EE270
Large scale matrix computation, optimization and learning

Instructor : Mert Pilanci

Stanford University

Thursday, Jan 23 2020

Randomized Linear Algebra
Lecture 6: Johnson Lindenstrauss Lemma and Applications

Dimension Reduction

- map a high dimensional vector to low dimensions such that certain properites are preserved
- examples so far:
- Approximate Matrix Multiplication $AS^TSB \approx AB$ where S is random
- Freivalds Algorithm $ABr - Mr$ where r is random
- Trace estimation $r^TMr \approx \mathbf{tr}(M)$ where r is random

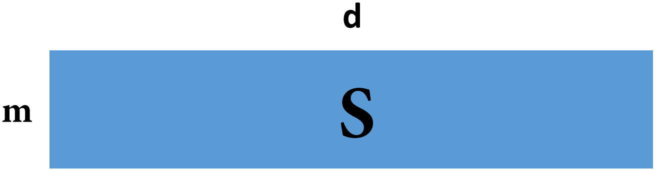
Dimension Reduction

- map a high dimensional vector to low dimensions such that certain properites are preserved
- examples so far:
- Approximate Matrix Multiplication $AS^TSB \approx AB$ where S is random
- Freivalds Algorithm $ABr - Mr$ where r is random
- Trace estimation $r^T Mr \approx \mathbf{tr}(M)$ where r is random

- Generic dimension reduction problem
- Given vectors $x_1, ..., x_n \in \mathbb{R}^d$, compress the data points into low dimensional representation $y_1, ..., y_n \in \mathbb{R}^m$ where $m < d$
- another instance is Principal Component Analysis

Randomized Dimension Reduction

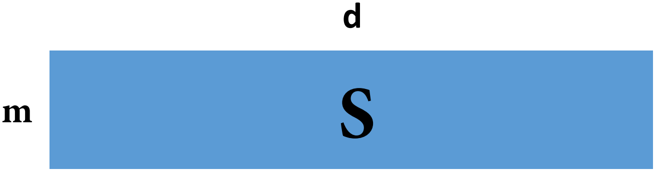
- Given vectors $x_1, \dots, x_n \in \mathbb{R}^d$, compress the data points into low dimensional representation $y_1, \dots, y_n \in \mathbb{R}^m$ where $m < d$



- Linear transformation $y_i = Sx_i$ for $i = 1, \dots, n$
- S is chosen randomly

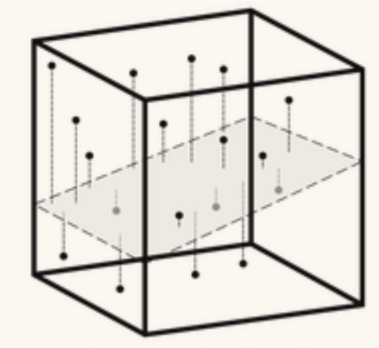
Randomized Dimension Reduction

- Given vectors $x_1, \dots, x_n \in \mathbb{R}^d$, compress the data points into low dimensional representation $y_1, \dots, y_n \in \mathbb{R}^m$ where $m < d$



- Linear transformation $y_i = Sx_i$ for $i = 1, \dots, n$
- S is chosen randomly
- Approximate Matrix Multiplication: $AS^TSB \approx AB$ where S is random matrix

Geometry of Random Projections



Johnson Lindenstrauss Lemma

- Let $\epsilon \in (0, \frac{1}{2})$. Given any set of points $\{x_1, ..., x_n\}$ in $\mathbb{R}^d$, there exists a map $S : \mathbb{R}^n \to \mathbb{R}^m$ with $m = \frac{9\log(n)}{\epsilon^2 - \epsilon^3}$ such that

$$1 - \epsilon \leq \frac{\|Sx_i - Sx_j\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \epsilon$$

Johnson Lindenstrauss Lemma

- Let $\epsilon \in (0, \frac{1}{2})$. Given any set of points $\{x_1, ..., x_n\}$ in $\mathbb{R}^d$, there exists a map $S : \mathbb{R}^n \to \mathbb{R}^m$ with $m = \frac{9 \log(n)}{\epsilon^2 - \epsilon^3}$ such that

$$1 - \epsilon \leq \frac{\|Sx_i - Sx_j\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \epsilon$$

- Note that the target dimension m is **independent of the original dimension** d, and depends **only on the number of points** n and the accuracy parameter.

Johnson Lindenstrauss Lemma

- Let $\epsilon \in (0, \frac{1}{2})$. Given any set of points $\{x_1, ..., x_n\}$ in $\mathbb{R}^d$, there exists a map $S : \mathbb{R}^n \to \mathbb{R}^m$ with $m = \frac{9\log(n)}{\epsilon^2 - \epsilon^3}$ such that

$$1 - \epsilon \leq \frac{\|Sx_i - Sx_j\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \epsilon$$

- Note that the target dimension m is **independent of the original dimension** d, and depends **only on the number of points** n and the accuracy parameter.
- more surprises: picking an $m \times d$ random matrix $S = \frac{1}{\sqrt{m}}G$ with $G_{ij} \sim N(0, 1)$ standard normal works with high probability!

Johnson Lindenstrauss (JL) Lemma

- Define $u_{ij} \triangleq \frac{x_i - x_j}{\|x_i - x_j\|_2}$.
- note that $\|u_{ij}\|_2 = 1$

- JL Lemma:
  $\mathbb{P}\left[\|Su_{ij}\|_2^2 \in (1 \pm \epsilon) \text{ for all } i, j \in \{1, ..., n\}\right] \geq 1 - \delta$

  where $\delta \in (0, 1)$ for large enough m

Proof of JL Lemma

- We need to show $\|Su_{ij}\|_2^2$ is concentrated around 1
- **Lemma** Let $S_{ij} \sim \frac{1}{\sqrt{m}} N(0,1)$ and u be any fixed vector. Then

$$\mathbb{E}\|Su\|_2^2 = \|u\|$$

- implies that the distance between two points is preserved in expectation
- **Proof:**

Concentration of Measure for Uniform Distribution on the Sphere

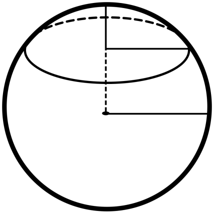
- Suppose $m = 1$, i.e., we project to dimension one
- S is a row vector $S = g^T \in \mathbb{R}^d \sim N(0, I)$
- $\mathbb{P}\left[|g^T u| \geq \epsilon\right] = \mathbb{P}\left[|g^T e_1| \geq \epsilon\right] = \mathbb{P}\left[|g_1| \geq \epsilon\right]$
  where e_1 is the first ordinary basis vector

Concentration of Measure for Uniform Distribution on the Sphere

- Suppose $m = 1$, i.e., we project to dimension one
- S is a row vector $S = g^T \in \mathbb{R}^d \sim N(0, I)$
- $\mathbb{P}\left[|g^T u| \geq \epsilon\right] = \mathbb{P}\left[|g^T e_1| \geq \epsilon\right] = \mathbb{P}\left[|g_1| \geq \epsilon\right]$ where e_1 is the first ordinary basis vector
- **Lemma:** $\mathbb{P}\left[|s_1| \geq \frac{t\|g\|_2}{\sqrt{d}}\right] \leq 2e^{-\frac{t^2}{2}}$.
- Note that $\frac{g}{\|g\|_2}$ is distributed uniformly on the unit sphere

Concentration of Measure for Uniform Distribution on the Sphere

- **Lemma:** $\mathbb{P}\left[|g_1| \geq \frac{t\|g\|_2}{\sqrt{d}}\right] \leq 2e^{-\frac{t^2}{2}}$.
- Note that $\frac{g}{\|g\|_2}$ is distributed uniformly on the unit sphere



- Pythagorean theorem: $\frac{t^2}{d} + R_{\text{cap}}^2 = 1$ implies $R_{\text{cap}} = \sqrt{1 - \frac{t^2}{d}}$
- $\mathbb{P}\left[|\frac{g_1}{\|g\|_2}| \geq \frac{t}{\sqrt{d}}\right] \leq \frac{\text{area of the spherical cap}}{\text{area of the sphere}} \leq \frac{\left(\sqrt{1-\frac{t^2}{d}}\right)^{d-1}}{1^{d-1}}$
- using the fact $(1 - \frac{x}{n})^n \leq e^{-x}$ we get
  $\mathbb{P}\left[|\frac{g_1}{\|g\|_2}| \geq \frac{t}{\sqrt{d}}\right] \leq 2e^{-\frac{t^2}{2}}$.

Proof of JL Lemma

- Back to the general case $S \in \mathbb{R}^{m \times d}$
- Consider the probability that $\|Su\|_2^2$ deviates from 1, i.e., projected vectors are stretched more than their expectation

$$\mathbb{P}\left[\|Su\|_2^2 \geq (1+\epsilon)\|u\|_2^2\right] \leq e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}}$$

Proof of JL Lemma

- Back to the general case $S \in \mathbb{R}^{m \times d}$
- Consider the probability that $\|Su\|_2^2$ deviates from 1, i.e., projected vectors are stretched more than their expectation

$$\mathbb{P}\left[\|Su\|_2^2 \geq (1+\epsilon)\|u\|_2^2\right] \leq e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}}$$

- $\mathbb{P}\left[\|Su_{ij}\|_2^2 \geq (1+\epsilon)\|u_{ij}\|_2^2\right] \leq \sum_{i,j} e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}} = n^2 e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}}$

Proof of JL Lemma

- Back to the general case $S \in \mathbb{R}^{m\times d}$
- Consider the probability that $\|Su\|_2^2$ deviates from 1, i.e., projected vectors are stretched more than their expectation

$$\mathbb{P}\left[\|Su\|_2^2 \geq (1+\epsilon)\|u\|_2^2\right] \leq e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}}$$

- $\mathbb{P}\left[\|Su_{ij}\|_2^2 \geq (1+\epsilon)\|u_{ij}\|_2^2\right] \leq \sum_{i,j} e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}} = n^2 e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}}$

  Set error probability $= \frac{1}{2} = n^2 e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}}$
- $m = \frac{9\log n}{\epsilon^2-\epsilon^3}$

Proof of JL Lemma

- Back to the general case $S \in \mathbb{R}^{m \times d}$
- Consider the probability that $\|Su\|_2^2$ deviates from 1, i.e., projected vectors are stretched more than their expectation

$$\mathbb{P}\left[\|Su\|_2^2 \geq (1+\epsilon)\|u\|_2^2\right] \leq e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}}$$

- $\mathbb{P}\left[\|Su_{ij}\|_2^2 \geq (1+\epsilon)\|u_{ij}\|_2^2\right] \leq \sum_{i,j} e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}} = n^2 e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}}$

  Set error probability $= \frac{1}{2} = n^2 e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}}$
- $m = \frac{9 \log n}{\epsilon^2 - \epsilon^3}$

  for smaller error probability $0.01 = n^2 e^{-(\epsilon^2-\epsilon^3)\frac{m}{4}}$
- $m = \frac{\text{constant} \times \log n}{\epsilon^2 - \epsilon^3}$

True 'projections': random subspaces also work

- Pick $S_{(i)}$ uniformly random on the unit sphere
- Pick $S_{(i+1)}$ uniformly random on the unit sphere and $\perp S_{(i)}, ...S_{(1)}$
- S is a projection matrix, which projects onto a uniformly random subspace

$$\mathbb{P}\left\{\left|\|Su\|_2 - \sqrt{\frac{m}{d}}\right| > t\right\} \leq 2e^{\frac{-t^2 d}{2}}$$

- Applying union bound for all points $i, j = 1, ..., d$ gives a similar result
- Random i.i.d. S matrices are easier to generate and approximately orthogonal: $\mathbb{E}S^T S = I$

Computationally cheaper random matrices

- Gaussian $S_{ij} = \frac{1}{\sqrt{m}} N(0,1)$
- Rademacher

$$S_{ij} = \begin{cases} +\frac{1}{m} & \text{with probability } \frac{1}{\sqrt{m}} \\ -\frac{1}{\sqrt{m}} & \text{with probability } \frac{1}{2} \end{cases} \tag{1}$$

- Bernoulli-Rademacher

$$S_{ij} = \begin{cases} +\frac{\sqrt{3}}{\sqrt{m}} & \text{with probability } \frac{1}{2} \\ 0 & \text{with probability } \frac{2}{3} \\ -\frac{\sqrt{3}}{\sqrt{m}} & \text{with probability } \frac{1}{2} \end{cases} \tag{2}$$

- other sparse matrices (e.g. one non-zero per column)
- Fourier transform based matrices

Optimality of the JL Embedding

- Let $\epsilon \in (0, \frac{1}{2})$. Given any set of points $\{x_1, ..., x_n\}$ in $\mathbb{R}^d$, there exists a map $S : \mathbb{R}^n \to \mathbb{R}^m$ with $m = \frac{9\log(n)}{\epsilon^2 - \epsilon^3}$ such that

$$1 - \epsilon \leq \frac{\|Sx_i - Sx_j\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \epsilon \qquad (\star)$$

- Can we embed to a **smaller dimension**?
- maybe using a **nonlinear** embedding?

Optimality of the JL Embedding

- Let $\epsilon \in (0, \frac{1}{2})$. Given any set of points $\{x_1, ..., x_n\}$ in $\mathbb{R}^d$, there exists a map $S : \mathbb{R}^n \to \mathbb{R}^m$ with $m = \frac{9\log(n)}{\epsilon^2 - \epsilon^3}$ such that

$$1 - \epsilon \leq \frac{\|Sx_i - Sx_j\|_2^2}{\|x_i - x_j\|_2^2} \leq 1 + \epsilon \qquad (\star)$$

- Can we embed to a **smaller dimension**?
- maybe using a **nonlinear** embedding?
- **No**

  Johnson-Lindenstrauss Embedding is optimal
- There exists a set of n points $\{x_1, ..., x_n\}$ such that any linear/nonlinear embedding satisfying $(\star)$ must have $m \geq O(\frac{\log n}{\epsilon^2})$.

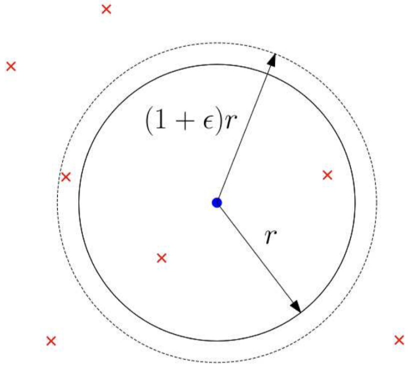
Optimality of the Johnson-Lindenstrauss Lemma, Larsen and Nelson, 2016

Applications of JL Embeddings

- General idea: run algorithms on $Sx_1, \ldots, Sx_n \in \mathbb{R}^m$ instead of $x_1, \ldots, x_n$
- Examples:
  - approximate nearest neighbor search
  - estimating norms and frequency moments
  - regression
  - classification
  - randomized matrix operations (matrix multiplication, decomposition etc)
  - optimization
  - ...

Approximate Nearest Neighbors

- Given a point set $P = \{x_1, ..., x_n\} \in \mathbb{R}^d$
- and a query point $q \in \mathbb{R}^d$
- Find an ϵ-approximate nearest neighbor to q from P



Estimating p-norms

- Streaming data

  $x_{t+1} = x_t + \delta_t$
- Estimate $\|x\|_2$
- second moment

Estimating p-norms

- Streaming data

  $x_{t+1} = x_t + \delta_t$
- Estimate $\|x\|_2$
- second moment
- **linear sketch**

  Generate S randomly such that $\mathbb{E} S^T S = I$

  Let $y_t = Sx_t$

  $y_t = Sx_t + S\delta_t$
- $\|Sy\|_2^2 \approx \|Sx\|_2^2$

Estimating p-norms

- Streaming data

  $x_{t+1} = x_t + \delta_t$
- Estimate $\|x\|_2$
- second moment
- **linear sketch**

  Generate S randomly such that $\mathbb{E}S^T S = I$

  Let $y_t = Sx_t$

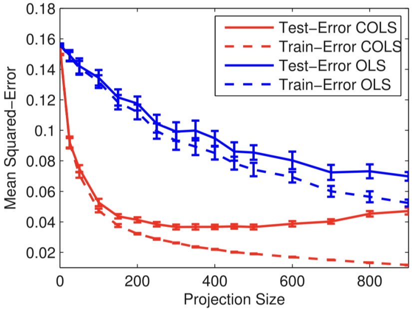
  $y_t = Sx_t + S\delta_t$
- $\|Sy\|_2^2 \approx \|Sx\|_2^2$
- Can also be extended to $\|x\|_p$

Music similarity prediction

- Predict the similarity score $\in [0, 1]$ between 30 second tracks
- Frequency based features from each 200ms segment results in 10^6 features
- OLS: randomly pick m features
- COLS: apply random projection to dimension m

Fard et al. Compressed Least-Squares Regression on Sparse Spaces, 2012

Music similarity prediction



- Predict the similarity score $\in [0, 1]$ between 30 second tracks
- Frequency based features from each 200ms segment results in 10^6 features
- OLS: randomly pick m features
- COLS: apply random projection to dimension m

Fard et al. Compressed Least-Squares Regression on Sparse Spaces, 2012

Random Projection for Approximate Matrix Multiplication

- matrices A and B
  output AS^TSB where S is a random projection matrix

Random Projection for Approximate Matrix Multiplication

- matrices A and B
  output AS^TSB where S is a random projection matrix
- need to characterize $\|Sx\|_2^2 - \|x\|_2^2$ for vectors x
- **Definition:** (ϵ, δ, p) JL moment property

$$\mathbb{E}\left|\|Sx\|_2^2 - 1\right|^p \leq \epsilon^p \delta$$

for any unit norm x where $p \geq 2$

Random Projection for Approximate Matrix Multiplication

- matrices A and B
  output AS^TSB where S is a random projection matrix
- need to characterize $\|Sx\|_2^2 - \|x\|_2^2$ for vectors x
- **Definition:** (ϵ, δ, p) JL moment property

$$\mathbb{E}\left|\|Sx\|_2^2 - 1\right|^p \leq \epsilon^p \delta$$

  for any unit norm x where $p \geq 2$
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$random i.i.d. sub-Gaussian, e.g., ± 1, or $N(0,1)$ with $m = \frac{c_1}{\epsilon^2} \log \frac{1}{\delta}$ satisfies $(\epsilon, \delta, \log \frac{1}{\delta})$ JL moment property

Random Projection for Approximate Matrix Multiplication

- matrices A and B
  output AS^TSB where S is a random projection matrix
- need to characterize $\|Sx\|_2^2 - \|x\|_2^2$ for vectors x
- **Definition:** (ϵ, δ, p) JL moment property

$$\mathbb{E}\left|\|Sx\|_2^2 - 1\right|^p \leq \epsilon^p \delta$$

  for any unit norm x where $p \geq 2$
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$random i.i.d. sub-Gaussian, e.g., ± 1, or $N(0,1)$ with $m = \frac{c_1}{\epsilon^2} \log \frac{1}{\delta}$ satisfies $(\epsilon, \delta, \log \frac{1}{\delta})$ JL moment property
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$CountSketch matrix (one nonzero per column, which is ± 1 at a uniformly random location) with $m = \frac{c_2}{\epsilon^2 \delta}$ satisfies $(\epsilon, \delta, 2)$ JL moment property

Random Projection for Approximate Matrix Multiplication

- matrices A and B
  output AS^TSB where S is a random projection matrix
- need to characterize $\|Sx\|_2^2 - \|x\|_2^2$ for vectors x
- **Definition:** (ϵ, δ, p) JL moment property

$$\mathbb{E}\left|\|Sx\|_2^2 - 1\right|^p \leq \epsilon^p \delta$$

  for any unit norm x where $p \geq 2$
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$random i.i.d. sub-Gaussian, e.g., ± 1, or $N(0,1)$ with $m = \frac{c_1}{\epsilon^2} \log \frac{1}{\delta}$ satisfies $(\epsilon, \delta, \log \frac{1}{\delta})$ JL moment property
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$CountSketch matrix (one nonzero per column, which is ± 1 at a uniformly random location) with $m = \frac{c_2}{\epsilon^2 \delta}$ satisfies $(\epsilon, \delta, 2)$ JL moment property
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$Fast JL Transform with $m = \frac{c_3}{\epsilon} \log \frac{1}{\delta}$ satisfies $(\epsilon, \delta, \log \frac{n}{\delta})$ JL moment property

Approximating inner products

- **Lemma**

$$\mathbb{E}\left|\|Sx\|_2^2 - 1\right|^p \leq \epsilon^p \delta$$

for any unit norm x implies that

$$\mathbb{E}\left|x^T S^T S y - x^T y\right|^p \leq 3\epsilon^p \delta$$

since

$$x^T y = \frac{1}{2}\left(\|x\|_2^2 + \|y\|_2^2 - \|x - y\|_2^2\right)$$

$$x^T S^T S y = \frac{1}{2}\left(\|Sx\|_2^2 + \|Sy\|_2^2 - \|S(x - y)\|_2^2\right)$$

Random Projection for Approximate Matrix Multiplication

- Let $C = AS^TSB$

$$\mathbb{P}\left[\|AB - C\|_F > 3\epsilon\|A\|_F\|B\|_F\right] = \left[\|AB - C\|_F^p > (3\epsilon)^p\|A\|_F^p\|B\|_F^p\right]$$
$$\leq \frac{\mathbb{E}\|AB - C\|_F^p}{(3\epsilon\|A\|_F\|B\|_F)^p}$$

- Let $a_i = A_{(i)}$ and $b_i = B_{(i)}$

$$\|AB - C\|_F^2 = \sum_{ij}\left|(Sa_i)^T(Sb_j) - a_i^Tb_j\right|^2$$

Random Projection for Approximate Matrix Multiplication

- Let $C = AS^T SB$

$$\mathbb{P}\left[\|AB - C\|_F > 3\epsilon \|A\|_F \|B\|_F\right] = \left[\|AB - C\|_F^p > (3\epsilon)^p \|A\|_F^p \|B\|_F^p\right]$$
$$\leq \frac{\mathbb{E}\|AB - C\|_F^p}{(3\epsilon \|A\|_F \|B\|_F)^p}$$

- Let $a_i = A_{(i)}$ and $b_i = B_{(i)}$

$$\|AB - C\|_F^2 = \sum_{ij} \left|(Sa_i)^T (Sb_j) - a_i^T b_j\right|^2$$

- we can normalize $\frac{a_i}{\|a_i\|_2}, \frac{b_i}{\|b_i\|_2}$ and apply JL moment property to get

$$\mathbb{P}\left[\|AB - C\|_F > 3\epsilon \|A\|_F \|B\|_F\right] \leq \delta$$

Final error bound for random projection

- Let the approximate product of AB be $C = AS^T SB$

$$\mathbb{P}\left[\|AB - C\|_F > 3\epsilon \|A\|_F \|B\|_F\right] \leq \delta$$

- Follows from JL Moment property
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$random i.i.d. sub-Gaussian, e.g., ± 1, or $N(0,1)$ with $m = \frac{c_1}{\epsilon^2} \log \frac{1}{\delta}$
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$CountSketch matrix (one nonzero per column, which is ± 1 at a uniformly random location) with $m = \frac{c_2}{\epsilon^2 \delta}$
- $S \in \mathbb{R}^{m \times n} \sim \frac{1}{\sqrt{m}} \times$Fast JL Transform with $m = \frac{c_3}{\epsilon} \log \frac{1}{\delta}$

Final error bound for random projection

- Let the approximate product of AB be $C = AS^T SB$

$$\mathbb{P}\left[\|AB - C\|_F > 3\epsilon\|A\|_F\|B\|_F\right] \leq \delta$$

- Follows from JL Moment property
- $S \in \mathbb{R}^{m\times n} \sim \frac{1}{\sqrt{m}} \times$random i.i.d. sub-Gaussian, e.g., ± 1, or $N(0,1)$ with $m = \frac{c_1}{\epsilon^2}\log\frac{1}{\delta}$
- $S \in \mathbb{R}^{m\times n} \sim \frac{1}{\sqrt{m}} \times$CountSketch matrix (one nonzero per column, which is ± 1 at a uniformly random location) with $m = \frac{c_2}{\epsilon^2\delta}$
- $S \in \mathbb{R}^{m\times n} \sim \frac{1}{\sqrt{m}} \times$Fast JL Transform with $m = \frac{c_3}{\epsilon}\log\frac{1}{\delta}$
- Sparse JL and Fast JL are more efficient
- advantages: doesn't require any knowledge about matrices A and B (**oblivious**)
- optimal sampling probabilities depend on the column/row norms of A and B

Questions?