CS 15-851: Algorithms for Big Data	Spring 2025
Lecture 3 — 01/30/2025	
Prof. David Woodruff	*Scribe: Shalini Panthangi*

1 Proof that Count-Sketch Satisfies the JL Property

From the previous scribe notes, we have seen the proof that if CountSketch satisfies the JL-moment property, then we are able to show that we now have an approximate matrix product. Let's quickly recall the definitions of the relevant properties below:

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 $\|x\|_2 = 1$,

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

Approximate Matrix Product Property

For $\epsilon, \delta \in \left(0, \frac{1}{2}\right)$, let $\mathcal{D}$ 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 matrices A, B with n rows,

$$\mathbb{P}_S \left[\left\| A^T S^T S B - A^T B \right\|_F \geq 3\epsilon \|A\|_F \|B\|_F \right] \leq \delta.$$

We want to show that the JL Property holds the distribution D with $\ell = 2$.

1.1 Defining Count-Sketch Succinctly

t-Wise Independent Hash Families Definition

This concept is usually called *k-wise independent hash families*, but we will be using the variable t in place of the variable k, since we have k defined as something else previously.
A **t-wise independent hash family** is a collection of hash functions with the property that, for any t distinct inputs, the hashed values are uniformly and independently distributed.
A family of hash functions $\mathcal{H} = \{h : U \to [m]\}$ is called **t-wise independent** if for any distinct $x_1, x_2, ..., x_t \in U$ and any $y_1, y_2, ..., y_t \in [m]$:

$$\mathbb{P}[h(x_1) = y_1, h(x_2) = y_2, ..., h(x_t) = y_t] = \frac{1}{m^t}$$

This ensures that the values $h(x_1), h(x_2), ..., h(x_t)$ are uniformly and independently distributed.

1.1.1 Hash funtions involved in CountSketch

We define $h : [n] \to [k]$ to be a **2-wise independent hash function** which takes in a column index and returns the row in which that column should have a non-zero entry.

We define $\sigma : [n] \to \{-1, 1\}$ to be a **4-wise independent hash function** which takes in the column index and returns 1 or -1, representing the sign of the non-zero entry in that column.

1.2 Proving the JL Property with $\ell = 1$

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

We start with the expected squared norm:

$$\mathbb{E}[|Sx|_2^2] = \sum_{j\in[k]} \mathbb{E}\left[\left(\sum_{i\in[n]} \delta(h(i) = j)\sigma_i x_i\right)^2\right]$$

We get the next step by expanding the square and introducing i_1 and i_2 in order to represent two indices,

$$= \sum_{j\in[k]} \sum_{i_1,i_2\in[n]} \mathbb{E}[\delta(h(i_1) = j)\delta(h(i_2) = j)]\sigma_{i_1}\sigma_{i_2}x_{i_1}x_{i_2}$$

Now, this can be written as

$$= \sum_{j\in[k]} \sum_{i_1,i_2\in[n]} \mathbb{E}[\delta(h(i_1) = j)\delta(h(i_2) = j)] \cdot \mathbb{E}[\sigma_{i_1}\sigma_{i_2}] \cdot x_{i_1}x_{i_2}$$

Now, we notice that there are two cases to simplify the above equation. If $i_1 \neq i_2$, then we can rewrite $\mathbb{E}[\sigma_{i_1}\sigma_{i_2}]$ as $\mathbb{E}[\sigma_{i_1}]\mathbb{E}[\sigma_{i_2}]$. The expectation of each of these is 0, since we choose 1 out of the set $\{-1, 1\}$, giving us a mean of 0. However, if $i_1 = i_2$, then the i_1'th element hashes to the j'th bucket and the i_2'th element also hashes to the same j'th bucket, indicating that $i_1 = i_2$. In that case, we only need to consider the case where we now have the same element index i to get

$$= \sum_{j\in[k]} \sum_{i\in[n]} \mathbb{E}[\delta(h(i) = j)^2]x_i^2$$

We note the property that the square of an indicator variable is the same as the indicator variable itself. So, to find $\mathbb{E}[\delta(h(i) = j)]$, we know that since h is a 2-wise independent hash function, the probability that any given element mapped to a particular row is $\frac{1}{k}$. We replace $\mathbb{E}[\delta(h(i) = j)]$ with $\frac{1}{k}$ and pull that factor out of the sum.

$$= \left(\frac{1}{k}\right) \sum_{j\in[k]} \sum_{i\in[n]} x_i^2 = |x|_2^2$$

Since we are adding k possible values of j, this then becomes the definition of the operator norm. ■

1.3 Proving the JL Property with $\ell = 2$

Proof. We start with the expected norm. This is the same as what we did in the previous proof, but this time, we are introducing two different variable for j.

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

We expand this with 2 i variables per j to get 4 different i indices. This comes from expanding both of the squared norms with 2 i indices each.

$$= \sum_{j_1,j_2}\sum_{i_1,i_2,i_3,i_4}\mathbb{E}[\sigma_{i_1}\sigma_{i_2}\sigma_{i_3}\sigma_{i_4}\delta(h(i_1)=j_1)\delta(h(i_2)=j_1)\delta(h(i_3)=j_2)\delta(h(i_4)=j_2)]x_{i_1}x_{i_2}x_{i_3}x_{i_4}$$

We must be able to partition $\{i_1, i_2, i_3, i_4\}$ into equal pairs. This is because if 3 of them are the same and 1 is different, then the expectation is 0 because of the mismatch in the signs. Therefore, we need to make sure to partition them into pairs.

- Suppose $i_1 = i_2 = i_3 = i_4$. This means that i_1 hashes to j_1 and i_3 hashes to j_2. But, since we assume that $i_1 = i_3$, we know that i_1 hashes to j_1 and i_1 hashes to j_2, which can only happen when $j_1 = j_2$. We can then only include one variable quantifier for i, and find that the probability that $\delta(h(i) = j)$ (i hashes to the j'th bucket) is $\frac{1}{k}$

$$\sum_j \frac{1}{k}\sum_i x_i^4 = |x|_4^4$$

- Suppose $i_1 = i_2$ and $i_3 = i_4$ but $i_1 \neq i_3$. In this case, $j_1 \neq j_2$. Here we know that i_1 hashes to j_1 and i_3 hashes to j_2. We find that the probability these two events happen is $\frac{1}{k^2}$, simplifying the rest of the equation accordingly. Lastly, we subtract off the term from the previous case where $i_1 = i_2 = i_3 = i_4$.

$$\sum_{j_1,j_2,i_1,i_3}\frac{1}{k^2}x_{i_1}^2 x_{i_3}^2 = |x|_2^4 - |x|_4^4$$

- Suppose $i_1 = i_3$ and $i_2 = i_4$ but $i_1 \neq i_2$. This must mean that $j_1 = j_2$ because i_1 hashes to j_1 and i_3 hashes to j_1, which can only happen when $j_1 = j_2$. Therefore, simplifying out the expression by only keeping one j, we get the upper bound,

$$\sum_j \frac{1}{k^2}\sum_{i_1,i_2}x_{i_1}^2 x_{i_2}^2 \leq \frac{1}{k}|x|_4^4$$

  We obtain the same bound if $i_1 = i_4$ and $i_2 = i_3$ by similar logic.

Adding all the cases together, we get that $\mathbb{E}[|Sx|_2^4]$ is in the range $[|x|_2^4, |x|_2^4(1+\frac{2}{k})] = [1, 1+\frac{2}{k}]$

So we set $k = \frac{2}{\epsilon^2\delta}$ to finish the proof and obtain the JL property for $\ell = 2$,

$$\mathbb{E}_S|Sx|_2^2 - 1|^2 \leq \left(1 + \frac{2}{k}\right) - 2 + 1 = \frac{2}{k}$$

The matrix product result we wanted was:

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

The approximate matrix product gives us the result:

$$\mathbb{P}\left[\|A^TS^TSB - A^TB\|_F^2 \geq 3\varepsilon^2\|A\|_F^2\|B\|_F^2\right] \leq \delta$$

By setting $C = A^T$ and $D = B$, we finish the proof.

■

2 Affine Embeddings

We want to solve $AX = B$, where A is tall and thin with d columns, but B has a large number of columns. Since we cannot directly apply subspace embeddings, we explore the properties needed for S such that:

$$\|SAx - SB\| = (1 \pm \varepsilon)\|AX - B\|_F \tag{1}$$

for all X.

Assuming A has orthonormal columns, let X^* be the optimal solution:

$$X^* = \arg\min_X \|AX - B\| \tag{2}$$

From subspace embedding properties:

$$\|AX^* - B\| \approx \|SAX^* - SB\| \tag{3}$$

If B has m columns, we consider sketching A and B and solving the sketched version, reducing computational complexity.

Let $B* = AX * -B$, where $X*$ is the optimum, and suppose that A has orthonormal columns.

2.1 Frobenius Norm Identity Proof

Proof. We begin with the given expression and rewrite it using the definition of Frobenius and Squared Euclidean norm:

$$\|A + B\|_F^2 = \sum_i |A_i + B_i|_2^2$$

We expanding the squared norm by following a similar format to $(a + b)^2 = a^2 + b^2 + 2 * a * b$:

$$\sum_i |A_i|_2^2 + \sum_i |B_i|_2^2 + 2\langle A_i, B_i\rangle$$

Using the definition trace being the sum of diagonal elements of a matrix, we get:

$$\|A\|_F^2 + \|B\|_F^2 + 2\,\mathrm{tr}(A^T B).$$

■

2.2 Cauchy-Schwarz inequality for matrix norms

Proof.

$$\mathrm{Tr}(AB) = \sum_i \langle A_i, B_i \rangle \quad \text{(where } A_i \text{ are rows and } B_i \text{ are columns)}$$

$$\leq \sum_i |A_i|_2 |B_i|_2 \quad \text{(Cauchy-Schwarz inequality)}$$

$$\leq \left(\sum_i |A_i|_2^2 \right)^{1/2} \left(\sum_i |B_i|_2^2 \right)^{1/2} \quad \text{(Cauchy-Schwarz inequality)}$$

$$= \|A\|_F \|B\|_F \quad \text{(by definition of the Frobenius norm)}$$

■

2.3 Proving that Affine Embeddings can be solved using Sketching Matrix

Now we go to show that this problem can be solved using a sketching matrix S.

Proof. We begin with the given expression:

$$|S(AX - B)|_F^2 - |SB^*|_F^2.$$

Rewriting using the optimum term X^*, we subtract that term from one side and add it to the other side, partitioning the use of the optimum:

$$|S(AX - B)|_F^2 - |SB^*|_F^2 = |SA(X - X^*) + S(AX^* - B)|_F^2 - |SB^*|_F^2.$$

Applying the identity $|C + D|_F^2 = |C|_F^2 + |D|_F^2 + 2\mathrm{Tr}(C^T D)$ from section 2.1, we get:

$$|SA(X - X^*)|_F^2 + 2\mathrm{tr}[(X - X^*)^T A^T S^T S B^*].$$

Using the inequality $\mathrm{tr}(CD) \leq |C|_F |D|_F$ from section 2.2, we get:

$$|SA(X - X^*)|_F^2 \pm 2|X - X^*|_F |A^T S^T S B^*|_F.$$

Under the assumption of an approximate matrix product:

$$|SA(X - X^*)|_F^2 \pm 2\epsilon |X - X^*|_F |B^*|_F.$$

Finally, using the subspace embedding property for A:

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

■

We have the following:

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

The normal equations indicate:

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

If we were to draw this out, for each column X_i in X, the vector AX_i represents a point in the column space of A. The columns B_i are any points that lie in the same geometric space. We know that AX_i^* is the closest point to B_i within the column space since the shortest distance is the line that forms a right angle on the column space and touches B_i. The distance between these two points is $B_i^* = AX_i^* - B_i$, which contributes to the term $\|B^*\|_F$. Similarly, we get the term $\|A(X - X^*)\|_F$. We start with the normal equations.

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

which simplifies to the following using the Pythagorean theorem:

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

Then, we get,

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

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

This leads to:

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

indicating that the error due to the subspace embedding is approximately $2\epsilon\|AX - B\|_F^2$.

Next, we use the fact that:

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

with constant probability. In other words, S preserves the norm of a fixed matrix with constant probability.

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

This simplifies further to:

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

Thus, we conclude that S is a $(1 + 3\epsilon)$-affine embedding for X.