Result for Vectors

- Show: \( \Pr[|C^{\top}SD - CD|_{F}^2 \leq [6/(\delta(\text{# rows of } S))] \ast |C|_{F}^2 |D|_{F}^2] \geq 1 - \delta \)

- (JL Property) A distribution on matrices \( S \in \mathbb{R}^{k \times n} \) has the \((\epsilon, \delta, \ell)\)-JL moment property if for all \( x \in \mathbb{R}^{n} \) with \( |x|_{2} = 1 \),
  \[
  \mathbb{E}_{S} \left| |Sx|_{2}^{\ell} - 1 \right|^{\ell} \leq \epsilon^{\ell} \cdot \delta
  \]

- (From vectors to matrices) For \( \epsilon, \delta \in \left(0, \frac{1}{2}\right) \), let \( D \) be a distribution on matrices \( S \) with \( k \) rows and \( n \) columns that satisfies the \((\epsilon, \delta, \ell)\)-JL moment property for some \( \ell \geq 2 \). Then for \( A, B \) matrices with \( n \) rows,
  \[
  \Pr_{S} \left[ |A^{\top}S^{\top}SB - A^{\top}B|_{F} \geq 3 \epsilon |A|_{F} |B|_{F} \right] \leq \delta
  \]

- Just need to show that the CountSketch matrix \( S \) satisfies JL property and bound the number \( k \) of rows
CountSketch Satisfies the JL Property

- (JL Property) A distribution on matrices $S \in \mathbb{R}^{k \times n}$ has the $(\epsilon, \delta, \ell)$-JL moment property if for all $x \in \mathbb{R}^n$ with $|x|_2 = 1$,
  \[
  E_S \left| |Sx|_2^2 - 1 \right|^{\ell} \leq \epsilon^\ell \cdot \delta
  \]

- We show this property holds with $\ell = 2$. First, let us consider $E_S[|Sx|_2^2]$.

- For CountSketch matrix $S$, let
  - $h: [n] \rightarrow [k]$ be a 2-wise independent hash function
  - $\sigma: [n] \rightarrow \{-1, 1\}$ be a 4-wise independent hash function

- Let $\delta(E) = 1$ if event $E$ holds, and $\delta(E) = 0$ otherwise

- \[
  E[|Sx|_2^2] = \sum_{j \in [k]} E\left[ \left( \sum_{i \in [n]} \delta(h(i) = j)\sigma_i x_i \right)^2 \right] \\
  = \sum_{j \in [k]} \sum_{i1, i2 \in [n]} E[\delta(h(i1) = j)\delta(h(i2) = j)\sigma_{i1}\sigma_{i2}x_{i1}x_{i2}] \\
  = \sum_{j \in [k]} \sum_{i \in [n]} E[\delta(h(i) = j)^2]x_i^2 \\
  = \left( \frac{1}{k} \right) \sum_{j \in [k]} \sum_{i \in [n]} x_i^2 = |x|_2^2
  \]
CountSketch Satisfies the JL Property

- \( E[|Sx|^4] = E[\sum_{i \in [n]} \delta(h(i) = j) \sigma_i x_i^2] = \sum_{j_1, j_2, i_1, i_2, i_3, i_4} E[\sigma_{i_1} \sigma_{i_2} \sigma_{i_3} \sigma_{i_4} \delta(h(i_1) = j_1) \delta(h(i_2) = j_1) \delta(h(i_3) = j_2) \delta(h(i_4 = j_2))] x_{i_1} x_{i_2} x_{i_3} x_{i_4} \)

- We must be able to partition \( \{i_1, i_2, i_3, i_4\} \) into equal pairs

- Suppose \( i_1 = i_2 = i_3 = i_4 \). Then necessarily \( j_1 = j_2 \). Obtain \( \sum_{j} \frac{1}{k} \sum_i x_i^4 = |x|^4 \)

- Suppose \( i_1 = i_2 \) and \( i_3 = i_4 \) but \( i_1 \neq i_3 \). Then get \( \sum_{j_1, j_2, i_1, i_3} \frac{1}{k^2} x_{i_1}^2 x_{i_3}^2 = |x|^4 - |x|^4 \)

- Suppose \( i_1 = i_3 \) and \( i_2 = i_4 \) but \( i_1 \neq i_2 \). Then necessarily \( j_1 = j_2 \). Obtain \( \sum_{j} \frac{1}{k^2} \sum_{i_1, i_2} x_{i_1}^2 x_{i_2}^2 \leq \frac{1}{k} |x|^4 \). Obtain same bound if \( i_1 = i_4 \) and \( i_2 = i_3 \).

- Hence, \( E[|Sx|^4] \in [|x|^4, |x|^4(1 + \frac{2}{k})] = [1, 1 + \frac{2}{k}] \)

- So, \( \mathbb{E}_S |Sx|^2 - 1|^2 \leq \left(1 + \frac{2}{k}\right) - 2 + 1 = \frac{2}{k} \). Setting \( k = \frac{2}{\varepsilon^2 \delta} \) finishes the proof.
Where are we?

- (JL Property) A distribution on matrices $S \in \mathbb{R}^{k \times n}$ has the $(\epsilon, \delta, \ell)$-JL moment property if for all $x \in \mathbb{R}^n$ with $|x|_2 = 1$,
  $$E_S||Sx|_2^2 - 1|_\ell \leq \epsilon^\ell \cdot \delta$$

- (From vectors to matrices) For $\epsilon, \delta \in \left(0, \frac{1}{2}\right)$, let $D$ be a distribution on matrices $S$ with $k$ rows and $n$ columns that satisfies the $(\epsilon, \delta, \ell)$-JL moment property for some $\ell \geq 2$. Then for $A, B$ matrices with $n$ rows,
  $$\Pr_S\left[|A^T S^T S B - A^T B|_F^2 \geq 3 \epsilon^2 |A|_F^2 |B|_F^2 \right] \leq \delta$$

- We showed CountSketch has the JL property with $\ell = 2$, and $k = \frac{2}{\epsilon^2 \delta}$

- Matrix product result we wanted was:
  $$\Pr[|CS^T S D - CD|_F^2 \leq (6/(\delta k)) \cdot |C|_F^2 |D|_F^2] \geq 1 - \delta$$

- We are now done with the proof CountSketch is a subspace embedding
Course Outline

- Subspace embeddings and least squares regression
  - Gaussian matrices
  - Subsampled Randomized Hadamard Transform
  - CountSketch
- Affine embeddings
  - Application to low rank approximation
- High precision regression
- Leverage score sampling
- Distributed low rank approximation
- L1 Regression
- M-Estimator regression
Affine Embeddings

- Want to solve \( \min_{X} |AX - B|_{F}^{2} \), A is tall and thin with d columns, but B has a large number of columns

- Can’t directly apply subspace embeddings

- Let’s try to show \( |SAX - SB|_{F} = (1 \pm \epsilon)|AX - B|_{F} \) for all X and see what properties we need of S

- Can assume A has orthonormal columns

- Let \( B^{*} = AX^{*} - B \), where \( X^{*} \) is the optimum

\[
|S(AX - B)|_{F}^{2} - |SB^{*}|_{F}^{2} = |SA(X - X^{*}) + S(AX^{*} - B)|_{F}^{2} - |SB^{*}|_{F}^{2}
= |SA(X - X^{*})|_{F}^{2} + 2\text{tr}[(X - X^{*})^{T}A^{T}S^{T}SB^{*}] \quad \text{(use } |C + D|_{F}^{2} = |C|_{F}^{2} + |D|_{F}^{2} + 2\text{tr}(C^{T}D))
\leq |SA(X - X^{*})|_{F}^{2} + 2|X - X^{*}|_{F}|A^{T}S^{T}SB^{*}|_{F} \quad \text{(use } \text{tr}(CD) \leq |C|_{F}|D|_{F})
\leq |SA(X - X^{*})|_{F}^{2} + 2\epsilon|X - X^{*}|_{F}|B^{*}|_{F} \quad \text{(if we have approx. matrix product)}
\leq |A(X - X^{*})|_{F}^{2} + \epsilon(|A(X - X^{*})|_{F}^{2} + 2|X - X^{*}|_{F}|B^{*}|_{F}) \quad \text{(subspace embedding for A)}
Affine Embeddings

- We have

$$|S(AX - B)|_F^2 - |SB^*|_F^2 \in |A(X - X^*)|_F^2 \pm \epsilon( |A(X - X^*)|_F^2 + 2|X - X^*|_F|B^*|_F)$$

- Normal equations imply that

$$|AX - B|_F^2 = |A(X - X^*)|_F^2 + |B^*|_F^2$$

- $$|S(AX - B)|_F^2 - |SB^*|_F^2 - \left( |AX - B|_F^2 - |B^*|_F^2 \right)$$

$$\in \pm \epsilon( |A(X - X^*)|_F^2 + 2|X - X^*|_F|B^*|_F)$$

$$\in \pm \epsilon \left( |A(X - X^*)|_F + |B^*|_F \right)^2$$

$$\in \pm 2\epsilon \left( |A(X - X^*)|_F^2 + |B^*|_F^2 \right)$$

$$= \pm 2\epsilon |AX - B|_F^2$$

- $$|SB^*|_F^2 = (1 \pm \epsilon)|B^*|_F^2$$ (this holds with constant probability)
Affine Embeddings

- Know: $|S(AX - B)|_F^2 - |SB^*|_F^2 - (|AX - B|_F^2 - |B^*|_F^2) \in \pm 2\epsilon|AX - B|_F^2$

- Know: $|SB^*|_F^2 = (1 \pm \epsilon)|B^*|_F^2$

- $|S(AX - B)|_F^2 = (1 \pm 2\epsilon)|AX - B|_F^2 \pm \epsilon|B^*|_F^2$
  \[= (1 \pm 3\epsilon)|AX - B|_F^2\]

- Completes proof of affine embedding!
Affine Embeddings: Missing Proofs

- Claim: \( |A + B|_F^2 = |A|_F^2 + |B|_F^2 + 2\text{Tr}(A^TB) \)

- Proof: \( |A + B|_F^2 = \sum_i |A_i + B_i|_2^2 \)

\[
= \sum_i |A_i|_2^2 + \sum_i |B_i|_2^2 + 2\langle A_i, B_i \rangle \\
= |A|_F^2 + |B|_F^2 + 2\text{Tr}(A^TB)
\]
Affine Embeddings: Missing Proofs

- **Claim:** $\text{Tr}(AB) \leq |A|_F|B|_F$

- **Proof:** $\text{Tr}(AB) = \sum_i \langle A^i, B_i \rangle$ for rows $A^i$ and columns $B_i$

  $\leq \sum_i |A^i|_2 |B_i|_2$ by Cauchy-Schwarz for each $i$

  $\leq \left( \sum_i |A^i|_2^2 \right)^{\frac{1}{2}} \left( \sum_i |B_i|_2^2 \right)^{\frac{1}{2}}$ another Cauchy-Schwarz

  $= |A|_F|B|_F$
Affine Embeddings: Homework Proof

- Claim: \(|SB^*|^2_F = (1 \pm \epsilon)|B^*|^2_F\) with constant probability if CountSketch matrix \(S\) has \(k = 0(\frac{1}{\epsilon^2})\) rows

- Proof is Homework #1 Problem 3 from 2017

- \(|SB^*|^2_F = \sum_i |SB_i^*|^2_2\)

- By our analysis for CountSketch and linearity of expectation, \(E[|SB^*|^2_F] = \sum_i E[|SB_i^*|^2_2] = |B^*|^2_F\)

- Bound the variance and apply Chebyshev’s inequality
Low rank approximation

- A is an n x d matrix
  - Think of n points in $\mathbb{R}^d$

- E.g., A is a customer-product matrix
  - $A_{i,j} =$ how many times customer i purchased item j

- A is typically well-approximated by low rank matrix
  - E.g., high rank because of noise

- **Goal:** find a low rank matrix approximating A
  - Easy to store, data more interpretable
What is a good low rank approximation?

**Singular Value Decomposition (SVD)**

Any matrix \( A = U \cdot \Sigma \cdot V \)
- \( U \) has orthonormal columns
- \( \Sigma \) is diagonal with non-increasing positive entries down the diagonal
- \( V \) has orthonormal rows

- Rank-k approximation: \( A_k = U_k \cdot \Sigma_k \cdot V_k \)
  - rows of \( V_k \) are the top k principal components

\[
\begin{pmatrix}
A
\end{pmatrix} =
\begin{pmatrix}
U_k
\end{pmatrix}
\begin{pmatrix}
\Sigma_k
\end{pmatrix}
\begin{pmatrix}
V_k
\end{pmatrix}
+ \begin{pmatrix}
E
\end{pmatrix}
\]
What is a good low rank approximation?

\[ A_k = \text{argmin}_{\text{rank} k \text{ matrices } B} |A-B|_F \]

\[ (|C|_F = (\sum_{i,j} C_{i,j}^2)^{1/2}) \]

Computing \( A_k \) exactly is expensive

\[
\begin{pmatrix}
A
\end{pmatrix} =
\begin{pmatrix}
U_k
\end{pmatrix}
\begin{pmatrix}
\Sigma_k
\end{pmatrix}
\begin{pmatrix}
V_k
\end{pmatrix}
+ \begin{pmatrix}
E
\end{pmatrix}
\]
Low rank approximation

- **Goal:** output a rank $k$ matrix $A'$, so that
  \[ |A - A'|_F \leq (1+\varepsilon) |A - A_k|_F \]

- Can do this in $\text{nnz}(A) + (n+d) \cdot \text{poly}(k/\varepsilon)$ time \([S,CW]\)
  - $\text{nnz}(A)$ is number of non-zero entries of $A$
Solution to low-rank approximation [S]

- Given n x d input matrix A
- Compute S*A using a random matrix S with k/ε << n rows. S*A takes random linear combinations of rows of A
  
  ![Diagram of matrix operations](image)

- Project rows of A onto SA, then find best rank-k approximation to points inside of SA.
What is the matrix $S$?

- $S$ can be a $k/\epsilon \times n$ matrix of i.i.d. normal random variables
- $[S]$ $S$ can be a $\tilde{O}(k/\epsilon) \times n$ Fast Johnson Lindenstrauss Matrix
- $[CW]$ $S$ can be a $\text{poly}(k/\epsilon) \times n$ CountSketch matrix

$$\begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

$S \cdot A$ can be computed in $\text{nnz}(A)$ time
Why do these Matrices Work?

- Consider the regression problem $\min_X |A_kX - A|_F$

- Let $S$ be an affine embedding

- Then $|SA_kX - SA|_F = (1 + \epsilon)|A_kX - A|_F$ for all $X$

- By normal equations, $\arg\min_X |SA_kX - SA|_F = (SA_k)^{-1}SA$

- So, $|A_k(SA_k)^{-1}SA - A|_F \leq (1 + \epsilon)|A_k - A|_F$

- But $A_k(SA_k)^{-1}SA$ is a rank-$k$ matrix in the row span of $SA$!

- Let’s formalize why the algorithm works now…
Why do these Matrices Work?

- \( \min_{\text{rank}-k} \|XSA - A\|^2_F \leq \|A_k(SA_k)^{-1}SA - A\|^2_F \leq (1 + \varepsilon)\|A - A_k\|^2_F \)

- By the normal equations,
  \[ \|XSA - A\|^2_F = \|XSA - A(SA)^{-1}SA\|^2_F + \|A(SA)^{-1}SA - A\|^2_F \]

- Hence,
  \[ \min_{\text{rank}-k} \|XSA - A\|^2_F = \|A(SA)^{-1}SA - A\|^2_F + \min_{\text{rank}} \|XSA - A(SA)^{-1}SA\|^2_F \]

- Can write \( SA = U \Sigma V^T \) in its thin SVD

- Then, \( \min_{\text{rank}-k} \|XSA - A(SA)^{-1}SA\|^2_F = \min_{\text{rank}} \|XU \Sigma - A(SA)^{-1}U \Sigma\|^2_F \)
  \[ = \min_{\text{rank}-k} \|Y - A(SA)^{-1}U \Sigma\|^2_F \]

- Hence, we can just compute the SVD of \( A(SA)^{-1}U \Sigma \)

- But how do we compute \( A(SA)^{-1}U \Sigma \) quickly?
Caveat: projecting the points onto SA is slow

- Current algorithm:
  1. Compute $S^*A$
  2. Project each of the rows onto $S^*A$
  3. Find best rank-k approximation of projected points inside of rowspace of $S^*A$

- Bottleneck is step 2

- [CW] Approximate the projection
  - Fast algorithm for approximate regression
    $$\min_{\text{rank-}k} |X(SA) - A|_F^2$$
    Can solve with affine embeddings
  - Want $\text{nnz}(A) + (n+d)\text{poly}(k/\epsilon)$ time
Using Affine Embeddings

- We know we can just output \( \arg \min_{\text{rank-}k} X |XSAR - A|_F^2 \)

- Choose an affine embedding \( R \):
  \[ |XSAR - AR|_F^2 = (1 \pm \epsilon)|XSA - A|_F^2 \] for all \( X \)

- Note: we can compute \( AR \) and \( SAR \) in \( \text{nnz}(A) \) time

- Can just solve \( \min_{\text{rank}} X |XSAR - AR|_F^2 \)

- \( \min_{\text{rank}} X |XSAR - AR|_F^2 = |AR(SAR)^-(SAR) - AR|_F^2 + \min_{\text{ran}} X |XSAR - AR(SAR)^-(SAR)|_F^2 \)

- Compute \( \min_{\text{rank-}} Y |Y - AR(SAR)^-(SAR)|_F^2 \) using SVD which is \( n \cdot \text{poly}\left(\frac{k}{\epsilon}\right) \) time

- Necessarily, \( Y = XSAR \) for some \( X \). Output \( Y(SAR)^-SA \) in factored form. We’re done!
Low Rank Approximation Summary

1. Compute SA

2. Compute SAR and AR

3. Compute \( \min_{\text{rank-}k \ Y} |Y - AR(SAR)^{-}(SAR)|_F^2 \) using SVD

4. Output \( Y(SAR)^{-}SA \) in factored form

Overall time: \( \text{nnz}(A) + (n+d)\text{poly}(k/\varepsilon) \)