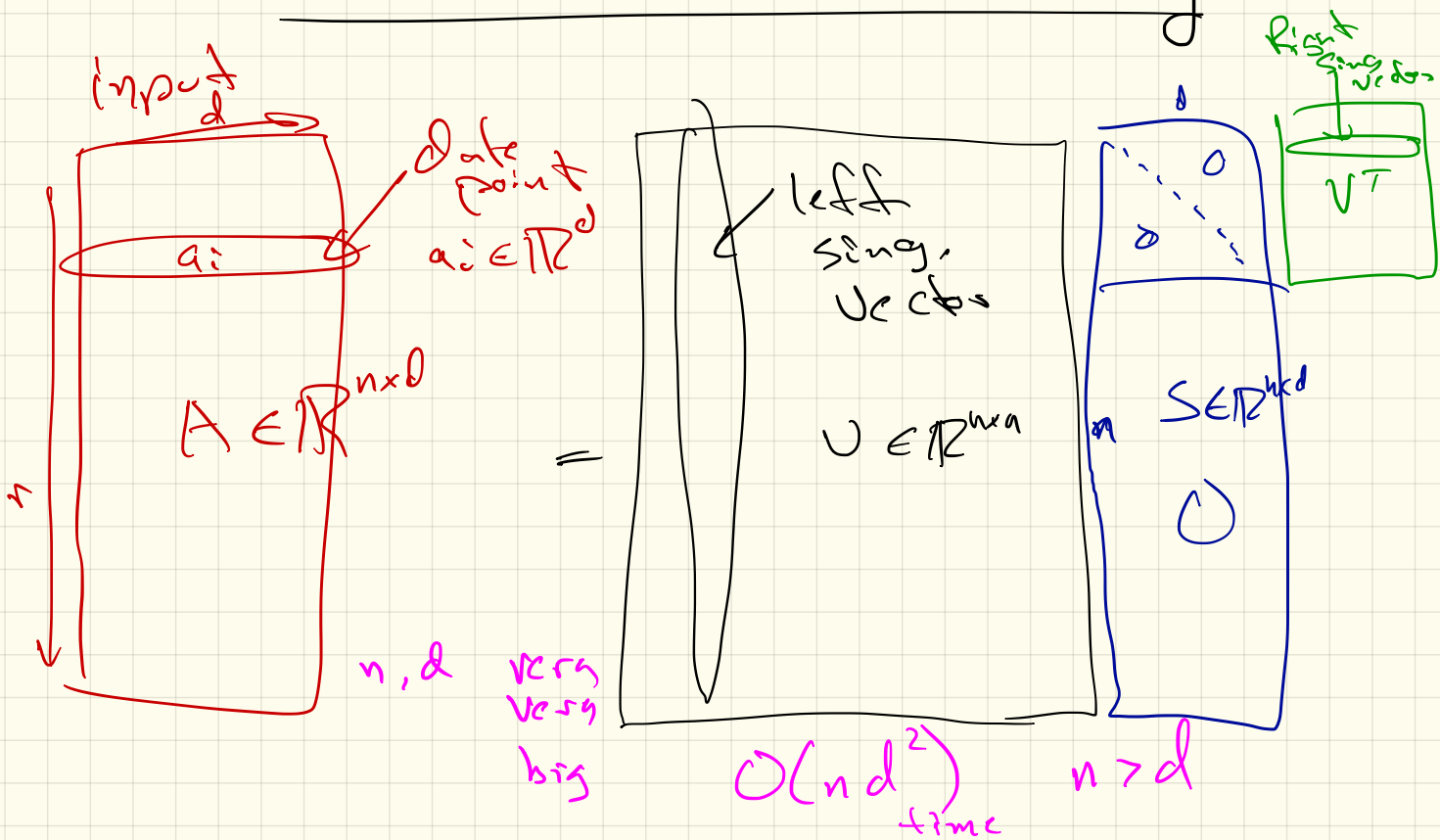
Matrix Sketching



Streaming Setting

$A = (a_1, a_2, \ldots a_t, \ldots a_n) \qquad a_i \in \mathbb{R}^d$

n very very big, d = manageable ≈ 1000

$C \in \text{zeros}(d, d)$

for $(a_i \in A)$

$C = C + a_i a_i^T \in \mathbb{R}^{d \times d}$ ← outerproduct

space $O(d^2)$

time $O(nd^2)$

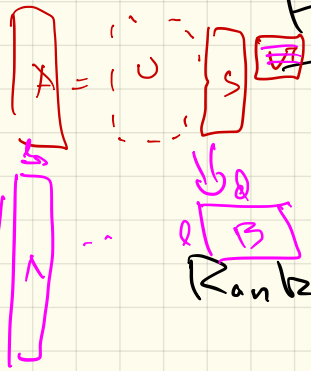
return C

$C = A^T A \in \mathbb{R}^d$

eigenvectors (C) = right sing vect (A)

eigenvalue (C) = squared sing. val (A)

Frequent Directions



$\Rightarrow d^2$ very very big

n very, very big $= 100$ mill

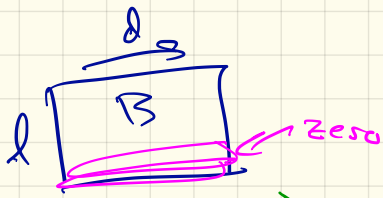
d very big ≈ 100 thousand

Rank-k apx $\quad k$ small $k = 10$ $\Rightarrow$ $O(dk)$ space $\quad O(ndk)$ time

Freq. Dir

$B = \text{zeros}(d \times \ell)$

$\ell = k + 1/\varepsilon$ or $k + k/\varepsilon$



for $(a_i \in A)$

Insert a_i into B

if (B no all zero rows) $\leftarrow O(d\ell^2)$ $\rightarrow O(nd\ell^2)$ $\rightarrow O(nd\ell)$

$[U, S, V^T] = \text{svd}(B)$

Set $\delta = s_{\ell\ell}^2 \leftarrow \ell$th sing val $(B)^2$

$S' = \text{diag}\left(\sqrt{s_{11}^2 - \delta}, \sqrt{s_{22}^2 - \delta}, \ldots, \sqrt{s_{\ell-1,\ell-1}^2 - \delta}, 0\right) 0 0 \ldots 0)$

Return $B = S'V^T$

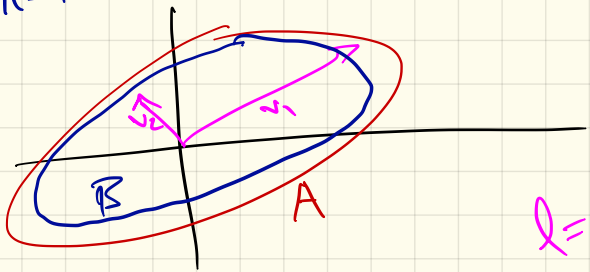
$B \leftarrow FD_\ell(A)$

$\|A\|_F^2 = \sum_i \sum_j a_{ij}^2 = \sum_{j=1}^{d} \sigma_j^2$

(1) $\forall x$, $\|x\|=1$

$$0 \le \|Ax\|^2 - \|Bx\|^2 \le \|A - A_k\|_F^2$$

$\ell = \frac{\ell}{2} + k$



$\ell = \frac{k}{\varepsilon} + k$

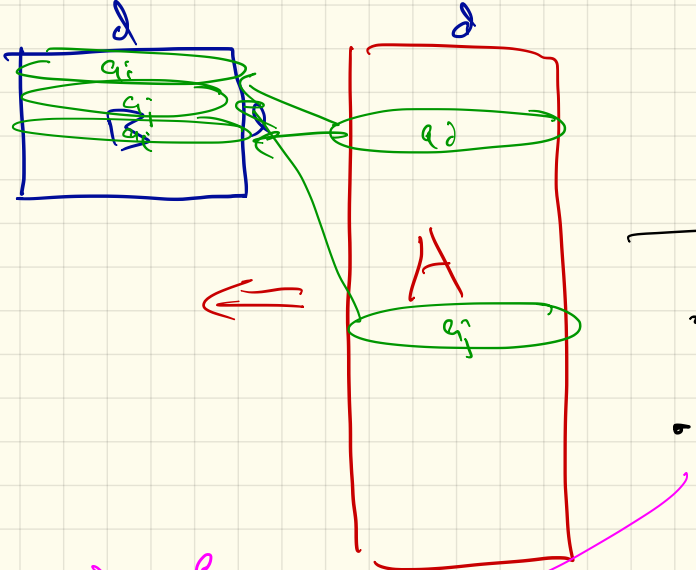
$$\le \frac{\|A\|_F^2}{\ell} \quad k=0, \ \ell = \frac{1}{\varepsilon}$$

(2) $$\|A - A\Pi_{B_k}\|_F^2 \le \frac{\ell}{\ell - k} \|A - A_k\|_F^2$$

$\ell = (1+\varepsilon)k$

Π_{B_k}: projection onto top k sing. vectors of B

Row Sampling



rows of B are "interpretable"

- do not sample uniform
- norm-squared sampling

  sample a_i w.p. $p_i \propto \|a_i\|^2$

  $p_i = \frac{\|a_i\|^2}{\|A\|_F^2}$

$\|A - AB\|_F^2 \le \|A - A_k\|_F^2 + \varepsilon \|A\|_F^2$

$\ell = (k/\varepsilon^2) \log \frac{1}{\delta}$ w/ δ prob fail

time $O(nd \log d)$

$\|A - AT_B\|_F^2 \le (1+\varepsilon)\|A - A_k\|_F^2$

- leverage score sampling

  $s_i(a_i) = \|v_k^T a_i\|^2$

Reservior Sampling

Sampling 1 item from stream proportional to weight w_i

$A = \langle a_1, a_2, \ldots a_t, \ldots a_n \rangle$

$w_1 \quad w_2 \quad \ldots w_t \quad \ldots w_n$

$\rightarrow W_t = \sum_{i=1}^{t} w_i$

$W_t = w_1$

$b \leftarrow a_1$

for ($a_i \in A$)

$W_i = W_{i-1} + w_i$

$v \sim Unif(0,1)$

if ($v > \frac{w_i}{W_i}$)

do nothing

else

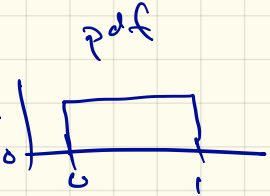
$b \leftarrow a_i$

Row Sampling

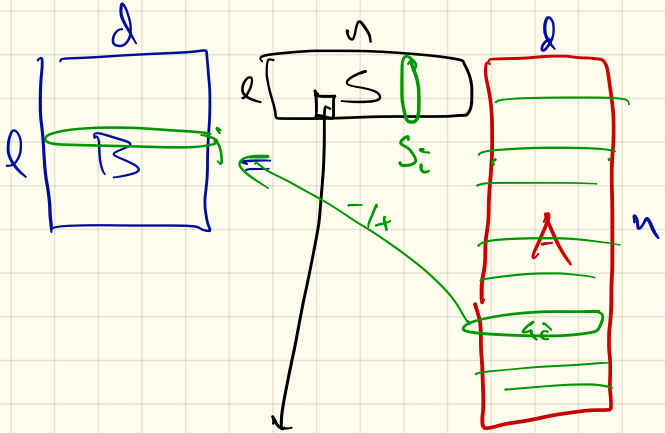
$w_i = \|a_i\|^2$

$v \sim Unif(0,1)$

random float in $[0,1)$



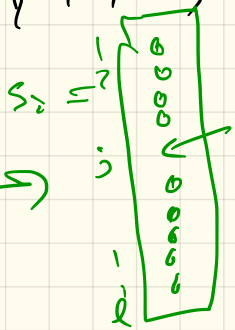
Count Sketch → time $O(nd)$ if $nnz(A)$ = number of non-zeros, time $O(nnz(A))$



JL

$S_{ij} \sim N(0,1)$

$S_{ij} = Unif\{-1, 0, +1\}$



Count sketch →

$k = O(d/\varepsilon^2)$

$\forall x \in \mathbb{R}^d$

$(1-\varepsilon) \le \frac{\|Ax\|}{\|Sx\|} \le (1+\varepsilon)$

$k = d^2/\varepsilon^2$ Count sketch

$\approx 10,000$