CSE 392: Matrix and Tensor Algorithms for Data Spring 2024

Lecture 8 — February 12, 2024

Instructor: Shashanka Ubaru *Scribe: Anish Pandya*

Last class, we were talking about leverage score sampling which improves the upper bound to $1 + \epsilon\left(\|\mathbf{A} - \mathbf{A_k}\|_2^2\right)$. This class, we began talking more about sketching and types of sketching matricies. We first discussed Gaussian sketching, then the Approximate Matrix Multiplication and JL Moment properties, and finally we discussed SRHT (Hadammard Transform). We plan to discuss the Countsketch algorithm next class.

Gaussian Embedding

Definition. Embedding. A matrix $\mathbf{S} \in \mathbb{R}^{n\times m}$ is an $\epsilon-$embedding of a set $\mathcal{P} \subset \mathbb{R}^n$ if for every $\mathbf{y} \in \mathcal{P}$,

$$\|\mathbf{Sy}\|_2 = (1 \pm \epsilon)\|y\|_2$$

We say that $\mathbf{S}$ is the sketching matrix. Essentially, for a matrix to be an ϵ-embedding, we need to show that it preserves the norm within ϵ for all $\|\mathbf{y}\|_2$. At a high level, a lot of problems boil down to finding a subspace embedding, as we will explore in the next couple classes. Problems like least-squares and low rank approximation are some problems where subspace embedding techniques come handy. Now, we consider vector embedding property also known as the **distributional JL lemma**:

Definition. Vector Embedding. Let $\mathbf{S} \in \mathbb{R}^{m\times d}$ have independent entries given by $s_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0,1)$. If $m = \mathcal{O}\left(\log\left(1/\delta\right)/\epsilon^2\right)$. Then, for any vector, $\mathbf{y} \in \mathbb{R}^d$, $\epsilon \in (0, 1]$:

$$\|\mathbf{Sy}\|_2^2 = (1 \pm \epsilon)\|y\|_2^2$$

with probability $(1 - \delta)$.

Proof. From the definition of the two norm, we have:

$$\begin{aligned}
\|\mathbf{Sy}\|_2^2 &= \frac{1}{m}\sum_{i=1}^{m}\left(\langle s_i, y\rangle\right)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{d} s_{ij}y_j\right)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{d} \frac{1}{\sqrt{m}}\mathcal{N}(0,1)y_j\right)^2 \\
&= \frac{1}{m^{3/2}}\sum_{i=1}^{m}\left(\mathcal{N}\left(0, \|\mathbf{y}\|_2^2\right)\right)^2
\end{aligned}$$

Notice that this is the χ^2 distribution. Now, let's use this property of the χ^2 distribution:

Remark 1. Let z be a χ^2 random variable with m degrees of freedom, then:

$$\mathbb{P}\left(|z - \mathbb{E}[z] \geq \epsilon\, \mathbb{E}[z]\right) \leq 2e^{-\epsilon^2 m/8}$$

Let $z = \|\mathbf{Sy}\|_2^2$. $\mathbb{E}[\|\mathbf{Sy}\|_2^2] = \|\mathbf{y}\|_2^2$.

$$\mathbb{P}\left(|\|\mathbf{Sy}\|_2^2 \geq (\epsilon+1)\|\mathbf{y}\|_2^2\right) \leq 2e^{-\epsilon^2 m/8} \leq \delta$$

Now, we can solve the inequality for m:

$$\begin{aligned} 2e^{-\epsilon^2 m/8} &< \delta \\ \frac{\epsilon^2 m}{8} &> -\ln\left(\frac{\delta}{2}\right) \\ m &> \frac{8}{\epsilon^2}\ln(\frac{2}{\delta}) \end{aligned}$$

So then, m must be $\mathcal{O}\left(\log(1/\delta)/\epsilon^2\right)$.

$$\begin{aligned} \mathbb{P}\left((\epsilon-1)\|\mathbf{y}\|_2^2 \leq \|\mathbf{Sy}\|_2 \leq (\epsilon+1)\|\mathbf{y}\|_2^2\right) &= 1 - \mathbb{P}\left(\|\mathbf{Sy}\|_2^2 \geq (\epsilon+1)\|\mathbf{y}\|_2^2\right) \\ \mathbb{P}\left((\epsilon-1)\|\mathbf{y}\|_2^2 \leq \|\mathbf{Sy}\|_2 \leq (\epsilon+1)\|\mathbf{y}\|_2^2\right) &\leq 1 - \delta \end{aligned}$$

Next, we considered the **JL-lemma**:

Lemma 1. ***JL-Lemma.*** *Let $\mathbf{S} \in \mathbb{R}^{m\times d}$ have independent entries, $s_{ij} \sim \frac{1}{m}\mathcal{N}(0,1)$. If $m = \mathcal{O}(\log(n)/\epsilon^2)$, then for any n data points, with probability at least $9/10$:*

$$(1-\epsilon)\left\|\mathbf{x_i} - \mathbf{x_j}\right\|_2 \leq \left\|\mathbf{Sx_i} - \mathbf{Sx_j}\right\|_2 \leq (1+\epsilon)\left\|\mathbf{x_i} - \mathbf{x_j}\right\|_2$$

We fix $i, j \in [d]$. Let $\mathbf{y} = \mathbf{x_i} - \mathbf{x_j}$. By the Distributional JL $\|\mathbf{S}(\mathbf{x_j} - \mathbf{x_i})\|_2 = (1 \pm \epsilon)\|\mathbf{x_i} - \mathbf{x_j}\|_2$. Let $\delta = 1/n^2$. Then, there are less than n^2 (i,j) pairs, by a union bound, we have ■

Theorem 1. ***Subspace Embedding.*** *Let $\boldsymbol{S} \in \mathbb{R}^{m\times n}$ have independent entries, $s_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0,1)$. If $m = \mathcal{O}(d\log(1/\delta)/\epsilon^2$, then for a given $\mathbf{A} \in \mathbb{R}^{n\times d}$, with probability of at least $1-\delta$.*

Embedding a $d-$dimensional subspace $\mathcal{U} = \text{span}(\mathbf{A}) = \text{span}(\mathbf{U}) \in \mathbb{R}^n$. Then,

$$\|\mathbf{SUx}\|_2 = (1 \pm \epsilon)\|\mathbf{x}\|_2$$

or,

$$\|\mathbf{U}^T\mathbf{S}^T\mathbf{SU} - \mathbf{I}\|_2 \leq \epsilon$$

From the epsilon-net argument, we know that $|\mathcal{N}(\epsilon)| \leq (1 + (2/\epsilon))^d$. If $\mathbf{S}$ is distributional JL with failure with probability δ', taking the union of the $\epsilon-$net size, we get the result:

$$m = \mathcal{O}\left(\frac{d\log(1/\delta)}{\epsilon^2}\right)$$

Approximate Matrix Multiplication and JL Moment

Theorem 2. *Given* $\mathbf{A} \in \mathbb{R}^{n\times d}, \mathbf{A}, n \geq d$, *and* $\mathbf{B} \in \mathbb{R}^{d'\times n}$. *Let* r *be the rank of* $\mathbf{A}$, *and let* ϵ *and* $\delta \in \mathbb{R}$ *be greater than zero. Let* $\mathbf{S}$ *be chosen such that, with probability of at least* $1-\delta$*:*

$$\|\mathbf{B}\mathbf{S}^{\mathbf{T}}\mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F \leq \epsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F$$

Then, $\mathbf{S}$ *is an* $\epsilon r-$*embedding of span*$(\mathbf{A})$.

Proof. Let $\mathbf{B} = \mathbf{A}^T$, and since $\mathbf{S}$ is chosen. Then,

$$\|\mathbf{A}^{\mathbf{T}}\mathbf{S}^{\mathbf{T}}\mathbf{S}\mathbf{A} - \mathbf{I}\|_2 \leq \|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{A} - \mathbf{I}\|_F \leq \epsilon\|\mathbf{A}\|_F^2 = \epsilon r$$

■

Fact 1. JL Moment. A distribution on $\mathbf{S} \in \mathbb{R}^{m\times d}$, has the $(\epsilon, \delta, \ell)-$JL moment property if for every $\mathbf{y} \in \mathbb{R}^d$ with $\|\mathbf{y}\|_2 = 1$,

$$\mathbb{E}\left[|\|\mathbf{S}\mathbf{y}\|_2^2 - 1|^\ell\right] \leq \epsilon^\ell\delta$$

Notably for $\ell = 2$ and $\mathbb{E}\left[\|\mathbf{S}\mathbf{y}\|_2\right] = 1$, we have:

$$\mathrm{Var}\left(\|\mathbf{S}\mathbf{y}\|_2^2\right) \leq \epsilon^2\delta$$

We will also need to recall the Polarization identity:

Fact 2. The polarization identity for two vectors $\mathbf{a}, \mathbf{b}$ in $\mathbb{R}^n$ is:

$$\begin{aligned}\langle\mathbf{a}, \mathbf{b}\rangle &= \frac{1}{4}\left(\|\mathbf{a}+\mathbf{b}\|_2^2 - \|\mathbf{a}-\mathbf{b}\|_2^2\right)\\ &= \frac{1}{2}\left(\|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 - \|\mathbf{a}-\mathbf{b}\|_2^2\right)\\ &= \frac{1}{2}\left(\|\mathbf{a}+\mathbf{b}\|_2^2 - \|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2\right)\end{aligned}$$

Theorem 3. ***JL Moment and AMM.*** *Given* $\mathbf{A} \in \mathbb{R}^{n\times d}, \mathbf{B} \in \mathbb{R}^{d'\times n}$, *and* $\epsilon, \delta > 0$ *in* $\mathbb{R}$, *and* $\mathbf{S}$ *satisfying the* $(\epsilon, \delta, \ell)-$*JL moment property for* $\ell \geq 2$, *then we have the following with probability at least* $1-\delta$*:*

$$\|\mathbf{S}\mathbf{S}^T\mathbf{S}\mathbf{A} - \mathbf{B}\mathbf{A}\|_F \leq 3\epsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F$$

Proof. We refer to the proof given of Theorem 2.8 in Dr. Woodruff's text, which is also in the paper by Kane and Nelson. Consider two vectors, $\mathbf{x}, \mathbf{y}$ in $\mathbb{R}^d$. Then, by the second polarization identity,

$$\frac{\langle\mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y}\rangle}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2} = \frac{\|\mathbf{S}\mathbf{x}\|_2^2 + \|\mathbf{S}\mathbf{y}\|_2^2 - \|\mathbf{S}(\mathbf{x}-\mathbf{y})\|_2^2}{2}$$

The moment norm is defined as $\|X\|_\ell = \left(\mathbb{E}[X^\ell]\right)^{1/\ell}$. From the moment norm, we can use Minkowski's inequality to show the result. Minkowski's inequality is the Triangle inequality for the moment

norm. Let $X = \|\langle \mathbf{S}\hat{\mathbf{x}}, \mathbf{S}\hat{\mathbf{y}}\rangle - \langle \hat{\mathbf{x}}, \hat{\mathbf{y}}\rangle\|_\ell$, where we consider the unit vectors $\hat{\mathbf{x}}, \hat{\mathbf{y}}$. Then, again from the second Polarization identity, we see:

$$\langle \mathbf{S}\hat{\mathbf{x}}, \mathbf{S}\hat{\mathbf{y}}\rangle - \langle \hat{\mathbf{x}}, \hat{\mathbf{y}}\rangle = \frac{(\|\mathbf{S}\hat{\mathbf{x}}\|_2^2 - 1) + (\|\mathbf{S}\hat{\mathbf{x}}\|_2^2 - 1) - (\|\mathbf{S}(\hat{\mathbf{x}} + \hat{\mathbf{y}})\|_2^2 - \|\hat{\mathbf{x}} + \hat{\mathbf{y}}\|_2^2)}{2}$$

So then using the Minkowski inequality and the **JL-moment property** we see:

$$\begin{aligned}
\|\langle \mathbf{S}\hat{\mathbf{x}}, \mathbf{S}\hat{\mathbf{y}}\rangle - \langle \hat{\mathbf{x}}, \hat{\mathbf{y}}\rangle\|_\ell &= \frac{1}{2}\left(\|(\|\mathbf{S}\hat{\mathbf{x}}\|_2^2 - 1) + (\|\mathbf{S}\hat{\mathbf{x}}\|_2^2 - 1) - \left(\|\mathbf{S}(\hat{\mathbf{x}} - \hat{\mathbf{y}})\|_2^2 - \|\hat{\mathbf{x}} + \hat{\mathbf{y}}\|_2^2\right)\|_\ell\right) \\
&\leq \frac{1}{2}\left(\|(\|\mathbf{S}\hat{\mathbf{x}}\|_2^2 - 1)\|_\ell + \|(\|\mathbf{S}\hat{\mathbf{x}}\|_2^2 - 1)\|_\ell - \|\left(\|\mathbf{S}(\hat{\mathbf{x}} - \hat{\mathbf{y}})\|_2^2 - \|\hat{\mathbf{x}} + \hat{\mathbf{y}}\|_2^2\right)\|_\ell\right) \\
&\leq \frac{1}{2}\left(\epsilon\delta^{1/\ell} + \epsilon\delta^{1/\ell} - \|\hat{\mathbf{x}} + \hat{\mathbf{y}}\|_2^2 \cdot \epsilon\delta^{1/\ell}\right) \\
&\leq 3\epsilon\delta^{1/\ell}
\end{aligned}$$

For arbitrary $\mathbf{x}, \mathbf{y}$, then we have the inequality:

$$\frac{\|\langle \mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y}\rangle - \langle \mathbf{x}, \mathbf{y}\rangle\|_\ell}{\|\mathbf{x}\|_2\mathbf{y}\|_2} \leq 3\epsilon\delta^{1/\ell}$$

Now, we define a random variable, where $\mathbf{A_i}$

$$X_{ij} = \frac{1}{\|\mathbf{A_i}\|_2\|\mathbf{B_j}\|_2} \cdot (\langle \mathbf{S}\mathbf{A_i}, \mathbf{S}\mathbf{B_j}\rangle - \langle \mathbf{A_i}, \mathbf{B_j}\rangle)$$

So then,

$$\begin{aligned}
\|\|\mathbf{A^T}\mathbf{S^T}\mathbf{S}\mathbf{B} - \mathbf{A^T}\mathbf{B}\|_F^2\|_{\ell/2} &= \|\sum_{i=1}^{n}\sum_{j=1}^{m}\|\mathbf{A_i}\|_2^2 \cdot \|\mathbf{B_j}\|_2^2 X_{ij}^2\|_{\ell/2} \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{m}\|\mathbf{B}_j\|_2^2\|\mathbf{A}_i\|_2^2\|\mathbf{X_{ij}^2}\|_{\ell/2} \\
&\leq \left(3\epsilon\delta^{1/\ell}\right)^2\|\mathbf{A}\|_F^2\|\mathbf{B}\|_F^2
\end{aligned}$$

If $\mathbb{E}[\|\mathbf{A^T}\mathbf{S^T}\mathbf{S}\mathbf{B} - \mathbf{A^T}\mathbf{B}\|_F^\ell] = \|\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B}\|_F^2\|_{\ell/2}^{\ell/2}$, following Markov's Inequality, we have:

$$\begin{aligned}
\mathbb{P}\left(\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B} - \mathbf{A}^T\mathbf{B}\|_F \geq 3\epsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F\right) &\leq \frac{1}{(3\epsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F)^\ell}\,\mathbb{E}[\|\mathbf{A^T}\mathbf{S^T}\mathbf{S}\mathbf{B} - \mathbf{A^T}\mathbf{B}\|_F^\ell] \\
&\leq \delta
\end{aligned}$$

■

Subsapled Randomized Hadamard Transform (SRHT).

The SRHT is a matrix: **PHD** Let $\mathbf{D} \in \mathbb{R}^{n\times n}, \mathbf{H} \in \mathbb{R}^{n\times n}, \mathbf{P} \in \mathbb{R}^{m\times n}$. D is a diagnol matrix. Then, D is a diagnol matrix with entries that are independent and identidcally distrbuted with entries either $+1$ or -1.$\mathbf{H}$ is a Hadamard Matrix, and $\mathbf{P}$ is a matrix that uniformly samples the rows of $\mathbf{HD}$.

Now, let's review some properties of $\mathbf{H}$, a Hadamard matrix. Hadamard matricies have a recursive structure, as defined by:

$$\mathbf{H_0} = [1]$$
$$\mathbf{H}_{i+1} = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{H}_i & \mathbf{H}_i \\ \mathbf{H}_i & -\mathbf{H}_i \end{pmatrix}$$

Another definition of the Hadamard matrix is $\mathbf{H_n H_n}^T = n\mathbb{I}_n$, where $\mathbb{I}_n$ is the $n \times n$ identity. Note some important properties of Hadamard matrixies:

Fact 3. Hadamard matricies are orthogonal:

$$\mathbf{H}_i^T \mathbf{H}_i = \mathbb{I}_n$$

Fact 4. For any $\mathbf{x} \in \mathbb{R}^n, n = 2^k, k \in \mathbb{N}$, we have $\|\mathbf{Hx}\|_2 = \|\mathbf{x}\|_2$ and $\|\mathbf{HDx}\|_2 = \|\mathbf{x}\|_2$

Fact 5. Matrix-vector multiplication can be computed in $\mathcal{O}(n \log(n)$ time for $\mathbf{x} \in \mathbb{R}^n$, $n = 2k$ for some $k \in \mathbb{N}$.

Fact 6. Let $\mathbf{x} = [\mathbf{x}_1, \mathbf{x_2}]^T \in \mathbb{R}^{2^k}$ for $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^{2^k - 1}$. Then:

$$\mathbf{H}_i \mathbf{x} = \begin{pmatrix} \mathbf{H}_i \mathbf{x_1} + \mathbf{H}_i \mathbf{x_2} \\ \mathbf{H}_i \mathbf{x_1} - \mathbf{H}_i \mathbf{x_2} \end{pmatrix}$$

This is a linear time operation from $\mathbf{H}_i \mathbf{x_1}, \mathbf{H}_i \mathbf{x_2}$.

Lemma 2. ***SRHT Mixing Lemma.*** *Let* $\mathbf{H}$ *be an* $n \times n$ *Hadamard Matrix,* $\mathbf{D}$*, a random diagnol* ± 1 *matrix. Let* $\mathbf{z} = \mathbf{HDx}$ *for* $\mathbf{x} \in \mathbb{R}^{\mathbf{n}}$*. With probability* $1 - \delta$*, for all* i*, simultaneously, we have:*

$$z_i^2 \leq \frac{c \log(n/\delta))}{n} \|\mathbf{z}\|_2^2$$

Theorem 4. ***Rademacher Concentration.*** *Let* $r_1, \ldots, r_n$ *be Rademacher random variables. Then for any* $\mathbf{a} \in \mathbb{R}^n$*,*

$$\mathbb{P}\left[\sum_{i=1}^{n} r_i a_i \geq t\|\mathbf{a}\|_2\right] \leq e^{-t^2/2}$$

Lemma 3. ***Fast JL.*** *Let* $\mathbf{S} = \mathbf{PHD} \in \mathbb{R}^{\mathbf{m} \times \mathbf{n}}$ *be a subsampled randomized Hadamard transform (SRHT), with* $m = \mathcal{O}\left(\frac{\log(n/\delta) \log 1/\delta}{\epsilon^2}\right)$*. Then, with any fixed* $\mathbf{x} \in \mathbb{R}^n$*, with probability* $1 - \delta$*,*

$$\|\mathbf{Sx}\|_2^2 = (1 \pm \epsilon)\|\mathbf{x}\|_2^2$$

Theorem 5. ***SRHT Embedding.*** *For* $\mathbf{S} = PHD \in \mathbb{R}^{m \times n}$*, and* $A \in \mathbb{R}^{n \times d}$*, if* $m = \mathcal{O}\left(\frac{d \log(n/\delta) \log 1/\delta}{\epsilon^2}\right)$*, then with probability of at least* $1 - \delta$*,*

$$\|\mathbf{SAx}\|_2 \leq (1 \pm \epsilon)\|\mathbf{Ax}\|_2$$