right)^K$$

take logs of both sides $\quad K = \frac{\ln n}{2 \ln\left(\frac{1}{\varepsilon}\right)}$

$$K \ln\left(\frac{e(d+K)}{K}\right) \geq \ln\left(\frac{n}{2}\right)$$

$$\frac{\ln n}{2\ln\left(\frac{1}{\varepsilon}\right)} \ln\left(\frac{e(d+K)}{K}\right) \geq \ln\left(\frac{n}{2}\right)$$

$$\ln\left(\frac{e(d+K)}{K}\right) \geq (1-o(1))\, 2\ln\left(\frac{1}{\varepsilon}\right)$$

$$\frac{e(d+K)}{K} \geq (1-o(1)) \frac{1}{\varepsilon^2}$$

$$\frac{d}{K} \geq (1-o(1)) \frac{1}{e\varepsilon^2} - 1$$

$$d \geq \Omega\left(\frac{\ln n}{\varepsilon^2 \ln\left(\frac{1}{\varepsilon}\right)}\right)$$

$$\frac{\ln(n/2)}{\ln n} = \frac{\ln(n) - c}{\ln(n)} = 1 - \frac{c}{\ln(n)} = 1 - o(1)$$

## Heavy Hitters:

length $m$ element appears $> \frac{m}{2}$ times

Misra-Gries '82

initialize $k$ bins each with elt (initially null)
and a counter (initially 0)

For each element $e$ in stream
    If $e$ is in a bin $b$ then
        increment $b$'s counter
    else if find a bin whose counter $= 0$
        set its element $= e$, counter to $1$
    else
        decrement counter of every bin

For each bin $b$ do
    $i \leftarrow$ element in bin $b$,
    return $\tilde{f}_i = b$'s counter

If $f_i$ is true frequency of element $i$
$\tilde{f}_i$ : frequency returned by algo

$$f_i - \frac{m}{k} \leq \tilde{f}_i \leq f_i$$