SEEM5020 Algorithms for Big Data
Matrix Sketching: Sampling and Projection

Sibo WANG

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Matrix Data

Matrix data are prevalent in machine learning and data science:

- Multiplication of matrices: Neural networks, multiplying the weight matrix with the input activation;
- SVD, PCA, Low-rank approximation
- ...

However, the running costs are extremely high.

- Matrix multiplication: $A_{m \times n}B_{n \times p}$ incurs $O(m \cdot n \cdot p)$ running time.
- SVD for a matrix $A_{m \times n}$: $O(\min\{mn^2, nm^2\})$.

If approximation is allowed, we can make the computation to be more efficient via matrix sketching: Replace the input matrix $A$ with a more compact matrix $B$ that preserves most of the properties of $A$. 
Popular methods in matrix sketching:

- Data dependent approach: Row/column sampling
- Data independent/oblivious approach: Subspace embedding, e.g., using JL transformation.

We are going to see how these approaches can be applied to facilitate matrix computation.
1 Approximate Matrix Multiplication
   - Random Sampling
   - Applying JL Lemma

2 Low-Rank Approximation
   - Additive Error Guarantee via Random Sampling
   - Additive Error Guarantee via JL Lemma
   - Relative Error Guarantee with Sparse JL Transform
Matrix Multiplication

Problem 1

Given matrices \( A \in \mathbb{R}^{m \times n} \) and \( B \in \mathbb{R}^{n \times p} \), compute the matrix \( AB \).

- Standard iterative approach: \( O(m \cdot n \cdot p) \).
- Faster algorithm via the nontrivial divide-and-conquer approach. For the \( n \times n \) matrix with \( n \times n \) matrix multiplication, the running cost can be reduced to \( O(n^\omega) \). Currently \( \omega \approx 2.37 \).

Problem 2 (Approximate Matrix Multiplication)

Compute a matrix \( D \in \mathbb{R}^{m \times p} \) such that \( \|D - AB\| \) is bounded by appropriate matrix norms related to \( A \) and \( B \).

We will derive the approximation result \( D \) such that:

\[
\|D - AB\| \leq \varepsilon \|A\| \cdot \|B\|
\]

for some norm \( \| \cdot \| \). We will focus on the Frobenius norm.
Let us define $A_{(*,i)}$ be the $i$-th column and $B_{(i,*)}$ be the $i$-th row of $B$. Then, we have:

$$AB = \sum_{i=1}^{n} A_{(*,i)} B_{(i,*)}.$$ 

If we sample $w$ from $[1\cdots n]$ uniformly at random and take the $w$-th column of $A$ and the $w$-th row of $B$. Then, let $D = A_{(*,w)} B_{(w,*)}$ we have that:

$$\mathbb{E}[D] = \frac{1}{n} AB.$$ 

However, we can not easily bound the variance of such a uniform sampling approach. Alternative: Importance sampling.
More generally, we can draw each column $w$ with probability $p_w$, such that $p_1 + p_2 + \cdots + p_n = 1$. Then, for an arbitrary sample $w$ following this distribution $[p_1, p_2, \cdots, p_n]$, we can provide an unbiased estimation of $AB$:

$$D = \frac{A(\ast, w)B(w, \ast)}{p_w},$$

$$\mathbb{E}[D] = AB.$$

For example, if we set $p_i = \frac{1}{n}$, we derive the same conclusion as on the last page. By drawing $t$ independent samples, and let $D_i$ be the $i$-th matrix we computed. We can derive the estimation as:

$$D = \frac{1}{t} \sum_{i=1}^{t} D_i.$$

Clearly, we still have $\mathbb{E}[D] = AB$. 
Importance Sampling

How to set the probabilities $p_1, p_2, \ldots, p_n$?

- If $A_{(\ast, w)}B_{(w, \ast)}$ contributes a lot to

  $$AB = \sum_{i=1}^{n} A_{(\ast, i)}B_{(i, \ast)}.$$ 

  we should set $p_w$ with a higher probability.

- Probability $p_w$ should not be too difficult to compute. We cannot directly use

  $$\|A_{(\ast, w)}B_{(w, \ast)}\|_F$$

  since it may need a cost no lower than $O(nmp)$.
Importance Sampling

We use the spectral normal of $A_{(\ast, w)}B_{(w, \ast)}$. The spectral normal of a matrix $A$ is the largest singular value of $A$. For rank-1 matrix, it is exactly:

$$
\|A_{(\ast, w)}B_{(w, \ast)}\|_2 = \|A_{(\ast, w)}\|_2 \|B_{(w, \ast)}\|_2.
$$

As suggested by Drineas, Kannan, and Mahoney [1], we set:

$$
p_w = \frac{\|A_{(\ast, w)}\|_2 \|B_{(w, \ast)}\|_2}{\sum_{j=1}^{n} \|A_{(\ast, j)}\|_2 \|B_{(j, \ast)}\|_2}
$$

**Remark:** This can be done by scanning matrix $A$ and matrix $B$ once. Denote $nnz(X)$ as the number of non-zero entries of a matrix $X$. Then, the weight can be computed in $O(nnz(A) + nnz(B))$. 


Total Running Time

Total running time of approximate Matrix Multiplication via random sampling:

- Deriving the probability $p_i$ for $i \in [1 \cdots n]$: $O(\text{nnz}(A) + \text{nnz}(B))$;
- Compute $D_i = A(\ast, w_i) B(w_i, \ast)$ for $i = 1, 2, \cdots, t$;
  - Each computation takes $O(mp)$ time. In total: $O(tmp)$ time.
- Final time complexity: $O(tmp + \text{nnz}(A) + \text{nnz}(B))$.

Next, we analyze how to set $t$ to derive an approximation guarantee.
Theoretical Analysis

We will show that the derived estimation has the following type of $(\varepsilon, \delta)$-approximation guarantee:

$$
\mathbb{P}[\|D - AB\|_F \geq \varepsilon \|A\|_F \|B\|_F] \leq \delta.
$$

Lemma 1

Let $D$ be the average of the matrix computed after $t$ times of random sampling. Then, we have that

$$
\mathbb{E}[D] = AB,
$$

$$
\mathbb{E}[\|D - AB\|_F^2] \leq \frac{1}{t} \|A\|_F^2 \|B\|_F^2.
$$

We will prove Lemma 1 later.
Given Lemma 1, we can derive the following result via Markov’s inequality.

\[
P[||D - AB||_F \geq \epsilon ||A||_F ||B||_F] \leq \frac{\mathbb{E}[||D - AB||^2_F]}{\epsilon^2 ||A||^2_F ||B||^2_F}
\]

\[
\leq \frac{1}{t} \frac{||A||^2_F ||B||^2_F}{\epsilon^2} \leq \frac{1}{t\epsilon^2}.
\]

By setting \( t = \lceil \frac{1}{3\epsilon^2} \rceil \), we can achieve a success probability of \( \frac{2}{3} \). By \( O(\log \frac{1}{\delta}) \) independent trials, we can further boost the success probability to \( 1 - \delta \) by median trick. But how? We do not know the ground-truth and the matrices are not scalar values that we can derive their "median".
Run $\ell$ independent trials of the random sampling based algorithm. In each trial, we further set $t = \lceil \frac{1}{3(\varepsilon/3)^2} \rceil = \lceil \frac{3}{\varepsilon^2} \rceil$ and we have that:

$$\mathbb{P}[\|D - AB\|_F \geq \frac{\varepsilon}{3} \|A\|_F \|B\|_F] \leq \frac{1}{3}.$$  

For the $\ell$ trials, denote the derived matrix as $D_1, D_2, \cdots, D_\ell$. Denote $Y_i$ as the indicator variable and it is 1 if $\|D_i - AB\|_F < \frac{\varepsilon}{3} \|A\|_F \|B\|_F$ and otherwise 0. Then, clearly, $\mathbb{E}[Y_i] \geq \frac{2}{3}$.

- For each $D_i$, derive the number $\omega_i$ as follows:

$$\omega_i = |\{j| j \neq i \land \|D_i - D_j\|_F \leq \frac{2\varepsilon}{3}\}|$$

- If we find a $D_i$ such that $\omega_i \geq \ell/2$, output $D_i$. Else report failure.
Theorem 1

If \( \ell = \lceil 48 \log(1/\delta) \rceil \) and we have \( \sum_{i=1}^{\ell} Y_i \geq \ell/2 + 1 \), then the algorithm outputs \( D_i \) such that

\[
\|D_i - AB\|_F \leq \epsilon \|A\|_F \|B\|_F
\]

holds with probability at least \( 1 - \delta \).

We show that if the majority of \( Y_i \) is 1, then the algorithm can indeed achieve the approximation guarantee. We use the triangle inequality:

\[
\begin{align*}
|D_i - D_j|_F &\leq |D_i - AB|_F + |D_j - AB|_F \\
|D_i - D_j|_F &\geq |D_i - AB|_F - |D_j - AB|_F
\end{align*}
\]

When the algorithm fails? If \( |D_i - AB|_F > \epsilon \|A\|_F \|B\|_F \) and we report \( D_i \). Here we show that we will not report such a matrix \( D_i \).

- More than half of them should have that \( |D_i - D_j|_F \geq \frac{2\epsilon}{3} \). Then, \( D_i \) will not be returned.
Median Trick for Matrix Data (Cont.)

Given a matrix that satisfies \( \|D_i - AB\|_F < \frac{\varepsilon}{3} \|A\|_F \|B\|_F \).

- Then, for a matrix \( D_j \) that belongs to this majority, we have:

\[
\|D_i - D_j\|_F \leq \|D_i - AB\|_F + \|D_j - AB\|_F \leq \frac{2\varepsilon}{3}.
\]

- Then, it should holds that \( \omega_i \geq \ell/2 \) and it will be outputed.

Total time complexity to achieve \( 1 - \delta \) success probability:

- Without median trick: \( O\left(\frac{mp}{\varepsilon^2 \delta} + \text{nnz}(A) + \text{nnz}(B)\right) \)

- We have \( O(\log(1/\delta)) \) independent trials. Then, we compare the matrices derived in these trials. In each trial, notice that we have \( O(\frac{1}{\varepsilon^2}) \) computation.

Cost: \( O\left(mp \cdot \log^2(1/\delta) + \frac{m \log(1/\delta)}{\varepsilon^2} + \text{nnz}(A) + \text{nnz}(B)\right) \).
Proof of Lemma 1

We prove a stronger version:

**Lemma 2**

\[
\mathbb{E}[\|D - AB\|_F^2] \leq \frac{1}{t} \left( \sum_{j=1}^{n} \|A_{(\ast,j)}\|_2 \|B_{(j,\ast)}\|_F \right)^2 - \frac{1}{t} \|AB\|_F^2.
\]

Then, we can derive that:

\[
\mathbb{E}[\|D - AB\|_F^2] \leq \frac{1}{t} \left( \sum_{j=1}^{n} \|A_{(\ast,j)}\|_2 \|B_{(j,\ast)}\|_F \right)^2 - \frac{1}{t} \|AB\|_F^2
\]

\[
\leq \frac{1}{t} \left( \sum_{j=1}^{n} \|A_{(\ast,j)}\|_2 \|B_{(j,\ast)}\|_2 \right)^2 \leq \frac{1}{t} \|A\|_F^2 \|B\|_F^2.
\]

Also, we only prove for the case when \( t = 1 \). The case for an arbitrary \( t \) can be easily extended using the independence of these \( t \) trials.
Proof of Lemma 1 (Cont.)

We define $M_{x,y}$ as the entry in the $x$-th row and the $y$-th column in $M$. Then, we have:

$$\mathbb{E}[\|D - AB\|_F^2] = \mathbb{E}\left[\sum_{x,y} (D_{x,y} - (AB)_{x,y})^2\right] = \sum_{x,y} \mathbb{E}[(D_{x,y} - (AB)_{x,y})^2]$$

Fix $x$ and $y$. Consider $\mathbb{E}[(D_{x,y} - (AB)_{x,y})^2]$. The random choice comes from the randomly chosen $i$ to get $i$-th column of $A$ and $i$-th row of $B$.

Given the random choice of $i$, the corresponding $(x,y)$-th entry of matrix $D$ will be $A_{x,i}B_{i,y}/p_i$. Then, note that $\mathbb{E}[D_{x,y}] = (AB)_{x,y}$. We have:

$$\mathbb{E}[(D_{x,y} - (AB)_{x,y})^2] = \mathbb{E}[D_{x,y}^2 + (AB)_{x,y}^2 - 2D_{x,y}(AB)_{x,y}]$$

$$= \mathbb{E}[D_{x,y}^2] - (AB)_{x,y}^2 = \sum_{i=1}^{n} p_i(A_{x,i}B_{i,y}/p_i)^2 - (AB)_{x,y}^2$$

$$= \sum_{i=1}^{n} (A_{x,i}B_{i,y})^2/p_i - (AB)_{x,y}^2$$
Aggregate all combinations of \( x \) and \( y \) and recap that \( p_i \) is set as 
\[
\frac{\|A_{(*)i}\|^2 \|B_{(i,*)}\|^2}{\sum_{i=1}^n \|A_{(*)i}\|^2 \|B_{(i,*)}\|^2}.
\]
We can derive:
\[
\mathbb{E}[\|D - AB\|^2_F] = \sum_{x,y} \left( \sum_{i=1}^n (A_{x,i}B_{i,y})^2 / p_i - (AB)_{x,y}^2 \right)
\]
\[
= \sum_{i=1}^n \sum_{x,y} (A_{x,i}B_{i,y})^2 / p_i - \|AB\|_F^2
\]
\[
= \sum_{i=1}^n \|A_{(*)i}\|^2 \|B_{(i,*)}\|^2 / p_i - \|AB\|_F^2
\]
\[
= \left( \sum_{i=1}^n \|A_{(*)i}\|^2 \|B_{(i,*)}\|^2 \right)^2 - \|AB\|_F^2.
\]
Proof done.
Rationale of Current Choice of $p_i$

The current setting of $p_i$ actually minimizes $\mathbb{E}[\|D - AB\|_F^2]$. Recap that:

$$\mathbb{E}[\|D - AB\|_F^2] = \sum_{i=1}^{n} \|A_{(\ast,i)}\|_2^2 \|B_{(i,\ast)}\|_2^2 / p_i - \|AB\|_F^2$$

We can define it as a constrained optimization problem:

$$\min \sum_{i=1}^{n} \|A_{(\ast,i)}\|_2^2 \|B_{(i,\ast)}\|_2^2 / p_i$$

subject to the constrain $p_1 + p_2 + \cdots + p_n = 1$. The steps are omitted and are left as self exercise.
Table of Contents

1 Approximate Matrix Multiplication
   • Random Sampling
   • Applying JL Lemma

2 Low-Rank Approximation
   • Additive Error Guarantee via Random Sampling
   • Additive Error Guarantee via JL Lemma
   • Relative Error Guarantee with Sparse JL Transform
Approximate Matrix Multiplication via JL Lemma [2]

We add an additional parameter $s$ to indicate the size of the subset that we will use the JL transformation, i.e., that will be multiplied with the matrix $\Pi$.

**Lemma 3**

A random matrix $\Pi \in \mathbb{R}^{k \times n}$ forms a JL transformation with parameters $\varepsilon, \delta, s$, or $\text{JLT}(\varepsilon, \delta, s)$ for short, if there exists a function $f$ such that for any $0 < \varepsilon, \delta < 1$, positive integer $s$ and a size $s$ subset $X \in \mathbb{R}^n$, where $k = O\left(\frac{\log s}{\varepsilon^2 f(\delta)}\right)$ for all $v \in X$ holds that

$$(1 - \varepsilon)\|v\|_2^2 \leq \|\Pi v\|_2^2 \leq (1 + \varepsilon)\|v\|_2^2.$$ 

As we have shown in previous lectures, setting $\Pi_{i,j} \sim \frac{1}{\sqrt{k}} \mathcal{N}(0, 1)$ and $k = O\left(\frac{\log(s/\delta)}{\varepsilon^2}\right)$, we can achieve $\text{JLT}(\varepsilon, \delta, s)$. 
Lemma 4

Let \( \Pi \in \mathbb{R}^{k \times n} \) be a matrix satisfying \( JLT(\varepsilon, \delta, s^2) \). Given a set \( X \) with size \( s \) consisting of unit vector only, let \( u \) and \( v \) be two vectors in \( X \). Then, the following holds:

\[
|\langle \Pi u, \Pi u \rangle - \langle u, v \rangle| \leq \varepsilon,
\]

where \( \langle u, v \rangle \) denotes the dot product of \( u \) and \( v \).

Proof.

Due to the JL property, we have:

\[
(1 - \varepsilon)\|u + v\|_2^2 \leq \|\Pi(u + v)\|_2^2 \leq (1 + \varepsilon)\|u + v\|_2^2 \tag{1}
\]

\[
(1 - \varepsilon)\|u - v\|_2^2 \leq \|\Pi(u - v)\|_2^2 \leq (1 + \varepsilon)\|u - v\|_2^2 \tag{2}
\]

Note that \( \|u\|_2 = 1 \) and \( \|v\|_2 = 1 \). Add Eqn. 1-2 finishes the proof.
Corollary 1

Let \( \Pi \in \mathbb{R}^{k \times n} \) be a matrix satisfying JLTS\((\varepsilon, \delta, s^2)\). Given an arbitrary vector \( u \) and \( v \) from set \( X \) of size \( s \), we have that:

\[
|\langle \Pi u, \Pi v \rangle - \langle u, v \rangle| \leq \varepsilon \|u\|_2 \|v\|_2,
\]

with \( 1 - \delta \) probability.

Theorem 2

If \( \Pi \in \mathbb{R}^{k \times n} \) satisfies JLTS\((\varepsilon, \delta, m + p)\), then for \( A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p} \)

\[
P[\|AB - A\Pi^T\Pi B\|_F \leq \varepsilon \|A\|_F \|B\|_F] \geq 1 - \delta.
\]
Proof of Theorem 2

Proof.

Consider entry \((AB)_{i,j}\). Then, for the entry \((A\Pi^T\Pi B)_{i,j}\), notice that it will be the \(i\)-th row of the matrix \(A\Pi^T\) and the \(j\)-th column of matrix \(\Pi B\).

Observe that the \(i\)-th row of \(A\Pi^T\) will be \(A_{(i,*)}\Pi^T\) and the \(j\)-th column of \(\Pi B\) is \(\Pi B_{(*,j)}\). Also, the \(i\)-th row and \(j\)-th column of \(A_{i,j}\) is \(A_{(i,*)}B_{(*,j)}\).

Denote \(a_i\) as the column vector by transposing the \(i\)-th row of \(A\), i.e., \((A_{(i,*)})^T\) and \(b_j\) as the column vector of the \(j\)-th column of \(B\), i.e., \(B_{(*,j)}\). Then, \((A\Pi^T\Pi B)_{i,j} = \langle \Pi a_i, \Pi b_j \rangle\) and \((AB)_{i,j} = \langle a_i, b_j \rangle\). Then,

\[
\|AB - A\Pi^T\Pi B\|_F^2 = \sum_{i,j} |\Pi a_i, \Pi b_j\rangle - \langle a_i, b_j \rangle|^2 \\
\leq \sum_{i,j} \epsilon^2 \|a_i\|_2^2 \cdot \|b_j\|_2^2 = \epsilon^2 \|A\|_F^2 \|B\|_F^2.
\]

This finishes the proof.
The key is to convert vectors of dimension $n$ into vectors of dimension $k = O\left(\frac{1}{\varepsilon^2} \log \left(\frac{m+p}{\delta}\right)\right)$.

- Time to compute $A\Pi^T$ and $\Pi B$: $O(kmn + knp)$.
  - If we use faster JL transform, we can compute with $O(k\text{nnz}(A) + k\text{nnz}(B))$

- Time to compute $A\Pi^T\Pi B$: $O(kmp)$

- Total time complexity: $O\left(\frac{1}{\varepsilon^2} \log \left(\frac{m+p}{\delta}\right) (mp + \text{nnz}(A) + \text{nnz}(B))\right)$. 
1. Approximate Matrix Multiplication
   - Random Sampling
   - Applying JL Lemma

2. Low-Rank Approximation
   - Additive Error Guarantee via Random Sampling
   - Additive Error Guarantee via JL Lemma
   - Relative Error Guarantee with Sparse JL Transform
Low-Rank Approximation

Given an input $k$ and an input matrix $A \in \mathbb{R}^{m \times n}$, the low-rank approximation aims to solve the following optimization problem:

$$\min_{B \in \mathbb{R}^{m \times n} : \text{rank}(B) = k} ||A - B||_F.$$ 

As we discussed in the last lecture, let $A = U\Sigma V$. $A_k = U_k\Sigma_k V_k$ is the optimal solution. However, this takes $O(\min(mn^2, nm^2))$ time.

To reduce the time complexity, we derive an approximate solution that provides additive error guarantee. In particular, we aim to derive a rank-$k$ matrix $A'$ such that:

$$||A - A'||_F \leq ||A - A_k|| + \varepsilon||A||_F$$
Low-Rank Approximation with Random Sampling

We make a connection with the sampling-based matrix multiplication:

- We consider matrix multiplication $\mathbf{A}\mathbf{A}^T$.
- The sampling process aims to sample a matrix $\mathbf{C}$ from columns from $\mathbf{A}$ so that $\mathbf{A}\mathbf{A}^T \approx \mathbf{C}\mathbf{C}^T$.

For column $\mathbf{A}_{(:,w)}$, we sample with probability $\frac{||\mathbf{A}_{(:,w)}||_2^2}{||\mathbf{A}||_F^2}$.

The random sampling-based solution LinearSVD works as follows:

- Initialize an all-zero matrix $\mathbf{C}$ of size $m \times c$.
- Let $i = 1 \cdots c$. In the $i$-th iteration, sample a column $\mathbf{A}_{(:,w)}$ with probability $p_w$. Copy $\frac{\mathbf{A}_{(:,w)}}{\sqrt{c \cdot p_w}}$ to $\mathbf{C}_{(:,i)}$.
- Compute $\mathbf{C}^T\mathbf{C}$ and derive the eigen-decomposition of $\mathbf{C}^T\mathbf{C}$. Let $\mathbf{y}_i$ be the eigenvector corresponding to the $i$-th largest eigenvalue of $\mathbf{C}^T\mathbf{C}$.
- $\mathbf{y}_i$ is the right singular vector. Derive $\mathbf{h}_i = \mathbf{C}\mathbf{y}_i/\sigma_i(\mathbf{C})$.
- Return $\mathbf{H}_k$ where the $i$-th column of $\mathbf{H}_k$ is $\mathbf{h}_i$. 
Theoretical Analysis

Theorem 3

Given an input matrix $A$ and the input $k$, let $H_k$ be constructed from the LinearSVD. Then, we have:

$$\|A - H_kH_k^T A\|_F^2 \leq \|A - A_k\|_F^2 + 2\sqrt{k}||AA^T - CC^T||_F.$$

Recap: To approximate $AB$, by draw $c = \lceil \frac{3}{\epsilon^2} \rceil$ samples (where each sample get a random $w \in [1 \cdots n]$ and compute $A(\ast, w)B(w, \ast)$), we have that

$$\|D - AB\|_F^2 \leq \epsilon^2||A||_F^2||B||_F^2$$

holds with at least $2/3$ probability. Here, let $B = A^T$. Also, we have that $CC^T = D$ according to how we have sampled the matrix $C$. We can thus bound $2\sqrt{k}||AA^T - CC^T||_F$ to $\epsilon||A||_F^2$. 
Proof of Theorem 3

We will use the following result from the Matrix Perturbation Theory \(^1\).

**Theorem 4 (Hoffman-Wielandt inequality)**

*Given a matrix \(A\), let \(E\) be a perturbation of matrix \(A\).*

\[
\sum_{i=1}^{n} (\sigma_i(A + E) - \sigma_i(A))^2 \leq \|E\|_F^2,
\]

*where \(\sigma_i(X)\) means the \(i\)-th largest singular value of matrix \(X\).*

Also, let \(\text{Tr}(X)\) be the trace of a square matrix, which is the sum of the entries at the diagonal line. We will use the fact that:

\[
\|X\|_F^2 = \text{Tr}(X^T X) = \text{Tr}(XX^T)
\]

\[
\text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B)
\]

---

\(^1\)Matrix Perturbation Theory includes many theoretical results concerning a perturbation \(E\) to matrix \(A\). Here we only list one result.
According to the connection between trace and the Frobenius norm, we can write $\|A - H_k H_k^T A\|_F^2$ as:

$$\|A - H_k H_k^T A\|_F^2 = \text{Tr} \left( (A - H_k H_k^T A)^T (A - H_k H_k^T A) \right)$$

$$= \text{Tr} \left( A^T A - 2A^T H_k H_k^T A + A^T H_k H_k^T H_k H_k^T A \right)$$

$$= \text{Tr} \left( A^T A - A^T H_k H_k^T A \right) = \|A\|_F^2 - \|A^T H_k\|_F^2$$

Next, we will make a connection between (i) $\|A^T H_k\|_F^2$ and $\sum_{i=1}^k \sigma_i^2(C)$; (ii) $\sum_{i=1}^k \sigma_i^2(C)$ and $\sum_{i=1}^k \sigma_i^2(A)$. 

Proof of Theorem 3 (Cont.)
Proof of Theorem 3 (Cont.)

We first make a connection between $\|A^T H_k\|_F^2$ and $\sum_{i=1}^{k} \sigma_i^2(C)$.

\[
\left| \|A^T H_k\|_F^2 - \sum_{i=1}^{k} \sigma_i^2(C) \right| = \left| \sum_{i=1}^{k} (\|A^T h_i\|_2^2 - \sigma_i^2(C)) \right| \leq \\
\sum_{i=1}^{k} \left| \|A^T h_i\|_2^2 - \sigma_i^2(C) \right| \leq \sqrt{k} \left( \sum_{i=1}^{k} (\|A^T h_i\|_2^2 - \sigma_i^2(C))^2 \right)^{\frac{1}{2}} \\
= \sqrt{k} \left( \sum_{i=1}^{k} (\|A^T h_i\|_2^2 - \|C^T h_i\|_2^2)^2 \right)^{\frac{1}{2}} = \sqrt{k} \left( \sum_{i=1}^{k} (h_i^T (AA^T - CC^T) h_i)^2 \right)^{\frac{1}{2}} \\
\leq \sqrt{k} \|AA^T - CC^T\|_F.
\]

For the last inequality, we will need to use the optimization problem that we tackled in SVD analysis.
Next, we make a connection between $\sum_{i=1}^{k} \sigma_i^2(C)$ and $\sum_{i=1}^{k} \sigma_i^2(A)$.

\[
\left| \sum_{i=1}^{k} \sigma_i^2(C) - \sum_{i=1}^{k} \sigma_i^2(A) \right| \leq \sqrt{k} \sqrt{\sum_{i=1}^{k} (\sigma_i^2(C) - \sigma_i^2(A))^2}
\]

\[
= \sqrt{k} \sqrt{\sum_{i=1}^{k} \left(\sigma_i(CC^T) - \sigma_i(AA^T)\right)^2}
\]

\[
\leq \sqrt{k} \sqrt{\sum_{i=1}^{m} \left(\sigma_i(CC^T) - \sigma_i(AA^T)\right)^2} \leq \sqrt{k} \|CC^T - AA^T\|_F.
\]

The last inequality is due to Theorem 4.
Proof of Theorem 3 (Cont.)

We have:

\[ \| \mathbf{A} - \mathbf{H}_k \mathbf{H}_k^T \mathbf{A} \|_F^2 = \| \mathbf{A} \|_F^2 - \| \mathbf{A}^T \mathbf{H}_k \|_F^2 \]  
(3)

\[ \left| \| \mathbf{A}^T \mathbf{H}_k^T \|_F^2 - \sum_{i=1}^{k} \sigma_i^2(\mathbf{C}) \right| \leq \sqrt{k} \| \mathbf{A} \mathbf{A}^T - \mathbf{C} \mathbf{C}^T \|_F \]  
(4)

\[ \left| \sum_{i=1}^{k} \sigma_i^2(\mathbf{C}) - \sum_{i=1}^{k} \sigma_i^2(\mathbf{A}) \right| \leq \sqrt{k} \| \mathbf{C} \mathbf{C}^T - \mathbf{A} \mathbf{A}^T \|_F \]  
(5)

Using the fact that \( \| \mathbf{A} - \mathbf{A}_k \|_F^2 = \| \mathbf{A} \|_F^2 - \sum_{i=1}^{k} \sigma_i^2(\mathbf{A}) \) and combining the above inequalities, we prove Theorem 3.
By setting $c = O\left(\frac{k}{\varepsilon^2}\right)$, we can have a constant success probability larger than $1/2$. The time complexity:

- $O(mc)$ to construct matrix $C$.
- Compute $C^T C$: $O(mc^2)$. Calculate the SVD of $C^T C$: $O(c^3)$.
- Compute $H_k$: $O(kcm)$.
- If $k$ and $c$ are $O(1)$, then the algorithm linearly depends on $m$.
- To boost the success probability to $1 - \delta$: the time complexity needs to be multiplied by a factor of $O(\log(1/\delta))$. 
# Table of Contents

## 1 Approximate Matrix Multiplication
- Random Sampling
- Applying JL Lemma

## 2 Low-Rank Approximation
- Additive Error Guarantee via Random Sampling
- Additive Error Guarantee via JL Lemma
- Relative Error Guarantee with Sparse JL Transform
Low-Rank Approximation via Random Projection \[3\]

Low-Rank approximation via JL transformation-based random projection:

- Build a matrix $\Pi \in \mathbb{R}^{c \times m}$ satisfying $JLT(\varepsilon/2, \delta, 2k)$.
- Compute $B = \Pi A$.
- Then, derive the SVD for matrix $B = U_{[c \times c]} \Sigma_{[c \times n]} V_{[n \times n]}^T$. Let $U = [a_1, a_2, \ldots, a_c]$, $\Sigma_{i,i} = \lambda_i$, and $V = [b_1, b_2, \ldots, b_n]$. We can also write $B$ as the sum of outer product:

$$B = \sum_{i=1}^{c} \lambda_i a_i b_i^T.$$

- Return $B_k$ or $\hat{A} = A \sum_{i=1}^{k} b_i b_i^T$.

Total Running Time: Step 1 takes $O(mc)$ time. Step 2 takes $O(cnnz(A))$. Step 3 takes $O(nc^2)$. The total cost: $O(mc + cnnz(A) + nc^2)$. 
Theorem 5

Let \( \Pi \) be a matrix that satisfies \( JLT(\varepsilon/2, \delta, 2k) \). Then, we have that:

\[
\| A - \hat{A} \|_F \leq \| A - A_k \|_F^2 + \varepsilon \| A_k \|_F^2
\]

holds with at least \( 1 - \delta \) probability.

Note that given a orthonormal matrix \( X \in \mathbb{R}^{n \times n} \), we have \( \| A \|_F = \| AX \|_F \).

Recall that \( B = U\Sigma V^T \) and \( V \) is a \( n \times n \) orthonormal matrix. Hence,

\[
\| A - \hat{A} \|_F^2 = \| (A - \hat{A})V \|_F^2 = \sum_{i=1}^{n} \| Ab_i - \hat{Ab}_i \|_2^2
\]

\[
= \sum_{i=1}^{n} \| Ab_i - (A \sum_{j=1}^{k} b_j b_j^T) b_i \|_2^2 = \sum_{i=k+1}^{n} \| Ab_i \|_2^2
\]

\[
= \| A \|_F^2 - \sum_{i=1}^{k} \| Ab_i \|_2^2
\]
Theoretical Analysis

Note that $\lambda_i^2 = \|Bb_i\|_2^2$. It is not difficult to prove that:

$$
\sum_{i=1}^{k} \lambda_i^2 = \sum_{i=1}^{k} \|Bb_i\|_2^2 \leq (1 + \varepsilon/2) \sum_{i=1}^{k} \|Ab_i\|_2^2,
$$

$$
\sum_{i=1}^{k} \lambda_i^2 = \sum_{i=1}^{k} \|Bb_i\|_2^2 \geq (1 - \varepsilon/2) \|A_k\|_F^2.
$$

The proof is left as self-exercise. Combining the above results, we have:

$$
\sum_{i=1}^{k} \|Ab_i\|_2^2 \geq \frac{1 - \varepsilon/2}{1 + \varepsilon/2} \|A_k\|_F^2 \geq (1 - \varepsilon) \|A_k\|_F^2.
$$

As $\|A - \hat{A}\|_F^2 = \|A\|_F^2 - \sum_{i=1}^{k} \|Ab_i\|_2^2$. We further have:

$$
\|A - \hat{A}\|_F^2 \leq \|A\|_F^2 - \|A_k\| + \varepsilon \|A_k\|_F^2 = \|A - A_k\|_F^2 + \varepsilon \|A_k\|_F^2.
$$
Table of Contents

1 Approximate Matrix Multiplication
   • Random Sampling
   • Applying JL Lemma

2 Low-Rank Approximation
   • Additive Error Guarantee via Random Sampling
   • Additive Error Guarantee via JL Lemma
   • Relative Error Guarantee with Sparse JL Transform
Kenneth Clarkson and David Woodruff present new results on the time complexity of low-rank approximation via a sparse JL Transform. The key idea: build a sparse matrix using the idea of Count-Sketch, firstly presented in [4]. In particular:

- For matrix $\Pi \in \mathbb{R}^{c \times m}$, each column has a single non-zero entry sampled randomly. The value is set to 1 and -1 with equal probability.
- The time complexity of matrix multiplication $\Pi A$: $O(\text{nnz}(A))$.

**Theorem 6 (Theorem 23 in [5])**

For $A \in \mathbb{R}^{n \times n}$, there is an algorithm that with failure probability $1/10$ finds matrices $L, W \in \mathbb{R}^{n \times k}$ with orthonormal columns, and diagonal $D \in \mathbb{R}^{k \times k}$, so that $\|A - LDW^T\| \leq (1 + \varepsilon)\|A - A_k\|_F$. The algorithm runs in time

$$O(\text{nnz}(A)) + \tilde{O}(nk^2 \varepsilon^{-4} + k^3 \varepsilon^{-5}).$$

---

\(^2\)It also includes result on other numerical linear algebra.
Petros Drineas, Ravi Kannan, and Michael W. Mahoney.
Fast monte carlo algorithms for matrices I: approximating matrix multiplication.

Tamás Sarlós.
Improved approximation algorithms for large matrices via random projections.

Latent semantic indexing: A probabilistic analysis.

Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós.
A sparse johnson–lindenstrauss transform.

Kenneth L. Clarkson and David P. Woodruff.
Low rank approximation and regression in input sparsity time.