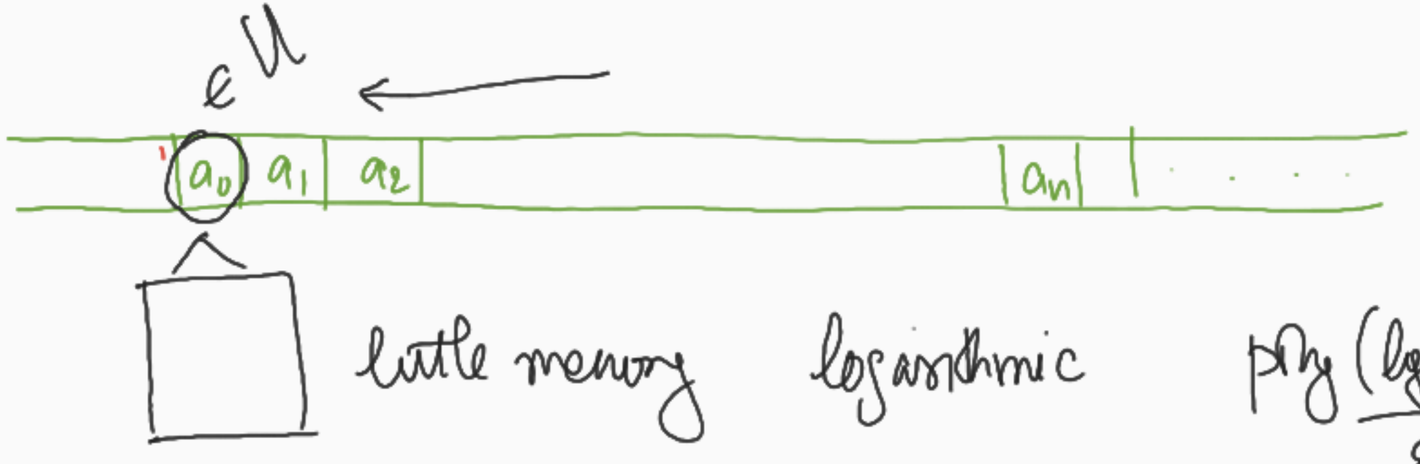
Lecture 13: The Data Stream Model

- Recap: dimension reduction → Approximation → Randomization
- data stream model
- computing 2nd moments
- distinct elements? ← Approx matrix multiplication

$O(\lg \lg n)$ bits

- how many elements?
- distinct?
- sum of elts?
- most frequent?
- median, mean,

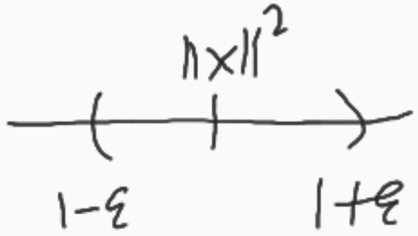


little memory logarithmic $\text{poly}(\lg n)$ $\frac{}{\varepsilon^{O(1)}}$

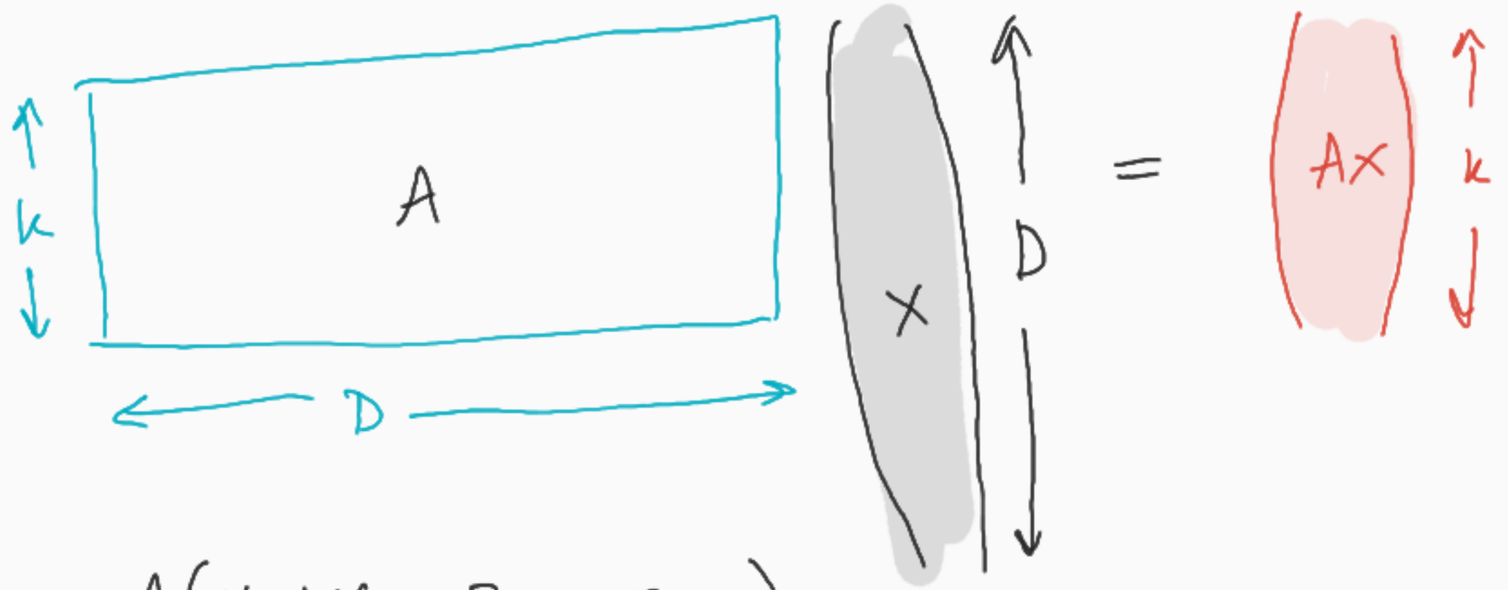
Recap: JL Lemma:

gen random matrix $A = \frac{1}{\sqrt{k}} M$ (k×D matrix) s.t. $\forall x \in \mathbb{R}^D$

$$Pr\left[\|Ax\|^2 \notin (1\pm\epsilon)\|x\|^2\right] \le \frac{1}{n^2}\delta$$



where $K = C\frac{\log n}{\epsilon^2} \log\frac{1/\delta}{\epsilon^2}$

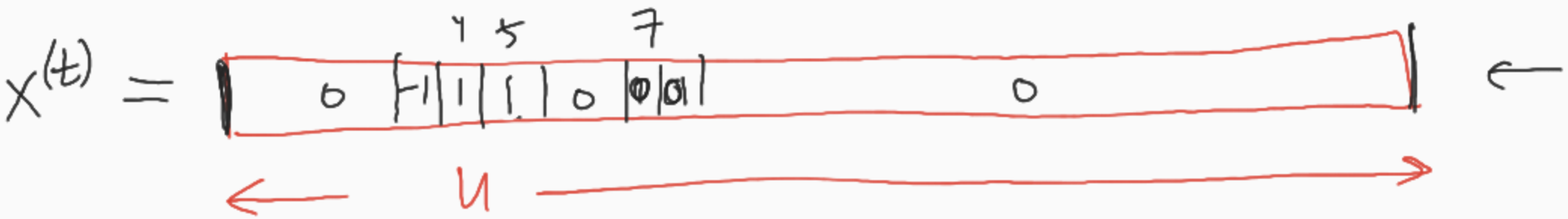


$$A(x_1 + x_2 - 3x_3 + 5x_4) = Ax_1 + Ax_2 - 3Ax_3 + 5Ax_4$$

Computing Second moments (F_2)

$1^2+1^2+2^2$ $\qquad$ 1^2+1^2

$(5,+)\ (4,+)\ (7,+)\ (5,+) \mid (5,-)\ (7,-)\ (8,+)\ (8,-) \mid$



#distinct elts $= \#\text{nnz}(x) = \|x\|_0$

#active conns $= \|x\|_1$

$$F_2 = \|x\|_2^2$$

heavy hitters $= \|x\|_\infty = \arg\max_i \{x_i\}$

at any time t, if i query, should get correct answer w.p. $\geq 1-\delta$

Computing F_2 [Alon Matias Szegedy]

$C \leftarrow 0$

Fix $h : U \to \{-1, 1\}$ 4wise independent hash fn.

if $(i, +)$ then $C \leftarrow C + h(i)$

$(i, -)$ $C \leftarrow C - h(i)$

if query then return C^2

$k$wise indep hash family $H = \{h : U \to [M]\}$ is family of fns.
$=$ of $U \to [M]$

$$\Pr_{h \leftarrow H}\left[h(a_1) = \alpha_1 \wedge h(a_2) = \alpha_2 \ldots \wedge h(a_k) = \alpha_k\right] = \frac{1}{M^k}$$

$\forall a_1 \neq a_2 \neq \ldots \neq a_k \in U$, $\alpha_1, \alpha_2 \ldots \alpha_k \in [M]$

Easy to sample from:
$k \log |U|$ bits of randomness

$h(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$

4-wise indp

all arithmetic over finite field

choose uniformly from $\mathbb{F}_{2^u}$

$U \to 2^m = M$

$2^u \to 2^u$

drop highest $u-m$ bits

$\Rightarrow$ 4-wise indp hash fn from $U \to [M]$

total randomness $= 4 \lg |U|$

Lemma 1: $\mathbb{E}[C^2] = F_2$

$$\mathbb{E}[C^2] = \mathbb{E}\left[\left(\sum_i x_i h_i\right)^2\right]$$

$$= \mathbb{E}\left[\left(\sum_i x_i h_i\right)\left(\sum_j x_j h_j\right)\right]$$

$$= \mathbb{E}\left[\sum_i x_i^2 h_i^2 + \sum_{i \neq j} x_i x_j h_i h_j\right]$$

$h_i^2 = 1$

$$= \sum x_i^2 + \sum x_i x_j \mathbb{E}[h_i]\mathbb{E}[h_j]$$

$\mathbb{E}[h_i] = 0$, $\mathbb{E}[h_j] = 0$

$$= \sum x_i^2 = F_2$$

$C^k \leftarrow \sum X_i h(i)$

h_i

$$\mathrm{Var}(C^2) = \mathbb{E}[(C^2)^2] - \mathbb{E}[C^2]^2$$

$\mathbb{E}[C^4]$

$$\mathbb{E}\left[\left(\sum x_i h_i\right)^4\right]$$

$$\mathbb{E}\left[\sum x_i^4 h_i^4 + c_1 \sum_{i,j} x_i^2 x_j^2 h_i^2 h_j^2\right]$$

$h_i^4 = 1$, $h_i^2 h_j^2 = 1$

$$+ c_2 \sum_{i,j,k} x_i^2 x_j x_k h_i^2 h_j h_k$$

$$+ \sum_{i,j,k,l} x_i x_j x_k x_l h_i h_j h_k h_l + \sum_{i,j} x_i^3 x_j h_i^3 h_j \Big]$$

$$4 \sum_{i \neq j} x_i^2 x_j^2 = \cancel{\sum x_i^4} + c \sum_{i,j} x_i^2 x_j^2 - \left(\cancel{\sum x_i^4} + 2 \sum_{i \neq j} x_i^2 x_j^2\right)$$

$$= \sum x_i^4 + c \sum_{i \neq j} x_i^2 x_j^2 - \left(\sum x_i^2\right)^2$$

Lemma 1: $E[C^2] = F_2 = \sum_i x_i^2$

Lemma 2: $Var(C^2) = 4 \sum'_{i,j} x_i^2 x_j^2$ $\quad i \neq j$

Paghuti Chebyshev's Ineq.

$$Pr[|C^2 - \mu| \geq \varepsilon\mu] \leq \frac{Var(C^2)}{\varepsilon^2 \mu^2} = \frac{4 \sum'_{i,j} x_i^2 x_j^2}{k\varepsilon^2 \left(\sum_i x_i^2\right)^2} \leq \frac{4}{\varepsilon^2 k}$$

(bigger than 1!) ☹

$C_1, C_2 \ldots C_k$

query $\leftarrow \frac{\sum_{i=1}^{k} c_i^2}{k}$ ← mean F_2, var $\leq \frac{4}{k}$

$= \delta \Rightarrow k = \frac{4}{\varepsilon^2 \delta}$

$X_1, X_2 \ldots X_k$ iid (μ, σ^2)

$\bar{X} = \left(\frac{\sum_i x_i}{k}\right)$ $\quad \mu, \sigma^2/k$

h_1 $\quad C_1 = \sum_i x_i h_1(i)$

h_2 $\quad h_2 \cdot x$

h_k $\quad C_k = \sum_i x_i h_k(i)$

---

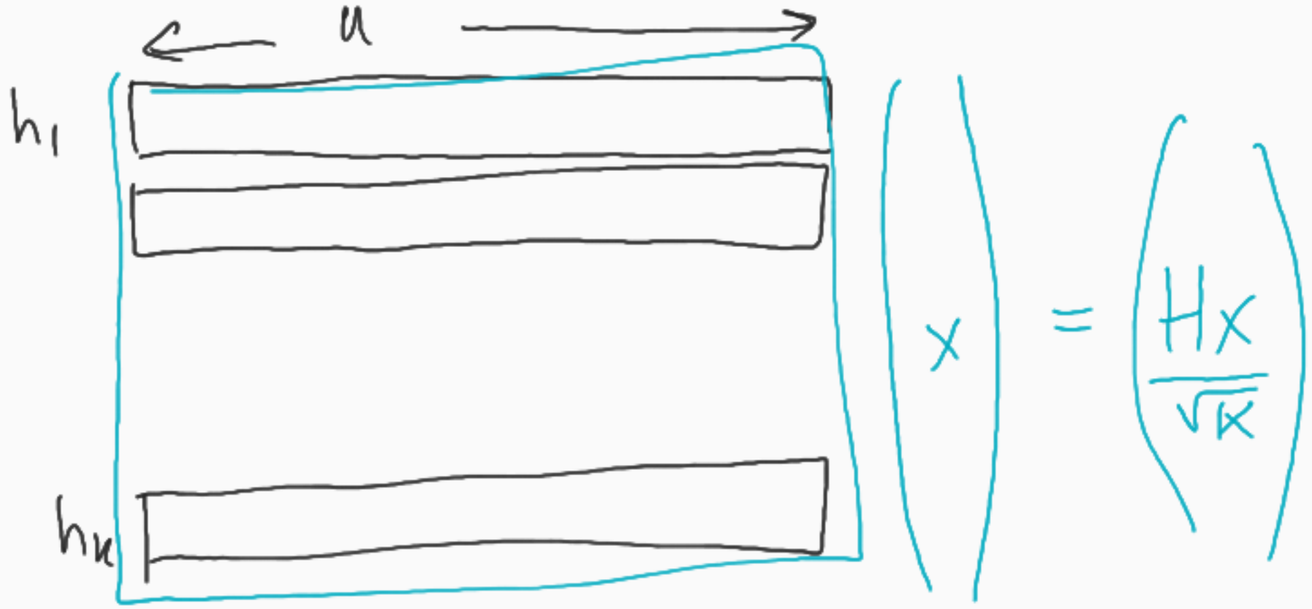
rehmed $\left(\frac{\sum_{j=1}^{k} C_j^2}{k} \right)$

JL: $k = \frac{\log 1/\delta}{\varepsilon^2}$

Cheby: $\frac{1/\delta}{\varepsilon^2}$

large space

sm space

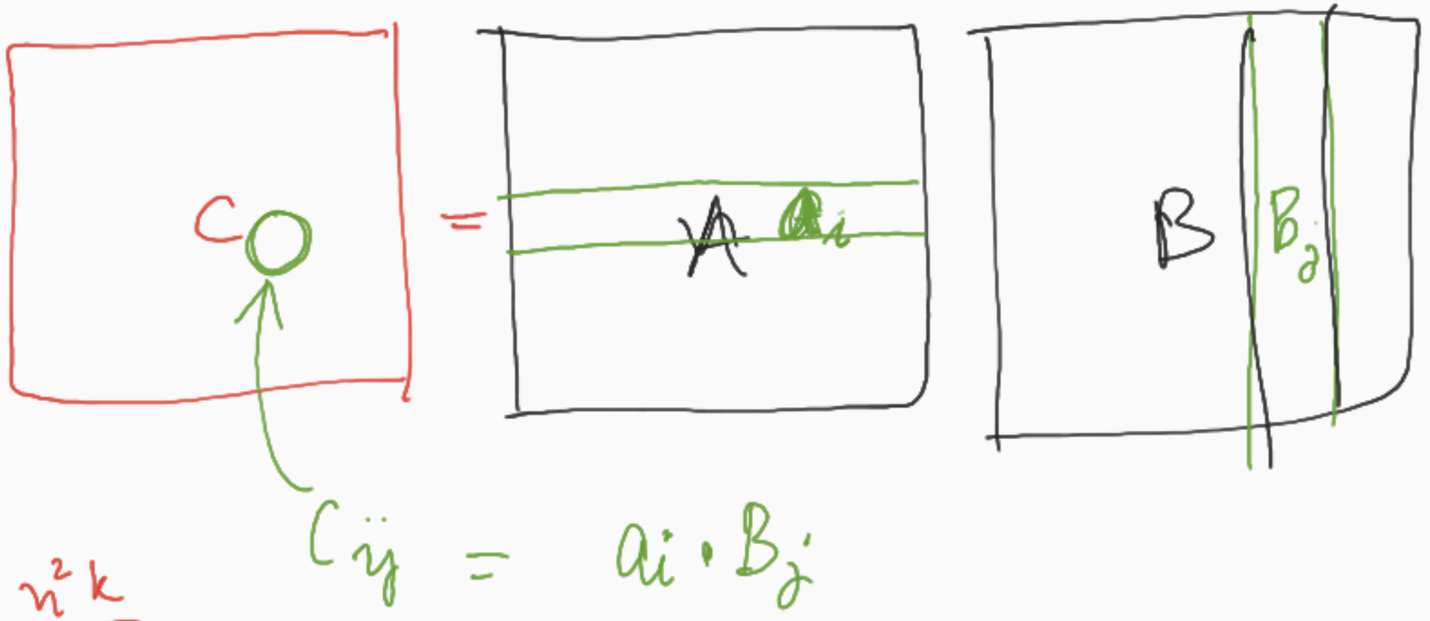


$\frac{1}{\sqrt{k}} H$

$\| x \| = \| \frac{1}{\sqrt{k}} H(x) \|^2$

return squared norm.

AB



approx

$\tilde{C} \approx$