Matrix Sketching

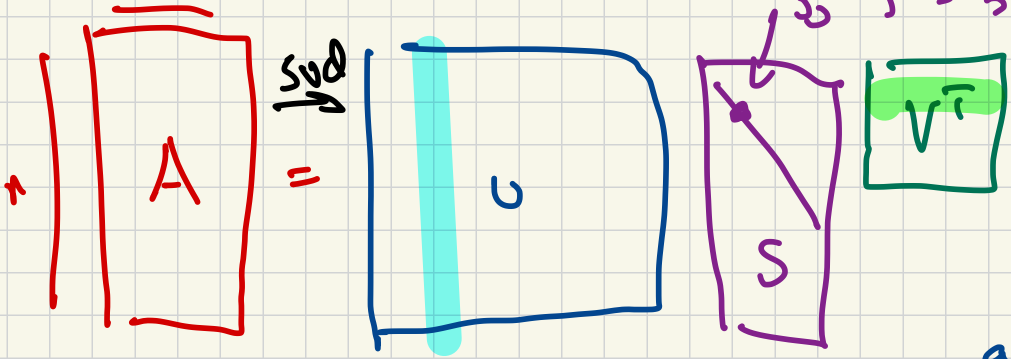
Mar 24, 2025

Data Mining

Jeff M. Phillips

Input $A \in \mathbb{R}^{n \times d}$ rows $\{a_1, a_2, \dots, a_n\} \subset \mathbb{R}^d$

$S_{jj} = \sigma_j$ sing. value



← right sing vector $v_j \in \mathbb{R}^d$

↑ left sing. vectors $u_j \in \mathbb{R}^n$

$$A_k = \underset{\substack{B \\ \mathrm{rank}(B)=k}}{\arg\min} \|A-B\|_2 \quad \text{or} \quad \|A-B\|_F$$

$$A = \sum_{j=1}^{d} u_j \sigma_j v_j^T \qquad A_k = \sum_{j=1}^{k} u_j \sigma_j v_j^T$$

$$\max_{\|x\|=1} \|(A-A_k)x\| = \|A-A_k\|_2 = \sigma_{k+1}$$

$$\|A-A_k\|_F^2 = \sum_{j=k+1}^{d} \sigma_j^2$$

How to compute A_k?

↳ call "svd" language

↳ LAPACK

2021 Jack Dongarra Turing Award.

runtime $O(nd \cdot \min\{n, d\}) = O(nd^2)$

Streaming see data in order

space $\ll n \cdot d$. $n \gg d$

$\underbrace{a_1 \in \mathbb{R}^d}_{d \text{ space}}, \quad a_2 \in \mathbb{R}^d, \quad a_3 \quad \ldots$

Setting

$n = 100$ million $\qquad d = 100 - 1000.$

Summed Covariance

space d^2

$\boxed{C \in \mathbb{R}^{d \times d}} \qquad C_{j,l} = 0$

for a_i $i = 1$ to n

$\quad C = C + a_i a_i^T$

$\mathrm{eig}(C) = V, L$

$V_k \rightarrow$ ~~top~~ top k RSV

$\sqrt{L} \rightarrow$ sing values

$C = AA^T$

$\mathrm{eig}(C) = \mathrm{svd}(A)^2$

Setting

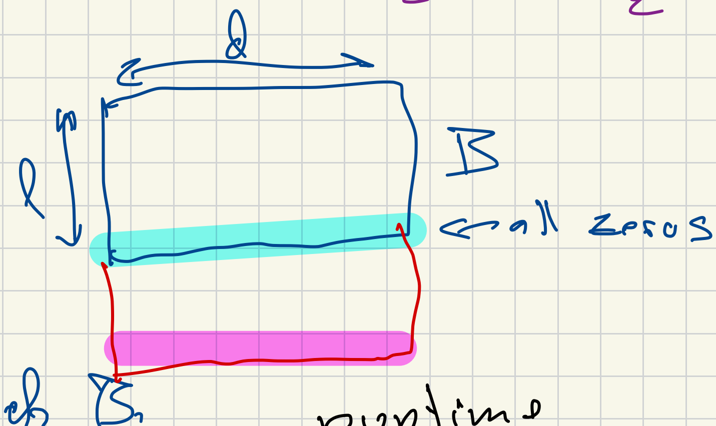
$n = 100$ million $\qquad d = 10k - 100k$

wants $k = 2, 10, 50$ $\qquad d^2$ too big

Need approximation parameter $\ell \approx \frac{1}{\varepsilon}$, $k + \frac{1}{\varepsilon}$

Frequent Directions

Initializing $B \in \mathbb{R}^{2\ell \times d}$



for a_i s.t. $i = 1$ to n

  insert a_i into empty row of B

  if (no all zero row)

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

    $\delta = \sigma_\ell^2$

    $S' = \mathrm{diag}\left(\sqrt{\sigma_1^2 - \delta}, \sqrt{\sigma_2^2 - \delta}, \sqrt{\sigma_3^2 - \delta} \cdots 0, 0, 0 \ldots 0\right)$

    $B = S' V^T$

return B.

runtime

$O\left(n/\ell \cdot \underbrace{(2\ell)^2 \cdot d}_{\text{svd}}\right)$

$= O(n d \cdot \ell)$

$$FD(A) \to B \qquad B \in \mathbb{R}^{\ell \times d}$$

(1) $\forall x \in \mathbb{R}^d \quad \|x\| = 1$

$$0 \le \|Ax\|^2 - \|Bx\|^2 \le \frac{\|A - A_k\|_F^2}{\ell - k}$$

$\ell = k + \frac{1}{\varepsilon}$

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

(2)

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

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

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

For Regression

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

Does not hold for FD.

Setting

Want: ① Have sparsity, want to maintain

$A \to B$

size $\neq n \cdot d \Rightarrow nnz(A)$

number of non-zeros.

② want to maintain example rows.

---

sample ℓ rows from $A \to$ as rows of B.

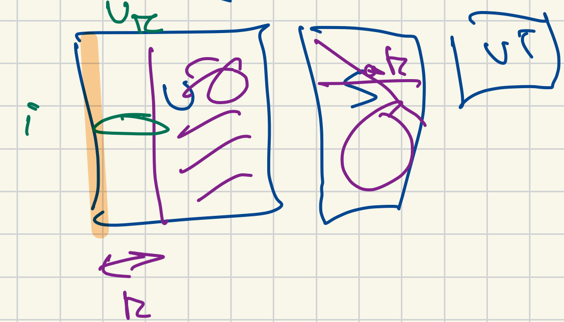
(weighted) Reservoir sampling?

row weight $w_i = \|a_i\|^2$

$\to \|A - \pi_B(A)\|_F \leq \|A - A_k\|_F + \varepsilon \|A\|_F$

$\ell = O\left(\frac{k}{\varepsilon^2}\right)$

Better Row Sampling w/ Leverage Scores



$A \xrightarrow{svd} U S V^T$

$lev(a_i) = \|U_k(i)\|^2$

if sample $\ell = O\left(\frac{k}{\varepsilon^2}\right) \sim lev(a_i)$

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

Random Projection

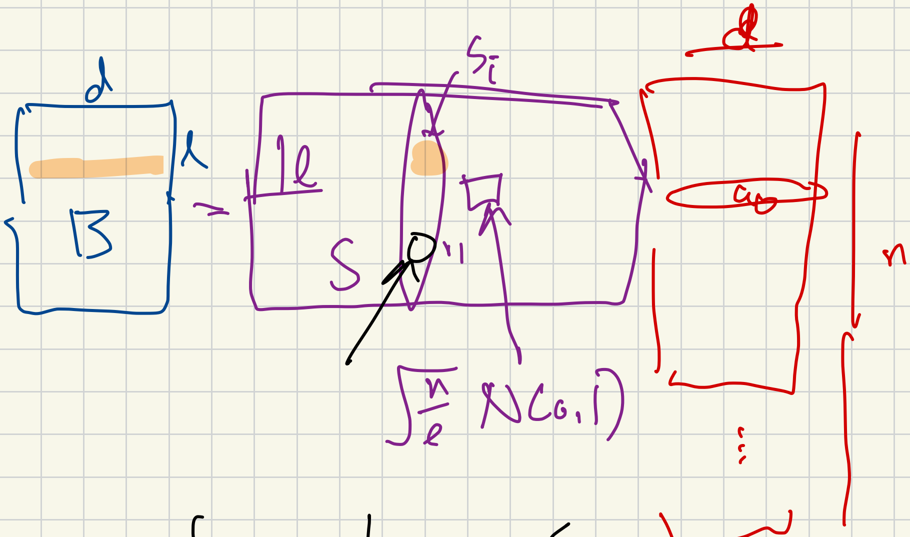
$B = SA$

predefine $S \in \mathbb{R}^{l \times n}$

$S_{ij} \sim_{iid} N(0,1)\sqrt{\frac{n}{l}}$

$l = \frac{d}{\varepsilon^2}$

$l << n$



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

each column S have 1 non-zero

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

Count Sketch

$l = \frac{d^2}{\varepsilon^2}$, sparsity $O\left(nnz(A) + \frac{1}{\varepsilon^2}\right)$

Count Sketch Alg.

$\ell = O\left(\frac{d^2}{\varepsilon^2}\right)$

Init $B = \mathbb{R}^{\ell \times d}$

$\|Bx\| = \sqrt{\sum_{j=1}^{d} \langle b_j, x\rangle^2}$

for a_i $i = 1$ to n

choose $j \sim \mathrm{Unif}(1, \ell)$

choose $s_i \sim \mathrm{Unif}(\{-1, +1\})$

jth row of B

$b_j = \underline{b_j + s_i \cdot a_i}$

return B.

Guarantee $\forall x$

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