L16 : Matrix Sketching

10.2.2025

Data Mining

Jeff M. Phillips

Eigendecomposition

Input: $M \in \mathbb{R}^{d \times d}$ assume positive definite (semi)

eigenvector $v \in \mathbb{R}^d$ eigenvalue $\lambda \in \mathbb{R}$

$$Mv = \lambda v$$

d eigenvector, value pairs

$$(v_1 \lambda_1 , v_2 \lambda_2 , \dots v_d \lambda_d)$$

$$\langle v_i , v_j \rangle = 0 \qquad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$$

p.d: $>$; p.s.d.: $\geq$

$M = A^T A \qquad A \in \mathbb{R}^{n \times d} \qquad n > d$

$\hookrightarrow M$ is p.s.d.

$\text{rank}(A) = d \implies M$ p.d.

say $\text{svd}(A) = U S V^T$

$$MV = A^T A V = (V S^T \cancel{U^T})(\cancel{U} S \cancel{V^T}) \cancel{V}$$

$$= V S^T S$$

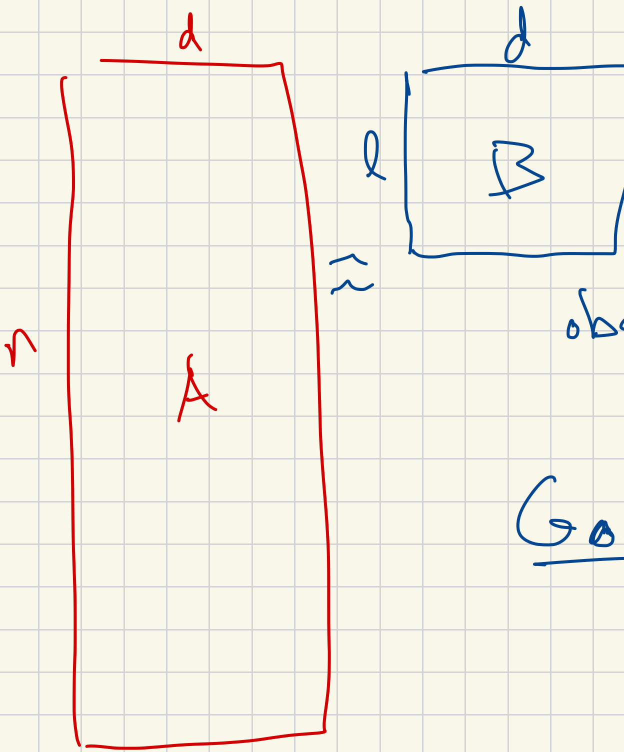
$$= V \,\text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$$

$\forall\, v_j$

$$M v_j = v_j \sigma_j^2$$

$$\sigma_j^2 = \lambda_j$$

Matrix Sketching in a Stream



n is very very large

d is large

observe 1 row a_i at a time

Goal: output $B \in \mathbb{R}^{\ell \times d}$

$$B_k \approx \text{svd}(B) = U' S' V'^T$$

$$v_1' \dots v_k' \approx v_1 \dots v_k$$

$$\sigma_1' \dots \sigma_k' \approx \sigma_1 \dots \sigma_k$$

$$USV^T = \text{svd}(A)$$

Covariance Matrix Summation

$n = 100$ million $\qquad d = 100$

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

$$M = \underbrace{\sum_{i=1}^{n} a_i^T a_i}$$

a_i = row vector

$\underbrace{a_i^T a_i}_{\text{rank } 1} \in \mathbb{R}^{d \times d}$

$C = \text{Zeros}_{d \times d}$

for $i = 1$ to n

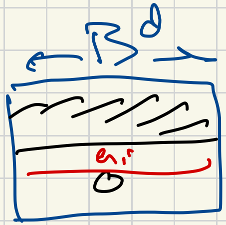
$C = C + a_i^T a_i$

return $M = C$

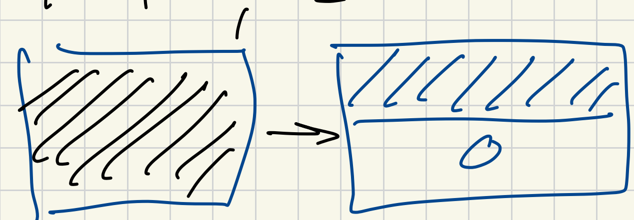
Frequent Directions

$n = 100$ mil $\quad d = 100$ floats

$k = 10 \quad 2$



when full



shrink op $(B) \leftarrow \mathbb{R}^{\ell \times d}$

$\ell \geq 2k$

$\ell = k/\varepsilon + k$

(ignore U)

$svd(B) = U \underline{S} \underline{V}^T$

set $S_{jj} = 0 \quad j \geq \ell/2$

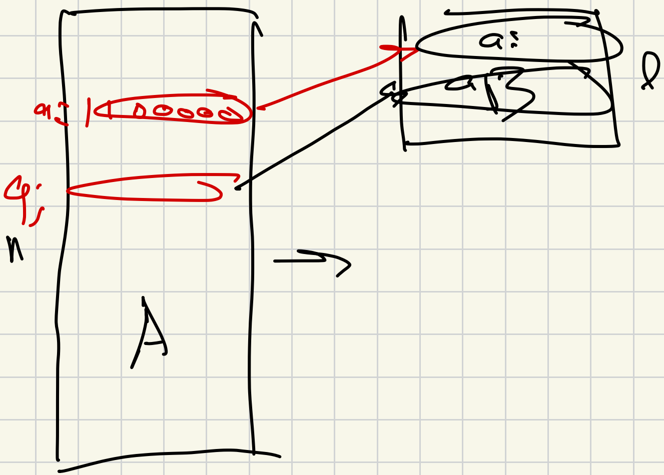
set $S_{jj} = \sqrt{S_{jj}^2 - S_{\ell/2}^2}$ if $j < \ell/2$

$B = S' V^T$

Misra-Gries

$$\| A - A \pi_{B_k} \|_F^2 \leq (1+\varepsilon) \| A - A_k \|_F^2$$

Row Sampling



$\ell << n$

$rsv(B) \approx rsv(A)$

sample ℓ rows of A

$a_i, a_{i'}, \ldots a_\ell$

put into sketch B.

proportional to a weight w_i

- $w_i = \|a_i\|^2$
- $w_i =$ leverage-score$_{(\text{rank-}k)}$ (a_i)

$\|U_k(i)\|^2$

$(1+\varepsilon)$-approx

$\ell \approx \frac{k^2}{\varepsilon^2}$

Count Sketch

t hash functions $\quad t = \log \frac{1}{\delta}$

$h_j : [n] \to [\ell]$

$s_j : [n] \to \{-, +\}$

$B \in \mathbb{R}^{\ell \times d}$ = all 0s.

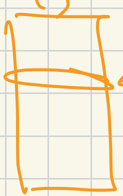
for $i = 1$ to n

add or subtract at random

$B_{\boxed{row = h_j[i]}} \mathrel{+}= s_j(i)\, a_i$

pick random row in sketch

B row of B



row $h[i]$

$\ell = \frac{k^2}{\varepsilon^2}$

For any $x \in \mathbb{R}^d$

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

das coefficients

model $\quad a_i \cdot x \approx y_i$