

CSE 392: Matrix and Tensor Algorithms for Data

Instructor: Shashanka Ubaru

University of Texas, Austin
Spring 2024

Lecture 6: Approximate matrix product and sampling

Outline

- 1 Randomization
- 2 Approximating Matrix Multiplication
- 3 Length-squared sampling
- 4 Leverage score sampling

Why randomization?

- *Modern data applications*: massive data, computationally expensive problems.
- *Approximate solutions* suffice in many situations.
- **Randomized sampling and sketching** allow us to design approximation algorithms with provable error guarantees.
- Probabilistic error bounds. E.g., the (ϵ, δ) type bounds.

Product and norms using randomization

If a random distribution on $\mathbf{s} \in \mathbb{R}^n$ has entries s_i with:

- $\mathbb{E}[s_i^2] = 1$ for $i = [n]$ and $\mathbb{E}[s_i s_j] = 0$ for $i, j = [n], i \neq j$.
- Then, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have

$$\mathbb{E}[\langle \mathbf{s} \cdot \mathbf{x}, \mathbf{s} \cdot \mathbf{y} \rangle] = \mathbb{E}[(\mathbf{s}^\top \mathbf{x}) \cdot (\mathbf{s}^\top \mathbf{y})] = \mathbb{E}[\mathbf{x}^\top \mathbf{s} \mathbf{s}^\top \mathbf{y}] = \mathbf{x}^\top \mathbf{y}$$

- In particular, $\mathbb{E}[\|\mathbf{s}^\top \mathbf{y}\|^2] = \mathbb{E}[\mathbf{y}^\top \mathbf{s} \mathbf{s}^\top \mathbf{y}] = \mathbf{y}^\top \mathbf{y} = \|\mathbf{y}\|^2$.

$$\mathbb{E}[\mathbf{s} \mathbf{s}^\top] = \begin{bmatrix} s_1^2 & s_1 s_2 & \cdots & s_1 s_n \\ s_2, s_1 & s_2^2 & & \vdots \\ \vdots & & \ddots & \\ s_n, s_1 & \cdots & & s_n^2 \end{bmatrix} = \mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & 1 \end{bmatrix}$$

Sketching and Sampling

Sketching:

- Suppose $s_i \sim \mathcal{N}(0, 1)$ and independent.
- We have $\mathbb{E}[s_i] = 0$, $\mathbb{E}[s_i^2] = \text{Var}(s_i) = 1$.
- For $i \neq j$, independence implies $\mathbb{E}[s_i s_j] = \mathbb{E}[s_i] \mathbb{E}[s_j] = 0$.

Sampling:

- Suppose we pick $i \in [n]$ uniformly with probability $\frac{1}{n}$ and set $s_i \leftarrow \sqrt{n}$, 0 o.w.
- We have $\mathbb{E}[s_i^2] = \frac{1}{n} \sqrt{n^2} + (1 - \frac{1}{n}) 0 = 1$.
- For $i \neq j$ if $s_i \neq 0 \implies s_j = 0$,
so $s_i s_j = 0$.

Randomized techniques

With repetitions and better distributions, randomization can be made highly accurate.

A random distribution on $\mathbf{S} \in \mathbb{R}^{c \times n}$ has independent rows, each row is $\frac{1}{\sqrt{c}}$ times a sample of $\mathbf{s} \in \mathbb{R}^n$, then

$$\mathbb{E}[\mathbf{S}^\top \mathbf{S}] = \mathbb{E}\left[\sum_{i \in [c]} \mathbf{s}_{i*}^\top \mathbf{s}_{i*}\right] = \sum_{i \in [c]} \mathbb{E}[\mathbf{s}_{i*}^\top \mathbf{s}_{i*}] = \sum_{i \in [c]} \frac{1}{c} \mathbf{I} = \mathbf{I},$$

so for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have $\mathbb{E}[\langle \mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y} \rangle] = \mathbb{E}[\mathbf{x}^\top \mathbf{S}^\top \mathbf{S} \mathbf{y}] = \mathbf{x}^\top \mathbb{E}[\mathbf{S}^\top \mathbf{S}] \mathbf{y} = \mathbf{x}^\top \mathbf{y}$.
In particular, $\mathbb{E}[\|\mathbf{S}\mathbf{y}\|^2] = \|\mathbf{y}\|^2$

Applications:

- Approximating matrix multiplication
- Least squares regression
- Low rank approximation

Approximating Matrix Multiplication (AMM)

Problem Statement:

Given an $m \times n$ matrix \mathbf{A} and an $n \times p$ matrix \mathbf{B} , approximate the product $\mathbf{A} \cdot \mathbf{B}$,

OR, equivalently,

Approximate the sum of n rank-one matrices.

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^n \underbrace{\begin{bmatrix} \mathbf{A}_{*i} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{B}_{i*} \end{bmatrix}}_{m \times p}$$

where \mathbf{A}_{*i} is the i th column of \mathbf{A} and \mathbf{B}_{i*} is the i th row of \mathbf{B} .

Sampling rows of a matrix

- If $\mathbf{S} \in \mathbb{R}^{c \times n}$ is a random row sampling matrix, then \mathbf{SA} :

$$\begin{bmatrix} 0 & s_{12} & 0 & 0 & \cdots & 0 \\ s_{21} & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & s_{33} & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \cdots & s_{cn} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{1*} \\ \mathbf{A}_{2*} \\ \vdots \\ \mathbf{A}_{n*} \end{bmatrix} = \begin{bmatrix} s_{12} \mathbf{A}_{2*} \\ s_{21} \mathbf{A}_{1*} \\ s_{33} \mathbf{A}_{3*} \\ \vdots \\ s_{cn} \mathbf{A}_{n*} \end{bmatrix}$$

- As above, for a single sampling vector \mathbf{s} , *uniform sampling* would pick $i \in [n]$ uniformly with probability $\frac{1}{n}$ and set $s_i \leftarrow \sqrt{n}$.
- Generally, given $\mathbf{p} \in [0, 1]^n$, $\sum_i p_i = 1$. Pick $i \in [n]$ with probability p_i , $s_i \leftarrow \sqrt{1/p_i}$. We have $\mathbb{E}[s_i^2] = p_i \sqrt{1/p_i}^2 + (1 - p_i)0 = 1$.
- In some instances, by choosing appropriate p_i 's, we can get improved results.

$$\begin{aligned}
 \mathbf{A} \cdot \mathbf{B} &= \sum_{i=1}^n \underbrace{\begin{bmatrix} \mathbf{A}_{*i} \end{bmatrix}}_{m \times p} \cdot \begin{bmatrix} \mathbf{B}_{i*} \end{bmatrix} \\
 &\approx \frac{1}{c} \sum_{t=1}^c \frac{1}{p_{j_t}} \underbrace{\begin{bmatrix} \mathbf{A}_{*j_t} \end{bmatrix}}_{m \times p} \cdot \begin{bmatrix} \mathbf{B}_{j_t*} \end{bmatrix}
 \end{aligned}$$

Pick c terms of the sum, with replacement, with respect to the p_i 's. I.e. set $j_t = i$, where $\Pr(j_t = i) = p_i$.

$$\underbrace{\begin{bmatrix} A \end{bmatrix}}_{m \times n} \cdot \underbrace{\begin{bmatrix} B \end{bmatrix}}_{n \times p} \approx \underbrace{\begin{bmatrix} C \end{bmatrix}}_{m \times c} \cdot \underbrace{\begin{bmatrix} R \end{bmatrix}}_{c \times p}$$

- We would like to estimate $AB \approx AS^\top SB$.
- Suppose S has just one row s_i . Then, we just get $A_{i*}s_i^2 B_{*i} = A_{*i}B_{i*}/p_i$ with probability p_i .
- If we pick uniformly with $p_i = 1/n$, and suppose one of the row norms $\|B_{1*}\|^2$ is much \gg norms of other rows, then the estimate will be poor, if we miss the row $i = 1$.
- One idea : catch the rows with large norms by setting $p_i \propto \|B_{1*}\|^2$. This is called **Length-squared sampling**.

$$\underbrace{\begin{bmatrix} A \end{bmatrix}}_{m \times n} \cdot \underbrace{\begin{bmatrix} B \end{bmatrix}}_{n \times p} \approx \underbrace{\begin{bmatrix} C \end{bmatrix}}_{m \times c} \cdot \underbrace{\begin{bmatrix} R \end{bmatrix}}_{c \times p}$$

- Create C and R by picking columns A_{*j_t} and rows B_{j_t*} with probability

$$\Pr(j_t = i) = \frac{\|A_{*i}\|_2 \|B_{i*}\|_2}{\sum_{j=1}^n \|A_{*j}\|_2 \|B_{j*}\|_2}$$

- Include $A_{*j_t}/\sqrt{cp_{j_t}}$ as a column of C , and $B_{j_t*}/\sqrt{cp_{j_t}}$ as a row of R .

Length-squared sampling

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times p}$. Let $\mathbf{S} \in \mathbb{R}^{c \times n}$ be the length squared sampling matrix. Then, $\mathbb{E}[\mathbf{C}\mathbf{R}] = \mathbf{A}\mathbf{B}$ (unbiased estimator), where $\mathbf{C} = \mathbf{A}\mathbf{S}^\top$, $\mathbf{R} = \mathbf{S}\mathbf{B}$, and

$$\mathbb{E}[\|\mathbf{C}\mathbf{R} - \mathbf{A}\mathbf{B}\|_F^2] \leq \frac{1}{c} \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2$$

Length-squared sampling

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times p}$. Let $\mathbf{S} \in \mathbb{R}^{c \times n}$ be the length squared sampling matrix. Then, $\mathbb{E}[\mathbf{CR}] = \mathbf{AB}$ (unbiased estimator), where $\mathbf{C} = \mathbf{AS}^\top$, $\mathbf{R} = \mathbf{SB}$, and

$$\mathbb{E}[\|\mathbf{CR} - \mathbf{AB}\|_F^2] \leq \frac{1}{c} \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2$$

Proof: First, for any probability p_i , we know that $\mathbb{E}[\mathbf{CR}_{ij}] = \mathbf{AB}_{ij}$. Elementwise is an unbiased estimator.

Next, note that for a single vector \mathbf{s} , $\mathbb{E}[\|\mathbf{Ass}^\top \mathbf{B} - \mathbf{AB}\|_F^2]$ is the sum of entry-wise variances.

Since $\text{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$, we have $\mathbb{E}[\|\mathbf{Ass}^\top \mathbf{B} - \mathbf{AB}\|_F^2] \leq \mathbb{E}[\|\mathbf{Ass}^\top \mathbf{B}\|_F^2]$

$$\begin{aligned}
\mathbb{E}[\|\mathbf{A}\mathbf{s}\mathbf{s}^\top\mathbf{B}\|_F^2] &= \sum_{j,k} \mathbb{E}[(\mathbf{A}_{j*}\mathbf{s}\mathbf{s}^\top\mathbf{B}_{*k})^2] = \sum_{j,k} \mathbb{E}[(\sum_i a_{ji}s_i^2 b_{ik})^2] \\
&= \sum_{j,k} \sum_i a_{ji}^2 p_i \frac{1}{p_i^2} b_{ik}^2 = \sum_i \sum_j a_{ji}^2 \frac{1}{p_i} \sum_k b_{ik}^2 = \sum_i \|\mathbf{A}_{*i}\|^2 \frac{1}{p_i} \|\mathbf{B}_{i*}\|^2 \\
&= \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2.
\end{aligned}$$

Next, for the case of c rows, the expected Frobenius norm error is sum of variance of the form

$$\text{Var}[\sum_{i \in [c]} \mathbf{x}^{(i)} / c] = \sum_{i \in [c]} \text{Var}[\mathbf{x}^{(i)} / c] = \text{Var}[\mathbf{x}^{(1)}] / c.$$

Thus, we get the result

$$\mathbb{E}[\|\mathbf{C}\mathbf{R} - \mathbf{A}\mathbf{B}\|_F^2] \leq \frac{1}{c} \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2.$$

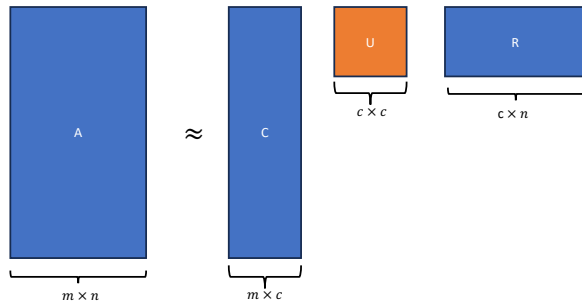
Using Markov's inequality, we can show that for $c \geq 1/\epsilon^2\delta$,

$$\Pr(\|\mathbf{C}\mathbf{R} - \mathbf{A}\mathbf{B}\|_F \geq \epsilon \|\mathbf{A}\|_F \|\mathbf{B}\|_F) \leq \delta.$$

CUR decomposition

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, a particular type of low rank approximation:

- A row sampling matrix $\mathbf{S}_1 \in \mathbb{R}^{c \times m}$, and $\mathbf{R} = \mathbf{S}_1 \mathbf{A} \in \mathbb{R}^{c \times n}$
- A column sampling matrix $\mathbf{S}_2 \in \mathbb{R}^{n \times c}$, and $\mathbf{C} = \mathbf{A} \mathbf{S}_2 \in \mathbb{R}^{m \times c}$
- A matrix $\mathbf{U} \in \mathbb{R}^{c \times c}$, such that $\mathbf{A} \approx \mathbf{C} \mathbf{U} \mathbf{R}$ and $c \ll \{m, n\}$.



CUR decomposition

- We can compute $\mathbf{U} = (\mathbf{A}\mathbf{S}_2)^\dagger \mathbf{S}_1^\top = (\mathbf{C}^\top \mathbf{C})^{-1} (\mathbf{S}_1 \mathbf{A} \mathbf{S}_2)^\top$.
- \mathbf{U} can be ill-conditioned.
- Typically, in applications, we are interested in random columns \mathbf{C} and rows \mathbf{R} of \mathbf{A} .
- We can also consider, $\mathbf{S}_1 \in \mathbb{R}^{r \times m}$ and $\mathbf{S}_2 \in \mathbb{R}^{n \times c}$, for different c, r .

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, row sampler $\mathbf{S}_1 \in \mathbb{R}^{r \times m}$, column $\mathbf{S}_2 \in \mathbb{R}^{n \times c}$, and with $\mathbf{C} = \mathbf{A}\mathbf{S}_2$, $\mathbf{R} = \mathbf{S}_1 \mathbf{A}$, $\mathbf{U} = (\mathbf{A}\mathbf{S}_2)^\dagger \mathbf{S}_1^\top$, then

$$\mathbb{E}[\|\mathbf{CUR} - \mathbf{A}\|_F^2] \leq 2\|\mathbf{A}\|_F^2 \left(\frac{1}{\sqrt{c}} + \frac{c}{r} \right) \leq \epsilon \|\mathbf{A}\|_F^2,$$

for $c = 16/\epsilon^2, r = 64/\epsilon^3$.

Matrix (low rank) approximations

- We can also consider sampling only the columns as $\mathbf{A} \approx \mathbf{C}\mathbf{X}$, or
- Sample only the rows $\mathbf{A} \approx \mathbf{X}\mathbf{R}$.
- More flexible structure can give better-conditioned \mathbf{X} .
- We need fast decaying spectrum.
- For

$$\Pr(\|\mathbf{C}\mathbf{U}\mathbf{R} - \mathbf{A}\|_F \geq \epsilon \|\mathbf{A}\|_F) \leq \delta,$$

we need $c = O(\delta^{-2}\epsilon^{-4})$, $r = O(\delta^{-3}\epsilon^{-6})$.

- Cost =?

Better variance reduction

- We want \mathbf{S} such that $\|\mathbf{S}\mathbf{A}\mathbf{x}\|$ is a good estimator of $\|\mathbf{A}\mathbf{x}\|$.
- Length-squared sampling : $p_i \propto \|\mathbf{A}_{i*}\|^2$ is good, but for some \mathbf{x} , we could have $\mathbf{A}_{i*}\mathbf{x} = 0$ even if $\|\mathbf{A}_{i*}\|^2$ is large.
- We want $(\frac{1}{\sqrt{p_i}}\mathbf{A}_{i*}\mathbf{x})^2$ to be “well-behaved” for all i and \mathbf{x} .
- “well-behaved” in one sense : bounded relative contribution to $\|\mathbf{A}\mathbf{x}\|^2 = \sum_i (\mathbf{A}_{i*}\mathbf{x})^2$.
- sampling using information related to $\text{span}(\mathbf{A})$.

Leverage scores

- **Leverage scores:** Given a linear subspace $L \subset \mathbb{R}^n$, for $i \in [n]$, the i th *leverage score* $\ell_i(L) = \sup_{\mathbf{y} \in L} y_i^2 / \|\mathbf{y}\|^2$.
- The leverage scores of $\mathbf{A} \in \mathbb{R}^{m \times n}$ are $\ell_i(\mathbf{A}) = \ell_i(\text{span}(\mathbf{A}))$.

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, and an orthonormal basis \mathbf{U} for $\text{span}(\mathbf{A})$, for $i \in [n]$, the i th *leverage score*

$$\ell_i(\mathbf{A}) = \sup_{\mathbf{x}} \frac{(\mathbf{A}_{i*} \mathbf{x})^2}{\|\mathbf{A} \mathbf{x}\|^2} = \|\mathbf{U}_{i*}\|^2.$$

Leverage scores

- **Leverage scores:** Given a linear subspace $L \subset \mathbb{R}^n$, for $i \in [n]$, the i th *leverage score* $\ell_i(L) = \sup_{\mathbf{y} \in L} y_i^2 / \|\mathbf{y}\|^2$.
- The leverage scores of $\mathbf{A} \in \mathbb{R}^{m \times n}$ are $\ell_i(\mathbf{A}) = \ell_i(\text{span}(\mathbf{A}))$.

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, and an orthonormal basis \mathbf{U} for $\text{span}(\mathbf{A})$, for $i \in [n]$, the i th *leverage score*

$$\ell_i(\mathbf{A}) = \sup_{\mathbf{x}} \frac{(\mathbf{A}_{i*} \mathbf{x})^2}{\|\mathbf{A} \mathbf{x}\|^2} = \|\mathbf{U}_{i*}\|^2.$$

For $L = \text{span}(\mathbf{A}) = \text{span}(\mathbf{U})$, and $\mathbf{z} \in L$ has $\mathbf{z} = \mathbf{A} \mathbf{x} = \mathbf{U} \mathbf{y}$ for some \mathbf{x}, \mathbf{y} . So,

$$\sup_{\mathbf{x}} \frac{(\mathbf{A}_{i*} \mathbf{x})^2}{\|\mathbf{A} \mathbf{x}\|^2} = \sup_{\mathbf{y}} \frac{(\mathbf{U}_{i*} \mathbf{y})^2}{\|\mathbf{U} \mathbf{y}\|^2} = \sup_{\mathbf{y}} \frac{(\mathbf{U}_{i*} \mathbf{y})^2}{\|\mathbf{y}\|^2} = \|\mathbf{U}_{i*}\|^2.$$

We have $\ell_i(\mathbf{A}) \in [0, 1]$ and $\sum_i \ell_i(\mathbf{A}) = \text{rank}(\mathbf{A})$.

Leverage score sampling

Leverage score sampling: sample rows with probability proportional to the square of the Euclidean norms of the rows of the left singular vectors of \mathbf{A} .

$$p_i = \frac{\|\mathbf{U}_{i*}\|^2}{\|\mathbf{U}\|_F^2} = \frac{\|\mathbf{U}_{i*}\|^2}{n}$$

Column sampling is equivalent to row sampling by focusing on \mathbf{A}^\top . So, we consider the right singular vectors \mathbf{V} .

$$p_j = \frac{\|\mathbf{V}_{j*}\|^2}{m}.$$

Leverage scores: general case

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and \mathbf{A}_k its best rank- k approximation (as computed by the SVD):

$$\underbrace{\mathbf{A} \approx \begin{bmatrix} \mathbf{A}_k \end{bmatrix}}_{m \times n} \approx \underbrace{\begin{bmatrix} \mathbf{U}_k \end{bmatrix}}_{m \times k} \cdot \underbrace{\begin{bmatrix} \Sigma_k \end{bmatrix}}_{k \times k} \cdot \underbrace{\begin{bmatrix} \mathbf{V}_k^\top \end{bmatrix}}_{k \times n}$$

Row Leverage scores and Column Leverage scores

$$p_i = \frac{\|(\mathbf{U}_k)_{i*}\|^2}{k} \qquad p_j = \frac{\|(\mathbf{V}_k)_{j*}\|^2}{k}$$

Leverage score sampling

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, if we randomly sample the columns $\mathbf{C} \in \mathbb{R}^{m \times c}$ using leverage scores, then, with probability at least 0.9,

$$\|\mathbf{A} - \mathbf{C}\mathbf{X}\|_F = \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\|_F \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F,$$

for sampling complexity

$$c = O\left(\frac{k}{\epsilon^2} \log\left(\frac{k}{\epsilon}\right)\right)$$

Leverage score sampling

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, if we randomly sample the columns $\mathbf{C} \in \mathbb{R}^{m \times c}$ using leverage scores, then, with probability at least 0.9,

$$\|\mathbf{A} - \mathbf{C}\mathbf{X}\|_F = \|\mathbf{A} - \mathbf{C}\mathbf{C}^\dagger \mathbf{A}\|_F \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F,$$

for sampling complexity

$$c = O\left(\frac{k}{\epsilon^2} \log\left(\frac{k}{\epsilon}\right)\right)$$

Proof uses Matrix Chernoff inequality.

Let \mathbf{X}_i for $i \in [m]$ be i.i.d copies of symmetric random $\mathbf{X} \in \mathbb{R}^{n \times n}$ with $\gamma, \sigma^2 > 0$, $\mathbb{E}[\mathbf{X}] = 0$, $\|\mathbf{X}\|_2 \leq \gamma$, and $\|\mathbb{E}[\mathbf{X}^2]\|_2 \leq \sigma^2$. Then for $\epsilon > 0$,

$$\Pr\left(\left\|\frac{1}{m} \sum_i \mathbf{X}_i\right\|_2 \geq \epsilon\right) \leq 2n \exp(-m\epsilon^2/(\sigma^2 + \gamma\epsilon/3)).$$

Further Reading

- Drineas, Petros, Ravi Kannan, and Michael W. Mahoney. “Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication.” *SIAM Journal on Computing* 36.1 (2006): 132-157.
- Drineas, Petros, Ravi Kannan, and Michael W. Mahoney. “Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix.” *SIAM Journal on computing* 36.1 (2006): 158-183.
- Kannan, Ravindran, and Santosh Vempala. “Randomized algorithms in numerical linear algebra.” *Acta Numerica* 26 (2017): 95-135.
- Boutsidis, Christos, and David P. Woodruff. “Optimal CUR matrix decompositions.” *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. 2014.

Questions?