

How to choose the right sketching matrix S ? [S]

- S is a Subsampled Randomized Hadamard Transform
 - $S = P^*H^*D$
 - D is a diagonal matrix with $+1, -1$ on diagonals
 - H is the Hadamard matrix: $H_{i,j} = (-1)^{\langle i,j \rangle} / n^{.5}$
 - P just chooses a random (small) subset of rows of H^*D
 - S^*A can be computed in $O(nd \log n)$ time

Why does it work?

Why does this work?

- We can again assume columns of A are orthonormal
- It suffices to show $|SAx|_2^2 = |PHDAx|_2^2 = 1 \pm \epsilon$ for all x
- HD is a rotation matrix, so $|HDAx|_2^2 = |Ax|_2^2 = 1$ for any x
 - Notation: let $y = Ax$
- Flattening Lemma: For any fixed y ,

$$\Pr [|HDy|_\infty \geq C \frac{\log^{.5}(\frac{nd}{\delta})}{n^{.5}}] \leq \frac{\delta}{2d}$$

Proving the Flattening Lemma

- **Flattening Lemma:** $\Pr [|\text{HDy}|_\infty \geq C \frac{\log^5 nd/\delta}{n^5}] \leq \frac{\delta}{2d}$
- Let $C > 0$ be a constant. We will show for a fixed i in $[n]$,

$$\Pr [|(HDy)_i| \geq C \frac{\log^5 nd/\delta}{n^5}] \leq \frac{\delta}{2nd}$$

- If we show this, we can apply a union bound over all i
- $|(HDy)_i| = \sum_j H_{i,j} D_{j,j} y_j$
- (Azuma-Hoeffding) For independent zero-mean random variables Z_j :

$$\Pr [|\sum_j Z_j| > t] \leq 2e^{-\frac{t^2}{2\sum_j \beta_j^2}}, \text{ where } |Z_j| \leq \beta_j \text{ with probability 1}$$

- $Z_j = H_{i,j} D_{j,j} y_j$ has 0 mean
- $|Z_j| \leq \frac{|y_j|}{n^5} = \beta_j$ with probability 1
- $\sum_j \beta_j^2 = \frac{1}{n}$

- $\Pr \left[|\sum_j Z_j| > \frac{C \log^5(\frac{nd}{\delta})}{n^5} \right] \leq 2e^{-\frac{C^2 \log^2(\frac{nd}{\delta})}{2}} \leq \frac{\delta}{2nd}$

Consequence of the Flattening Lemma

- Recall columns of A are orthonormal
- HDA has orthonormal columns
- Flattening Lemma implies $|HDAe_i|_\infty \leq C \frac{\log^5 nd/\delta}{n^{.5}}$ with probability $1 - \frac{\delta}{2d}$ for a fixed $i \in [d]$
- With probability $1 - \frac{\delta}{2}$, $|e_j HDAe_i| \leq C \frac{\log^5 nd/\delta}{n^{.5}}$ for all i, j
- Given this, $|e_j HDA|_2 \leq C \frac{d^{.5} \log^5 nd/\delta}{n^{.5}}$ for all j

(Can be optimized further)

Matrix Chernoff Bound

- Let X_1, \dots, X_s be independent copies of a symmetric random matrix $X \in \mathbb{R}^{d \times d}$ with $E[X] = 0$, $|X|_2 \leq \gamma$, and $|E[X^T X]|_2 \leq \sigma^2$. Let $W = \frac{1}{s} \sum_{i \in [s]} X_i$. For any $\epsilon > 0$,

$$\Pr[|W|_2 > \epsilon] \leq 2d \cdot e^{-s\epsilon^2 / (\sigma^2 + \frac{\gamma\epsilon}{3})}$$

(here $|W|_2 = \sup |Wx|_2 / |x|_2$)

- Let $V = HDA$, and recall V has orthonormal columns
- Suppose P in the $S = \text{PHD}$ definition samples s rows uniformly with replacement. If row i is sampled in the j -th sample, $P_{j,i} = \frac{\sqrt{n}}{\sqrt{s}}$, and is 0 otherwise
- Let Y_i be the i -th sampled row of $V = HDA$
- Let $X_i = I_d - n \cdot Y_i^T Y_i$
 - $E[X_i] = I_d - n \cdot \sum_j \left(\frac{1}{n}\right) V_j^T V_j = I_d - V^T V = 0^{d \times d}$
 - $|X_i|_2 \leq |I_d|_2 + n \cdot \max |e_j HDA|_2^2 = 1 + n \cdot C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} = \Theta\left(d \log\left(\frac{nd}{\delta}\right)\right)$

Matrix Chernoff Bound

- Recall: let Y_i be the i -th sampled row of $V = HDA$
- Let $X_i = I_d - n \cdot Y_i^T Y_i$
- $$E[X^T X + I_d] = I_d + I_d - 2n E[Y_i^T Y_i] + n^2 E[Y_i^T Y_i Y_i^T Y_i]$$

$$= 2I_d - 2I_d + n^2 \sum_i \left(\frac{1}{n}\right) \cdot v_i^T v_i v_i^T v_i = n \sum_i v_i^T v_i \cdot |v_i|_2^2$$
- Define $Z = n \sum_i v_i^T v_i C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} = C^2 d \log\left(\frac{nd}{\delta}\right) I_d$
- Note that $E[X^T X + I_d]$ and Z are real symmetric, with non-negative eigenvalues
- Claim: for all vectors y , we have: $y^T E[X^T X + I_d] y \leq y^T Z y$
- Proof: $y^T E[X^T X + I_d] y = n \sum_i y^T v_i^T v_i y |v_i|_2^2 = n \sum_i \langle v_i, y \rangle^2 |v_i|_2^2$ and

$$y^T Z y = n \sum_i y^T v_i^T v_i y C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} = d \sum_i \langle v_i, y \rangle^2 C^2 \log\left(\frac{nd}{\delta}\right)$$
- Hence, $|E[X^T X]|_2 \leq |E[X^T X] + I_d|_2 + |I_d|_2 = |E[X^T X + I_d]|_2 + 1$

$$\leq |Z|_2 + 1 \leq C^2 d \log\left(\frac{nd}{\delta}\right) + 1$$
- Hence, $|E[X^T X]|_2 = O\left(d \log\left(\frac{nd}{\delta}\right)\right)$

Matrix Chernoff Bound

- Hence, $|\mathbb{E}[X^T X]|_2 = O\left(d \log\left(\frac{nd}{\delta}\right)\right)$
- Recall: (Matrix Chernoff) Let X_1, \dots, X_s be independent copies of a symmetric random matrix $X \in \mathbb{R}^{d \times d}$ with $\mathbb{E}[X] = 0$, $|X|_2 \leq \gamma$, and $|\mathbb{E}[X^T X]|_2 \leq \sigma^2$. Let $W = \frac{1}{s} \sum_{i \in [s]} X_i$. For any $\epsilon > 0$, $\Pr[|W|_2 > \epsilon] \leq 2d \cdot e^{-s\epsilon^2/(\sigma^2 + \frac{\gamma\epsilon}{3})}$

$$\Pr\left[|I_d - (\text{PHDA})^T(\text{PHDA})|_2 > \epsilon\right] \leq 2d \cdot e^{-s\epsilon^2/(\Theta(d \log(\frac{nd}{\delta})))}$$

- Set $s = d \log\left(\frac{nd}{\delta}\right) \frac{\log(\frac{d}{\delta})}{\epsilon^2}$, to make this probability less than $\frac{\delta}{2}$

SRHT Wrapup

- Have shown $\|I_d - (\text{PHDA})^T(\text{PHDA})\|_2 < \epsilon$ using Matrix Chernoff Bound and with $s = d \log\left(\frac{nd}{\delta}\right) \frac{\log\left(\frac{d}{\delta}\right)}{\epsilon^2}$ samples
- Implies for every unit vector x ,
$$|1 - |\text{PHDA}x|_2^2| = |x^T x - x^T (\text{PHDA})^T (\text{PHDA}) x| < \epsilon,$$
so $|\text{PHDA}x|_2^2 \in 1 \pm \epsilon$ for all unit vectors x
- Considering the column span of A adjoined with b , we can again solve the regression problem
- The time for regression is now only $O(nd \log n) + \text{poly}\left(\frac{d \log(n)}{\epsilon}\right)$. Nearly optimal in matrix dimensions ($n \gg d$)

Faster Subspace Embeddings S [CW,MM,NN]

- CountSketch matrix
- Define $k \times n$ matrix S , for $k = O(d^2/\epsilon^2)$
- S is really sparse: single randomly chosen non-zero entry per column

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

Can compute $S \cdot A$ in $\text{nnz}(A)$ time!

- $\text{nnz}(A)$ is number of non-zero entries of A

Simple Proof [Nguyen]

- Can assume columns of A are orthonormal
- Suffices to show $|SAx|_2 = 1 \pm \varepsilon$ for all unit x
 - For regression, apply S to $[A, b]$
- SA is a $6d^2/(\delta\varepsilon^2) \times d$ matrix
- Suffices to show $\|A^T S^T SA - I\|_2 \leq \|A^T S^T SA - I\|_F \leq \varepsilon$
- Matrix product result shown below:
$$\Pr[|CS^TSD - CD|_F^2 \leq [6/(\delta(\# \text{ rows of } S))] * |C|_F^2 |D|_F^2] \geq 1 - \delta$$
- Set $C = A^T$ and $D = A$.
- Then $|A|_F^2 = d$ and $(\# \text{ rows of } S) = 6 d^2/(\delta\varepsilon^2)$

Matrix Product Result [Kane, Nelson]

- Show: $\Pr[|CS^TSD - CD|_F^2 \leq [6/(\delta(\# \text{ rows of } S))] * |C|_F^2 |D|_F^2] \geq 1 - \delta$
- (JL Property) A distribution on matrices $S \in \mathbb{R}^{k \times n}$ has the (ϵ, δ, ℓ) -JL moment property if for all $x \in \mathbb{R}^n$ with $|x|_2 = 1$,

$$E_S \left| |Sx|_2^2 - 1 \right|^\ell \leq \epsilon^\ell \cdot \delta$$
- (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 (ϵ, δ, ℓ) -JL moment property for some $\ell \geq 2$. Then for A, B matrices with n rows,

$$\Pr_S \left[|A^T S^T S B - A^T B|_F \geq 3 \epsilon |A|_F |B|_F \right] \leq \delta$$

From Vectors to Matrices

- (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 (ϵ, δ, ℓ) -JL moment property for some $\ell \geq 2$. Then for A, B matrices with n rows,

$$\Pr_S \left[\left| A^T S^T S B - A^T B \right|_F \geq 3 \epsilon |A|_F |B|_F \right] \leq \delta$$

- Proof: For a random scalar X , let $|X|_p = (E|X|^p)^{1/p}$
 - Sometimes consider $X = |T|_F$ for a random matrix T
 - $\left| |T|_F \right|_p = \left(E \left[|T|_F^p \right] \right)^{1/p}$
- Can show $|\cdot|_p$ is a norm if $p \geq 1$
 - Minkowski's Inequality: $|X + Y|_p \leq |X|_p + |Y|_p$
- For unit vectors x, y , we will bound $|\langle Sx, Sy \rangle - \langle x, y \rangle|_\ell$

Minkowski's Inequality

- Minkowski's Inequality: $|X + Y|_p \leq |X|_p + |Y|_p$

- Proof:

- If $|X|_p, |Y|_p$ are finite, then so is $|X + Y|_p$. **Why?**

- $f(x) = x^p$ is convex for $p \geq 1$, so for any fixed x, y :

$$|.5x + .5y|^p \leq |.5x| + |.5y|^p \leq .5|x|^p + .5|y|^p, \text{ so}$$
$$|x + y|^p \leq 2^{p-1}(|x|^p + |y|^p)$$

- So, $E[|X + Y|_p^p] \leq E[2^{p-1}(|X|_p^p + |Y|_p^p)]$

- $|X + Y|_p^p = \int |x + y|^p d\mu$

$$= \int |x + y| \cdot |x + y|^{p-1} d\mu$$

$$\leq \int (|x| + |y|) |x + y|^{p-1} d\mu$$

$$= \int |x| |x + y|^{p-1} d\mu + \int |y| |x + y|^{p-1} d\mu$$

$$\leq \left(\left(\int |x|^p d\mu \right)^{\frac{1}{p}} + \left(\int |y|^p d\mu \right)^{\frac{1}{p}} \right) \left(\int |x + y|^{(p-1) \left(\frac{p}{p-1} \right)} d\mu \right)^{\frac{p-1}{p}}$$

$$= (|X|_p + |Y|_p) |X + Y|_p^{p-1}$$

From Vectors to Matrices

- For unit vectors x, y , $|\langle Sx, Sy \rangle - \langle x, y \rangle|_\ell$

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

$$\leq \frac{1}{2} (||Sx|_2^2 - 1|_\ell + ||Sy|_2^2 - 1|_\ell + ||S(x-y)|_2^2 - |x-y|_2^2|_\ell)$$

$$\leq \frac{1}{2} (\epsilon \cdot \delta^\ell + \epsilon \cdot \delta^\ell + |x-y|_2^2 \epsilon \cdot \delta^\ell)$$

$$\leq 3 \epsilon \cdot \delta^\ell$$
- By linearity, for arbitrary x, y , $\frac{|\langle Sx, Sy \rangle - \langle x, y \rangle|_\ell}{|x|_2 |y|_2} \leq 3 \epsilon \cdot \delta^\ell$
- Suppose A has d columns and B has e columns. Let the columns of A be A_1, \dots, A_d and the columns of B be B_1, \dots, B_e
- Define $X_{i,j} = \frac{1}{|A_i|_2 |B_j|_2} \cdot (\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle)$
- $|A^T S^T S B - A^T B|_F^2 = \sum_i \sum_j |A_i|_2^2 \cdot |B_j|_2^2 X_{i,j}^2$

From Vectors to Matrices

- Have shown: for arbitrary x, y , $\frac{|\langle Sx, Sy \rangle - \langle x, y \rangle|}{\|x\|_2 \|y\|_2} \leq 3\epsilon \cdot \delta^{\frac{1}{\ell}}$
- For $X_{i,j} = \frac{1}{\|A_i\|_2 \|B_j\|_2} \cdot (\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle)$: $\|A^T S^T S B - A^T B\|_F^2 = \sum_i \sum_j \|A_i\|_2^2 \cdot \|B_j\|_2^2 X_{i,j}^2$
- $$\begin{aligned} \|A^T S^T S B - A^T B\|_F^2 &= \left| \sum_i \sum_j \|A_i\|_2^2 \cdot \|B_j\|_2^2 X_{i,j}^2 \right|_{\ell/2} \\ &\leq \sum_i \sum_j \|A_i\|_2^2 \cdot \|B_j\|_2^2 |X_{i,j}|_{\ell/2}^2 \\ &= \sum_i \sum_j \|A_i\|_2^2 \cdot \|B_j\|_2^2 |X_{i,j}|_{\ell}^2 \\ &\leq \left(3\epsilon\delta^{\frac{1}{\ell}}\right)^2 \sum_i \sum_j \|A_i\|_2^2 \|B_j\|_2^2 \\ &= \left(3\epsilon\delta^{\frac{1}{\ell}}\right)^2 \|A\|_F^2 \|B\|_F^2 \end{aligned}$$
- Since $E \left[\|A^T S^T S B - A^T B\|_F^{\ell} \right] = \left\| \|A^T S^T S B - A^T B\|_F^2 \right\|_{\ell/2}^{\ell/2}$, by Markov's inequality,
- $\Pr \left[\|A^T S^T S B - A^T B\|_F > 3\epsilon \|A\|_F \|B\|_F \right] \leq \left(\frac{1}{3\epsilon \|A\|_F \|B\|_F} \right)^{\ell} E \left[\|A^T S^T S B - A^T B\|_F^{\ell} \right] \leq \delta$

Result for Vectors

- Show: $\Pr[|CS^TSD - CD|_F^2 \leq [6/(\delta(\# \text{ rows of } S))] * |C|_F^2 |D|_F^2] \geq 1 - \delta$
- (JL Property) A distribution on matrices $S \in \mathbb{R}^{k \times n}$ has the (ϵ, δ, ℓ) -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 $\epsilon, \delta \in (0, \frac{1}{2})$, let D be a distribution on matrices S with k rows and n columns that satisfies the (ϵ, δ, ℓ) -JL moment property for some $\ell \geq 2$. Then for A, B matrices with n rows,

$$\Pr_S \left[|A^T S^T S B - A^T B|_F \geq 3 \epsilon |A|_F |B|_F \right] \leq \delta$$

- 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 (ϵ, δ, ℓ) -JL moment property if for all $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$,

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

- We show this property holds with $\ell = 2$. First, let us consider $\ell = 1$

- 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

- $$\begin{aligned} \mathbb{E}[\|Sx\|_2^2] &= \sum_{j \in [k]} \mathbb{E} \left[\left(\sum_{i \in [n]} \delta(h(i) = j) \sigma_i x_i \right)^2 \right] \\ &= \sum_{j \in [k]} \sum_{i_1, i_2 \in [n]} \mathbb{E} [\delta(h(i_1) = j) \delta(h(i_2) = j) \sigma_{i_1} \sigma_{i_2}] x_{i_1} x_{i_2} \\ &= \sum_{j \in [k]} \sum_{i \in [n]} \mathbb{E} [\delta(h(i) = j)^2] x_i^2 \\ &= \left(\frac{1}{k} \right) \sum_{j \in [k]} \sum_{i \in [n]} x_i^2 = \|x\|_2^2 \end{aligned}$$

CountSketch Satisfies the JL Property

- $$E[|Sx|_2^4] = E[\sum_{j \in [k]} \sum_{j' \in [k]} (\sum_{i \in [n]} \delta(h(i) = j) \sigma_i x_i)^2 (\sum_{i' \in [n]} \delta(h(i') = j') \sigma_{i'} x_{i'})^2] =$$

$$\sum_{j_1, j_2, i_1, i_2, i_3, i_4} E[\sigma_{i_1} \sigma_{i_2} \sigma_{i_3} \sigma_{i_4} \delta(h(i_1) = j_1) \delta(h(i_2) = j_1) \delta(h(i_3) = j_2) \delta(h(i_4) = j_2)] x_{i_1} x_{i_2} x_{i_3} x_{i_4}$$
- 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 $\sum_j \frac{1}{k} \sum_i x_i^4 = |x|_4^4$
- Suppose $i_1 = i_2$ and $i_3 = i_4$ but $i_1 \neq i_3$. Then get $\sum_{j_1, j_2, i_1, i_3} \frac{1}{k^2} x_{i_1}^2 x_{i_3}^2 = |x|_2^4 - |x|_4^4$
- Suppose $i_1 = i_3$ and $i_2 = i_4$ but $i_1 \neq i_2$. Then necessarily $j_1 = j_2$. Obtain $\sum_j \frac{1}{k^2} \sum_{i_1, i_2} x_{i_1}^2 x_{i_2}^2 \leq \frac{1}{k} |x|_2^4$. Obtain same bound if $i_1 = i_4$ and $i_2 = i_3$.
- Hence, $E[|Sx|_2^4] \in [|x|_2^4, |x|_2^4(1 + \frac{2}{k})] = [1, 1 + \frac{2}{k}]$
- So, $E_S ||Sx|_2^2 - 1|^2 \leq (1 + \frac{2}{k}) - 2 + 1 = \frac{2}{k}$. Setting $k = \frac{2}{\epsilon^2 \delta}$ finishes the proof

Where are we?

- (JL Property) A distribution on matrices $S \in \mathbb{R}^{k \times n}$ has the (ϵ, δ, ℓ) -JL moment property if for all $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$,

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

- (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 (ϵ, δ, ℓ) -JL moment property for some $\ell \geq 2$. Then for A, B matrices with n rows,

$$\Pr_S \left[\left| \|A^T S^T S B\|_F^2 - \|A^T B\|_F^2 \right| \geq 3 \epsilon^2 \|A\|_F^2 \|B\|_F^2 \right] \leq \delta$$

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

- Matrix product result we wanted was:

$$\Pr \left[\|CS^TSD - CD\|_F^2 \leq \frac{6}{\delta k} \|C\|_F^2 \|D\|_F^2 \right] \geq 1 - \delta$$

- 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