Approximate Matrix Multiplication

Lecture 18
October 27, 2022

Matrix data

Lot of data can be viewed as defining a matrix. We have already seen vectors modeling data/signals. More generally we can use tensors too.

n data items and each data item a_i is a vector over some features (say m features)

A is the matrix defined by the n data items.

Assuming $a_1, \ldots, a_n$ are columns then A is a $m \times n$ matrix

Combinatorial objects such as graphs can also be modeled via graphs

Numerical Linear Algebra

Basic problems in linear algebra:

- Matrix vector product: compute $\boldsymbol{Ax}$
- Matrix multiplication: compute $\boldsymbol{AB}$
- Linear equations: solve $\boldsymbol{Ax} = \boldsymbol{b}$
- Matrix inversion: compute $\boldsymbol{A}^{-1}$
- Least squares: solve $\min_{\boldsymbol{x}} \|\boldsymbol{Ax} - \boldsymbol{b}\|$
- Singular value decomposition, eigen values, principal component analysis, low-rank approximations
- $\ldots$

Fundamental in all areas of applied mathematics and engineering. Many applications to statistics and data analysis.

Numerical Linear Algebra

NLA has a vast literature

In practice iterative methods are used that converge to an optimum solution. They can take advantage of sparsity in the input data better than exact methods

Some TCS contributions in the recent past:

- randomized NLA for faster algorithms with provable approximation guarantees - sampling and JL based techniques and others
- revisit preconditioning methods for Laplacians and beyond
- Many powerful applications in theory and practice

Norms and matrix norms

Definition

A norm $\|\|$ in a real vector space V is a real valued function that has three properties: (i) $\|x\| \geq 0$ for all $x \in V$ and $\|x\| = 0$ implies $x = 0$, (ii) $\|ax\| = |a|\|x\|$ for all scalars a (iii) $\|x + y\| \leq \|x\| + \|y\|$

Familiar vector norms: $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$

If A is a injective linear transformation $\|Ax\|$ is also a norm in the original space.

Norms and metrics: $d(x, y) = \|x - y\|$ is a metric

Matrix norms

Consider vector space of all matrices $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{n}}$

What are useful norms over matrices?

- Treat matrix like a vector of dimension $\boldsymbol{m} \times \boldsymbol{n}$ and apply vector norm. For instance $\|\boldsymbol{A}\|_{\boldsymbol{F}}$ (Frobenius norm) is $(\sum_{\boldsymbol{i},\boldsymbol{j}} |\boldsymbol{A}_{\boldsymbol{i},\boldsymbol{j}}|^2)^{1/2}$.
- Treat matrix as linear operator and see what it does to norms of vectors it operates on. Spectral norm is $\sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{Ax}\|_2$.
- Schatten $\boldsymbol{p}$-norms based on singular values of $\boldsymbol{A}$
- Trace norm, nuclear norm, $\ldots$
- Norms are related in some cases (different perspective on the same norm)

Frobenus and Spectral norms

Some other properties:

$$\|AB\|_F \leq \|A\|_F \|B\|_F$$

$$\|AB\|_2 \leq \|A\|_2 \|B\|_2$$

Above property is calle sub-multiplicative property and is typical for matrix norms See more at https://en.wikipedia.org/wiki/Matrix_norm

Matrix Multiplication

Problem: Given matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times h}$ compute the matrix AB

- Standard algorithm based on definition: $O(mnh)$ time
- Faster algorithms via non-trivial Strassen-like divide and conquer.

Matrix Multiplication

Problem: Given matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times h}$ compute the matrix AB

- Standard algorithm based on definition: $O(mnh)$ time
- Faster algorithms via non-trivial Strassen-like divide and conquer.

Approximation: Compute $D \in \mathbb{R}^{m \times h}$ such that $\|D - AB\|$ is small in some appropriate matrix norm.

Two methods

- random sampling
- random projections (fast JL)

Matrix Multiplication

Problem: Given matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times h}$ compute the matrix AB

Notation: $M^{(j)}$ for j'th column of M and $M_{(i)}$ for i'th row of M both interpreted as vectors

From textbook definition: $D_{i,j} = \langle A_{(i)}, B^{(j)} \rangle = \sum_{k=1}^{n} A_{i,k} B_{k,j}$

Consider A^T consisting of m column vectors from $\mathbb{R}^n$ and B as h column vectors from $\mathbb{R}^n$

We want to compute all mh inner products of these vectors.

Part I

Random Sampling for Approx Matrix Mult

Approximate Matrix Multiplication

Want to approximate AB in the Frobenius norm "additively". Want D such that $\|D - AB\|_F \leq \epsilon\|AB\|_F$ but instead will settle for $\|D - AB\|_F \leq \epsilon\|A\|_F\|B\|_F$

Alternate definition of matrix multiplication based on outer product:

$$AB = \sum_{j=1}^{n} A^{(j)} B_{(j)}$$

$A^{(j)} B_{(j)}$ is a $m \times h$ matrix of rank 1

Importance Sampling

$$AB = \sum_{j=1}^{n} A^{(j)} B_{(j)}$$

- Pick a probability distribution over $[n]$, $p_1 + p_2 + \ldots + p_n = 1$
- For $\ell = 1$ to t do pick an index $j_\ell \in [n]$ according to distribution p (independent with replacement)
- Output $C = \frac{1}{t} \sum_{\ell=1}^{t} \frac{1}{p_{i_\ell}} A^{(j_\ell)} B_{(j_\ell)}$

Importance Sampling

$$AB = \sum_{j=1}^{n} A^{(j)} B_{(j)}$$

- Pick a probability distribution over $[n]$, $p_1 + p_2 + \ldots + p_n = 1$
- For $\ell = 1$ to t do pick an index $j_\ell \in [n]$ according to distribution p (independent with replacement)
- Output $C = \frac{1}{t} \sum_{\ell=1}^{t} \frac{1}{p_{i_\ell}} A^{(j_\ell)} B_{(j_\ell)}$

$C = \frac{1}{t} \sum_\ell C_\ell$ where $\mathrm{E}[C_\ell] = AB$.
By linearity of expectation: $\mathrm{E}[C] = AB$

Importance Sampling

Question: How should we choose $p_1, p_2, \ldots, p_n$?

Importance Sampling

Question: How should we choose $p_1, p_2, \ldots, p_n$? p_j should correspond to contribution of $A^{(j)}B_{(j)}$ to $\|AB\|_F$

Importance Sampling

Question: How should we choose $p_1, p_2, \ldots, p_n$? p_j should correspond to contribution of $A^{(j)}B_{(j)}$ to $\|AB\|_F$

Use spectral norm of $A^{(j)}B_{(j)}$ which is $\|A^{(j)}B_{(j)}\|_2$

Importance Sampling

Question: How should we choose $p_1, p_2, \ldots, p_n$? p_j should correspond to contribution of $A^{(j)}B_{(j)}$ to $\|AB\|_F$

Use spectral norm of $A^{(j)}B_{(j)}$ which is $\|A^{(j)}B_{(j)}\|_2$

Claim: $\|A^{(j)}B_{(j)}\|_2 = \|A^{(j)}\|_2\|B_{(j)}\|_2$.

Choose $p_j = \frac{\|A^{(j)}\|_2\|B_{(j)}\|_2}{\sum_\ell \|A^{(\ell)}\|_2\|B_{(\ell)}\|_2}$

Due to [Drineas-Kannan-Mahoney]

Running time

- For all j compute $\|A^{(j)}\|_2$ and $\|B_{(j)}\|_2$. Takes one pass over A and B
- Allows one to compute $p_1, p_2, \ldots, p_n$
- $C = \frac{1}{t}\sum_{\ell=1}^{t} \frac{1}{p_{i_\ell}} A^{(j_\ell)} B_{(j_\ell)}$
- At most $O(tmh + N_A + N_B)$ time where N_A and N_B is number of non-zeroes in A and B.
- Full computation takes $O(nmh)$ time.

Analysis of approximation

We would ideally like to know $\Pr[\|C - AB\|_F \geq \epsilon \|AB\|_F]$. Hard to understand $\|AB\|_F$. Instead analyse $\Pr[\|C - AB\|_F \geq \epsilon \|A\|_F \|B\|_F]$ since $\|AB\|_F \leq \|A\|_F \|B\|_F$.

Analysis of approximation

We would ideally like to know $\Pr[\|C - AB\|_F \geq \epsilon\|AB\|_F]$. Hard to understand $\|AB\|_F$. Instead analyse $\Pr[\|C - AB\|_F \geq \epsilon\|A\|_F\|B\|_F]$ since $\|AB\|_F \leq \|A\|_F\|B\|_F$.

Using Markov:

$$\Pr[\|C - AB\|_F \geq \epsilon\|A\|_F\|B\|_F] \leq \frac{\mathrm{E}\left[\|C - AB\|_F^2\right]}{\epsilon^2\|A\|_F^2\|B\|_F^2}$$

Lemma

$$\mathrm{E}\left[\|C - AB\|_F^2\right] \leq \frac{1}{t}\left(\sum_{j=1}^{n}\|A^{(j)}\|_2\|B_{(j)}\|_2\right)^2 - \frac{1}{t}\|AB\|_F^2$$

Proof later.

Analysis continued

Using Lemma on previous slide:

$$
\begin{aligned}
\mathrm{E}\left[\|C - AB\|_F^2\right] &\leq \frac{1}{t}\left(\sum_{j=1}^{n} \|A^{(j)}\|_2 \|B_{(j)}\|\right)^2 - \frac{1}{t}\|AB\|_F^2 \\
&\leq \frac{1}{t}\left(\sum_{j=1}^{n} \|A^{(j)}\|_2 \|B_{(j)}\|\right)^2 \\
&\leq \frac{1}{t}(\sum_{j=1}^{n} \|A^{(j)}\|_2^2)(\sum_{j=1}^{n} \|B_{(j)}\|_2^2) \quad \text{(Cauchy-Schwartz)} \\
&= \frac{1}{t}\|A\|_F^2 \|B\|_F^2 .
\end{aligned}
$$

Analysis contd

$$\begin{aligned}\Pr[\|C - AB\|_F \geq \epsilon \|A\|_F \|B\|_F] &\leq \frac{\mathrm{E}\left[\|C - AB\|_F^2\right]}{\epsilon^2 \|A\|_F^2 \|B\|_F^2} \\ &\leq \frac{1}{t\epsilon^2}\end{aligned}$$

Thus, if $t = \frac{1}{\epsilon^2 \delta}$ then

$$\Pr[\|C - AB\|_F \geq \epsilon \|A\|_F \|B\|_F] \leq \delta.$$

Median trick

Recall that we used median trick to improve dependence on δ from $1/\delta$ to $\log(1/\delta)$.

If $t = \frac{3}{\epsilon^2}$ then

$$\Pr[\|C - AB\|_F \geq \epsilon\|A\|_F\|B\|_F] \leq 1/3.$$

Repeat independently to obtain $C_1, C_2, \ldots, C_r$ where $r = \Theta(\log(1/\delta))$

By Chernoff bounds majority of estimators are good. How do we pick the "median" matrix?

Median trick

If $t = \frac{3}{\epsilon^2}$ then $\Pr[\|C - AB\|_F \geq \epsilon\|A\|_F\|B\|_F] \leq 1/3$.

[Clarkson-Woodruff]

- Repeat independently to obtain $C_1, C_2, \ldots, C_r$ where $r = \Theta(\log(1/\delta))$
- For each $1 \leq i \leq r$ compute

$$\rho_i = |\{j \mid j \neq i, \|C_i - C_j\| \leq 2\epsilon\|A\|_F\|B\|_F\}|$$

- If there is s such that $\rho_s \geq r/2$ output C_s. Else report failture.

Median trick

Lemma

Suppose $\|C_i - AB\|_F \leq \epsilon\|A\|_F\|B\|_F$ *for a majority of the indices (which happens with probability at least* $1-\delta$*) then algorithm correctly outputs an* s *such that* $\|C_s - AB\|_F \leq 3\epsilon\|A\|_F\|B\|_F$*.*

Median trick

Lemma

Suppose $\|C_i - AB\|_F \leq \epsilon\|A\|_F\|B\|_F$ *for a majority of the indices (which happens with probability at least* $1-\delta$*) then algorithm correctly outputs an* s *such that* $\|C_s - AB\|_F \leq 3\epsilon\|A\|_F\|B\|_F$*.*

Use triangle inequality:

$$\|C_i - C_j\|_F \leq \|C_i - AB\|_F + \|C_j - AB\|_F$$

and

$$\|C_i - C_j\|_F \geq \|C_i - AB\|_F - \|C_j - AB\|_F.$$

Proof

Suppose C_s satisfies $\|C_s - AB\|_F \leq \epsilon \|A\|_F \|B\|_F$. Then by triangle inequality and the fact that majority are within $\epsilon \|A\|_F \|B\|_F$ distance from AB, $\rho_s \geq r/2$. Thus there is always a candidate that is output when majority of indices are good.

Suppose C_s is bad, that is $\|C_s - AB\|_F > 3\|A\|_F \|B\|_F$. Claim is that $\rho_s < r/2$ since majority are within distance $\epsilon \|A\|_F \|B\|_F$ of AB. Thus, C_s will not be a candidate for output.

Note that when majority are within $\epsilon \|A\|_F \|B\|_F$ distance from AB, the above properties show that a good candidate exists, and no bad candidate is output. Thus the output C_s satisfies the property that $\|C_s - AB\|_F \leq 3\epsilon \|A\|_F \|B\|_F$.

Running time again

- For all j compute $\|A^{(j)}\|_2$ and $\|B_{(j)}\|_2$. $O(N_A + N_B)$ time.
- Allows one to compute $p_1, p_2, \ldots, p_n$
- $C = \frac{1}{t}\sum_{\ell=1}^{t} \frac{1}{p_{i_\ell}} A^{(j_\ell)} B_{(j_\ell)}$
- Total time: $O(tmh + N_A + N_B)$ time

Two options for ensuring output C satisfies $\|C - AB\|_F \leq \epsilon \|A\|_F \|B\|_F$ with probability at least $(1-\delta)$:

- Choose $t = \frac{1}{\epsilon^2 \delta}$ via Chebyshev
- Choose $t = \frac{1}{\epsilon^2}\ln(1/\delta)$ but median trick is weaker. We need to do pairwise matrix computations so effectivel $t = \frac{1}{\epsilon^4}\log^2(1/\delta)$.

Proof of Lemma

Lemma

$\mathrm{E}\left[\|C - AB\|_F^2\right] \leq \frac{1}{t}\left(\sum_{j=1}^{n}\|A^{(j)}\|_2\|B_{(j)}\|_2\right)^2 - \frac{1}{t}\|AB\|_F^2$

Recall C is sum of t independent estimators so the lemma is basically about $t = 1$.

$\mathrm{E}\left[\|C - AB\|_F^2\right] = \sum_{x,y}\mathrm{E}[(C_{x,y} - (AB)_{x,y})^2]$

Fix x, y. Let $Z = C_{x,y}$. We have $\mathrm{E}[Z] = (AB)_{x,y}$. Hence

$$Var[Z] = \mathrm{E}\left[Z^2\right] - (AB)_{x,y}^2 = \mathrm{E}\left[(C_{x,y} - (AB)_{x,y})^2\right]$$

Proof of Lemma

$\mathrm{E}[Z^2] = \sum_{j=1}^{n} p_j (A_{x,j} B_{j,y})^2 / p_j^2 = \sum_{j=1}^{n} (A_{x,j} B_{j,y})^2 / p_j$

Thus $\mathrm{E}\left[\|C - AB\|_F^2\right] = \sum_{x,y} \sum_{j=1}^{n} (A_{x,j} B_{j,y})^2 / p_j - \|AB\|_F^2$.

Simplifying the first term:
$\sum_{x,y} \sum_{j=1}^{n} (A_{x,j)})^2 (B_{j,y})^2 / p_j = \sum_{j=1}^{n} \frac{1}{p_j} \|A^{(j)}\|^2 \|B_{(j)}\|^2$

Proof of Lemma

$\mathrm{E}[Z^2] = \sum_{j=1}^{n} p_j (A_{x,j} B_{j,y})^2 / p_j^2 = \sum_{j=1}^{n} (A_{x,j} B_{j,y})^2 / p_j$

Thus $\mathrm{E}\left[\|C - AB\|_F^2\right] = \sum_{x,y} \sum_{j=1}^{n} (A_{x,j} B_{j,y})^2 / p_j - \|AB\|_F^2$.

Simplifying the first term:
$\sum_{x,y} \sum_{j=1}^{n} (A_{x,j)})^2 (B_{j,y})^2 / p_j = \sum_{j=1}^{n} \frac{1}{p_j} \|A^{(j)}\|^2 \|B_{(j)}\|^2$

Recall: $p_j = \frac{\|A^{(j)}\| \|B_{(j)}\|}{\sum_\ell \|A^{(\ell)}\| \|B_{(\ell)}\|}$

Thus $\mathrm{E}\left[\|C - AB\|_F^2\right] = \left(\sum_{j=1}^{n} \|A^{(j)}\| \|B_{(j)}\|\right)^2 - \|AB\|_F^2$.

Proof of Lemma

$\mathrm{E}[Z^2] = \sum_{j=1}^{n} p_j (A_{x,j} B_{j,y})^2 / p_j^2 = \sum_{j=1}^{n} (A_{x,j} B_{j,y})^2 / p_j$

Thus $\mathrm{E}\left[\|C - AB\|_F^2\right] = \sum_{x,y} \sum_{j=1}^{n} (A_{x,j} B_{j,y})^2 / p_j - \|AB\|_F^2$.

Simplifying the first term:
$\sum_{x,y} \sum_{j=1}^{n} (A_{x,j)})^2 (B_{j,y})^2 / p_j = \sum_{j=1}^{n} \frac{1}{p_j} \|A^{(j)}\|^2 \|B_{(j)}\|^2$

Recall: $p_j = \frac{\|A^{(j)}\| \|B_{(j)}\|}{\sum_\ell \|A^{(\ell)}\| \|B_{(\ell)}\|}$

Thus $\mathrm{E}\left[\|C - AB\|_F^2\right] = \left(\sum_{j=1}^{n} \|A^{(j)}\| \|B_{(j)}\|\right)^2 - \|AB\|_F^2$.

One can show that the choice of p_j values is optimum for reducing variance in the simple importance sampling scheme.

Sampling matrix view

$$AB = \sum_{j=1}^{n} A^{(j)} B_{(j)}$$

- Pick a probability distribution over $[n]$, $p_1 + p_2 + \ldots + p_n = 1$
- For $\ell = 1$ to t do pick an index $j_\ell \in [n]$ according to distribution p (independent with replacement)
- Output $C = \frac{1}{t} \sum_{\ell=1}^{t} \frac{1}{p_{i_\ell}} A^{(j_\ell)} B_{(j_\ell)}$

$C = (AS^T)(SB)$ where $S \in \mathbb{R}^{n \times t}$ is a sampling matrix:

$S_{i,j} = \frac{1}{\sqrt{tp_j}}$ if column j is picked in i'th sample else $S_{i,j} = 0$

Part II

Random Projection for Approx Matrix Mult

JL Approach for Approx Matrix Multiplication

[Sarlos]

We have $(AB)_{i,j} = \langle A_{(i)}, B^{(j)} \rangle$. Require mh dot products each of which takes n time. Project all rows and columns into lower dimensions and then compute the dot product.

Output $C = (AS^T)(SB)$ where S is a (fast) JL matrix. Works!

Advantage?

JL Approach for Approx Matrix Multiplication

[Sarlos]

We have $(AB)_{i,j} = \langle A_{(i)}, B^{(j)} \rangle$. Require mh dot products each of which takes n time. Project all rows and columns into lower dimensions and then compute the dot product.

Output $C = (AS^T)(SB)$ where S is a (fast) JL matrix. Works!

Advantage? Oblivious to A, B. Can update them etc.

Recalling JL

Lemma (Distributional JL Lemma)

Fix vector $x \in \mathbb{R}^d$ *and let* $\Pi \in \mathbb{R}^{k \times d}$ *matrix where each entry* Π_{ij} *is chosen independently according to standard normal distribution* $\mathcal{N}(0,1)$ *distribution. If* $k = \Omega(\frac{1}{\epsilon^2}\log(1/\delta))$, *then with probability* $(1-\delta)$ *we have* $\|\frac{1}{\sqrt{k}}\Pi x\|_2 = (1 \pm \epsilon)\|x\|_2$.

Can choose entries from $\{-1, 1\}$ as well.

Definition

Let $\mathcal{D}$ be a distribution over $m \times n$ matrices. $\mathcal{D}$ is said to have (ϵ, δ) JL moment property if for any unit vector $x \in \mathbb{R}^n$,

$$E_{\Pi \sim \mathcal{D}} |\|\Pi x\|_2^2 - 1| \leq \epsilon\delta.$$

JL Property

Angles are approximately preserved by JL projection:

Lemma

If Π *comes from* (ϵ, δ) *JL moment distribution then for all unit vectors* $x, y \in \mathbb{R}^n$, $\mathrm{E}[|\langle \Pi x, \Pi y\rangle - \langle x, y\rangle|^2] \leq c\epsilon^2\delta$.

Using JL

Theorem

Suppose Π *is chosen from a distribution* $\mathcal{D}$ *that satisfies* (ϵ, δ) *JL moment property then*

$$\Pr_{\Pi \sim \mathcal{D}}[\|AB - (A\Pi^T)(B\Pi)\|_F > 3\epsilon\|A\|_F\|B\|_F] \le \delta.$$

Using JL

Theorem

Suppose Π is chosen from a distribution $\mathcal{D}$ that satisfies (ϵ, δ) JL moment property then

$$\Pr_{\Pi \sim \mathcal{D}}[\|AB - (A\Pi^T)(B\Pi)\|_F > 3\epsilon \|A\|_F \|B\|_F] \leq \delta.$$

Let $C = (A\Pi^T)(B\Pi)$.

$$\Pr[\|AB - C\|_F^2 > 3\epsilon \|A\|_F \|B\|_F^2] \leq \frac{\mathrm{E}\left[\|AB - C\|_F^2\right]}{(3\epsilon \|A\|_F \|B\|_F)^2}.$$

Analysis

$C_{i,j} = \langle \Pi A_{(i)}, \Pi B^{(j)} \rangle$ while $(AB)_{i,j} = \langle A_{(i)}, B^{(j)} \rangle$

Notation: a_i for $A_{(i)}$ and b_j for $B^{(j)}$

Analysis

$C_{i,j} = \langle \Pi A_{(i)}, \Pi B^{(j)} \rangle$ while $(AB)_{i,j} = \langle A_{(i)}, B^{(j)} \rangle$

Notation: a_i for $A_{(i)}$ and b_j for $B^{(j)}$

$\|AB - C\|_F^2 = \sum_{i,j} |\langle \Pi a_i, \Pi b_j \rangle - \langle a_i, b_j \rangle|^2$

Analysis

$C_{i,j} = \langle \Pi A_{(i)}, \Pi B^{(j)} \rangle$ while $(AB)_{i,j} = \langle A_{(i)}, B^{(j)} \rangle$

Notation: a_i for $A_{(i)}$ and b_j for $B^{(j)}$

$\|AB - C\|_F^2 = \sum_{i,j} |\langle \Pi a_i, \Pi b_j \rangle - \langle a_i, b_j \rangle|^2$

Term by term:
$|\langle \Pi a_i, \Pi b_j \rangle - \langle a_i, b_j \rangle|^2 = \alpha |\langle \Pi \frac{a_i}{\|a_i\|_2}, \Pi \frac{b_j}{\|b_j\|_2} \rangle - \langle \frac{a_i}{\|a_i\|_2}, \frac{b_j}{\|b_j\|_2} \rangle|^2$
where $\alpha = \|a_i\|_2^2 \|b_j\|_2^2$

Analysis

$C_{i,j} = \langle \Pi A_{(i)}, \Pi B^{(j)} \rangle$ while $(AB)_{i,j} = \langle A_{(i)}, B^{(j)} \rangle$

Notation: a_i for $A_{(i)}$ and b_j for $B^{(j)}$

$\|AB - C\|_F^2 = \sum_{i,j} |\langle \Pi a_i, \Pi b_j \rangle - \langle a_i, b_j \rangle|^2$

Term by term:
$|\langle \Pi a_i, \Pi b_j \rangle - \langle a_i, b_j \rangle|^2 = \alpha |\langle \Pi \frac{a_i}{\|a_i\|_2}, \Pi \frac{b_j}{\|b_j\|_2} \rangle - \langle \frac{a_i}{\|a_i\|_2}, \frac{b_j}{\|b_j\|_2} \rangle|^2$
where $\alpha = \|a_i\|_2^2 \|b_j\|_2^2$

Applying JL property and linearity of expectation

$$\mathrm{E}\left[\|AB - C\|_F^2\right] \leq (c\epsilon)^2 \delta \sum_{i,j} \|a_i\|_2^2 \|b_j\|_2^2 \leq (c\epsilon^2)\delta \|A\|_F^2 \|B\|_F^2$$

Analysis

$$\Pr[\|AB - C\|_F^2 > 3\epsilon\|A\|_F\|B\|_F^2] \leq \frac{\mathrm{E}\left[\|AB - C\|_F^2\right]}{(3\epsilon\|A\|_F\|B\|_F)^2}.$$

and

$$\mathrm{E}\left[\|AB - C\|_F^2\right] \leq (c\epsilon)^2\delta\sum_{i,j}\|a_i\|_2^2\|b_j\|_2^2 \leq (c\epsilon^2)\delta\|A\|_F^2\|B\|_F^2$$

hence

$$\Pr[\|AB - C\|_F^2 > 3\epsilon\|A\|_F\|B\|_F^2] \leq \delta.$$

Running time

Roughly speaking: Π converts vectors of dimension n into vectors of dimension $d = O(\frac{1}{\epsilon^2} \log(1/\delta))$.

Need to compute ΠA^T and ΠB and then compute dot products.

mh inner products of vectors of dimension d which is $O(mhd^2)$ time in the worst case

Using Fast JL with very sparse Π one can improve running time