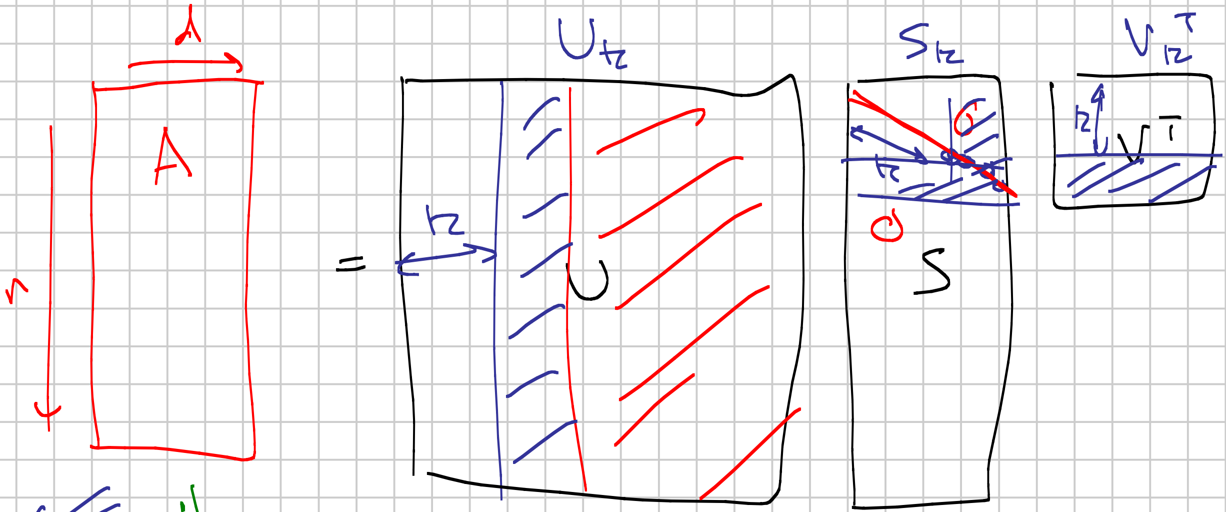
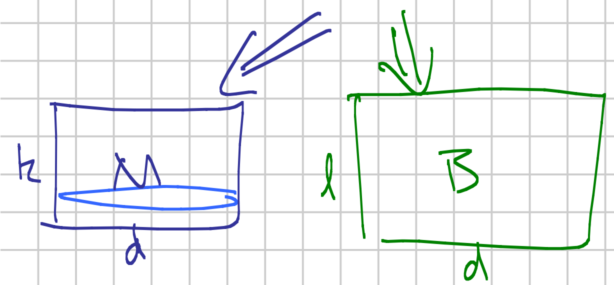
Matrix Sketching

Note Title 3/9/2016

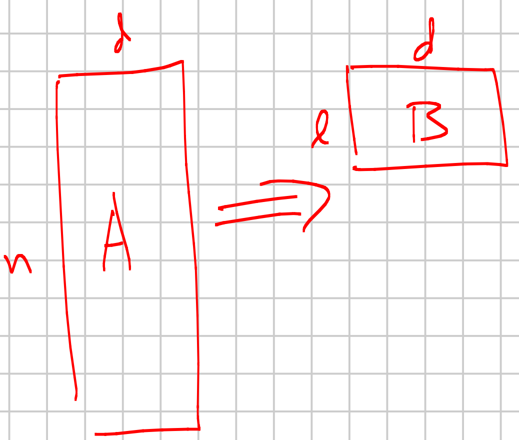




$A_k = U_k S_k V_k^T$

$S_k V_k^T = k \times d \Rightarrow M$

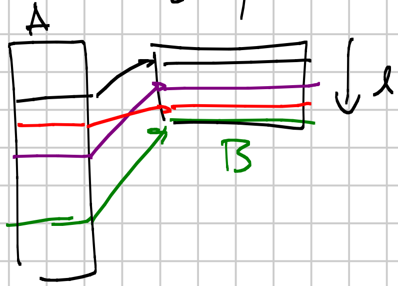
- slow for large n and d
  svd(A) takes $O(\min\{nd^2, n^2d\})$
- interpretability
  ↳ can we use actual rows of A?
- non-centralized computation
  + in a stream
  + data too big for one machine?



- Row Sampling
- Iterative (Frequent Directions)
  Misra-Gries
- Random Projections

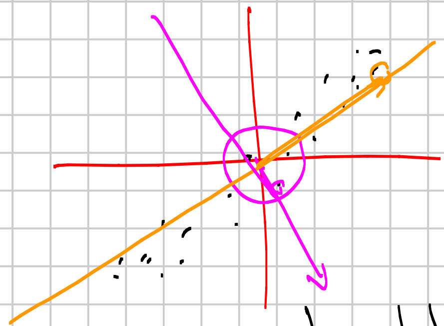
Row Sampling

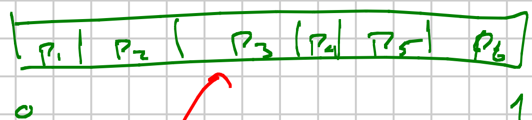
Have $\ell = O\left(\frac{k}{\varepsilon^2}\right)$ rows independently



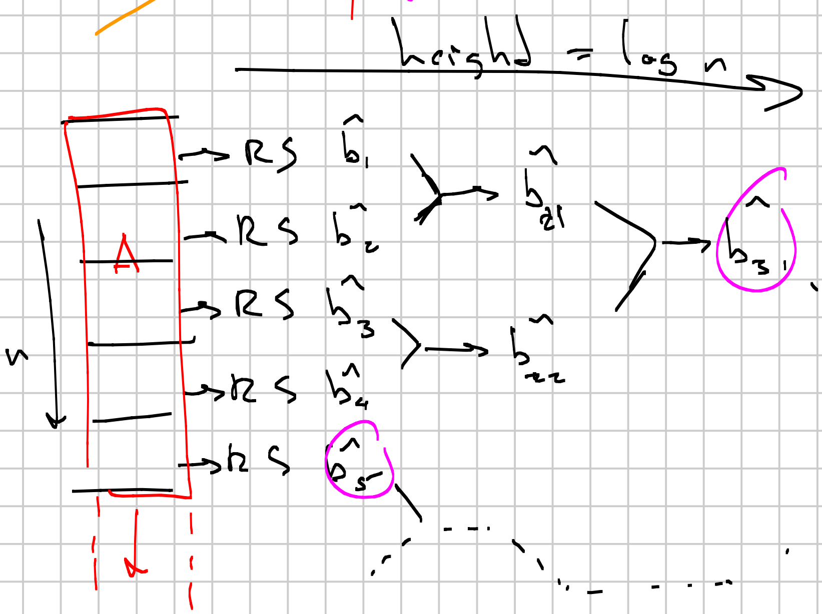
$a_i \in A$ row $\quad w_i = \|a_i\|^2$

sample proportional to $w_i \to p_i = \frac{w_i}{W}$





$u \leftarrow \mathrm{Unif}(0,1)$



Reservoir Sampling

$$W_i = \sum_{j=1}^{i} w_j$$

$b \leftarrow$ sample

see $a_i \leftarrow$ row (A)

with prob $\frac{w_i}{W_i}$ set $b = a_i$

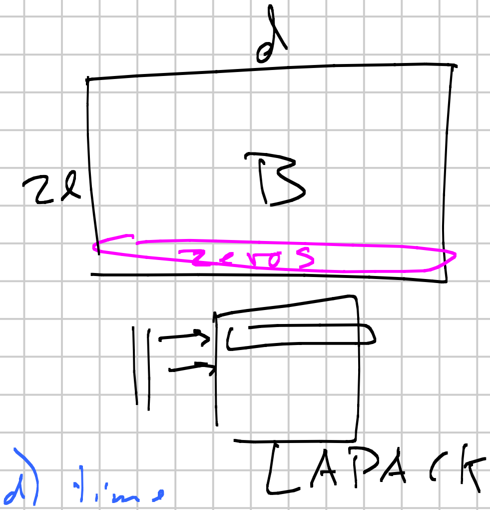
Priority Sampling

$$\underbrace{\|A - \pi_B(A)\|_F^2}_{\text{Projection of } A \text{ onto } B} \le \|A - \underbrace{A_k}_{\text{projection of } A \text{ onto } V_k}\|_F^2 + \varepsilon \underbrace{\|A\|_F^2}$$

$$\|A\|_F^2 = \sum_{i,j} (A_{i,j})^2$$

Frequent Directions

$B \leftarrow \text{zeros}(2\ell \times d)$ $O(nd \cdot \ell)$



for ($a_i \in \text{rows}(A)$)

- Insert a_i into B
- if (no all zero rows in B)
  - $[U, S, V^T] = \text{svd}(B)$ $O(\ell^2 d)$ time
  - Set $\delta_i = s_\ell^2$ ← ℓth sing. value of B, squared
  - $S' \leftarrow \text{diag}\left(\sqrt{s_1^2 - \delta_i}, \sqrt{s_2^2 - \delta_i}, \ldots, \sqrt{s_{\ell-1}^2 - \delta_i}, 0, \ldots 0\right)$ ← shrinks $\|B\|_F^2 \geq \ell \cdot \delta_i$
  - $B \leftarrow S' \cdot V^T$

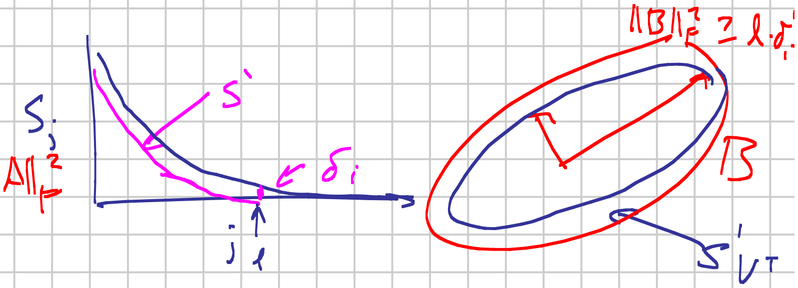
$S \in$ sing val of A

$S = s_1, s_2, \ldots$

$\|A\|_F^2 = \sum_{i=1}^{r} s_i^2$

$\Delta = \sum_{i=1}^{} \delta_i$

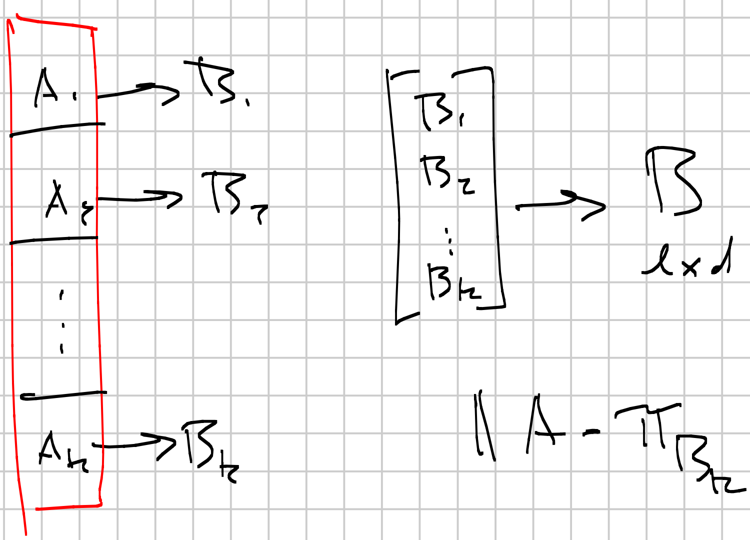
total shrink $\geq \Delta \cdot \ell \leq \|A\|_F^2$



$\forall$ all $x \in \mathbb{R}^d$ $\|x\| = 1$

$$0 \leq \|Ax\|^2 - \|Bx\|^2 \leq \Delta \leq \frac{\|A\|_F^2}{\ell} \quad \left(\leq \frac{\|A - A_k\|_F^2}{\ell - k}\right)$$

$\ell = k + 1/\varepsilon$ $\quad \leq \varepsilon \cdot \|A - A_k\|_F^2$



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

$\ell = k + k/\varepsilon$