**CS 15-859: Algorithms for Big Data** Fall 2019

Lecture 3.1 — September 19

*Prof. David Woodruff* *Scribe: Dongyu Li*

# 1 CountSketch Satisfies JL Property

Recall from the previous lecture we showed that CountSketch is a subspace embedding via approximate matrix product result, and all it remains to show is the assumption we made that the CountSketch matrix $S$ satisfies the JL property for some $\ell \geq 2$.

**Definition.** (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 $|x|_2 = 1$,

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

When $\ell = 1$, the JL Property checks whether the matrix $S$ preserves norms. Let's prove it for $\ell = 2$.

**Lemma 1.** *CountSketch matrix satisfies the JL-property for $\ell = 2$.*

*Proof.* First of all, we can view every CountSketch matrix $S$ as being described by two hash functions

- $h : [n] \to [k]$ is a 2-wise independent hash function. $h(i)$ is the row index in the $i$th column of matrix $S$ that has nonzero entry.
- $\sigma : [n] \to \{-1, 1\}$ is a 4-wise independent hash function. $\sigma(i)$ is the sign of the nonzero entry in the $i$th column of matrix $S$.

We define the notation $\delta(E) = 1$ if event $E$ holds, and $\delta(E) = 0$ otherwise. Let's first consider the expression $\mathbb{E}[|Sx|_2^2]$.

$$\begin{aligned}
\mathbb{E}[|Sx|_2^2] &= \mathbb{E}[\sum_{j=1}^{k} (S_{j*} \cdot x)^2] \\
&= \mathbb{E}[\sum_{j=1}^{k} (\sum_{i=1}^{n} \delta(h(i) = j)\sigma_i x_i)^2] \\
&= \mathbb{E}[\sum_{j=1}^{k} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \delta(h(i_i) = j)\delta(h(i_2) = j)\sigma_{i_1}\sigma_{i_2}x_{i_1}x_{i_2})] \\
&= \sum_{j=1}^{k} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \mathbb{E}[\delta(h(i_i) = j)\delta(h(i_2) = j)\sigma_{i_1}\sigma_{i_2}]x_{i_1}x_{i_2} \qquad \text{(linearity of expectation)} \\
&= \sum_{j=1}^{k} \sum_{i=1}^{n} \mathbb{E}[\delta(h(i) = j)^2]x_i^2 \qquad (1)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{k} \sum_{i=1}^{n} \mathbf{Pr}[\delta(h(i) = j)] x_i^2 \\
&= \sum_{j=1}^{k} \frac{1}{k} \sum_{i=1}^{n} x_i^2 && \text{(definition of hash function } h\text{)} \\
&= \sum_{i=1}^{n} x_i^2 \\
&= |x|_2^2
\end{aligned}
$$

Note that $\sigma_i = \sigma(i)$ in the above derivation for clarity. To arrvie at (1) from the previous step, we made the observation that when $i_1 \neq i_2$, $\mathbb{E}[\sigma_{i_1}] = \mathbb{E}[\sigma_{i_2}] = 0$ by definition of hash function $\sigma$, and else when $i_1 = i_2$, $\sigma_i^2 = 1$, as the range of $\sigma$ is $\{-1, 1\}$. In the former case, all terms become 0, and in the latter case, we can write it as shown in (1).

Since $\ell = 2$, we would also need to consider $\mathbb{E}[|Sx|_2^4]$.

$$
\begin{aligned}
\mathbb{E}[|Sx|_2^4] &= \mathbb{E}[(|Sx|_2^2)^2] \\
&= \mathbb{E}[\sum_{j=1}^{k} \sum_{j'=1}^{k} (\sum_{i=1}^{n} \delta(h(i) = j)\sigma_i x_i)^2 (\sum_{i'=1}^{n} \delta(h(i') = j')\sigma_{i'} x_{i'})^2] \\
&= \sum_{j_1, j_2, i_1, i_2, i_3, i_4} \mathbb{E}[\sigma_{i_1}\sigma_{i_2}\sigma_{i_3}\sigma_{i_4}\delta(h(i_1) = j_1)\delta(h(i_2) = j_1)\delta(h(i_3) = j_2)\delta(h(i_4) = j_2)] x_{i_1} x_{i_2} x_{i_3} x_{i_4}
\end{aligned}
$$

Observe that we must be able to partition $\{i_1, i_2, i_3, i_4\}$ into equal pairs, because if there is one that's distinct from each of the rest, then its expected value of $\sigma$ is 0 and makes the entire term 0. Hence, we have these cases to consider:

- Case $i_1 = i_2 = i_3 = i_4$: implies $h(i_i) = h(i_2) = h(i_3) = h(i_4)$ and only when $j_1 = j_2$, all events simultaneously happen, so these terms become

$$
\begin{aligned}
&\sum_{j=1}^{k} \sum_{i=1}^{n} \mathbb{E}[\delta(h(i) = j)^4] x_i^4 \\
&= \sum_{j=1}^{k} \sum_{i=1}^{n} \mathbf{Pr}[h(i) = j] x_i^4 \\
&= \sum_{j=1}^{k} \frac{1}{k} \sum_{i=1}^{n} x_i^4 \\
&= \sum_{i=1}^{n} x_i^4 \\
&= |x|_4^4
\end{aligned}
$$

- Case $i_1 = i_2$ and $i_3 = i_4$, but $i_1 \neq i_3$: these terms in the sum becomes:

$$\begin{aligned}
&\sum_{j_1,j_2=1}^{k} \sum_{i_1,i_3=1,i_1\neq i_3}^{n} \mathbb{E}[\delta(h(i_1) = j_1)^2 \delta(h(i_3) = j_2)^2] x_{i_1}^2 x_{i_3}^2 \\
&= \sum_{j_1,j_2=1}^{k} \sum_{i_1,i_3=1,i_1\neq i_3}^{n} \mathbf{Pr}[h(i_1) = j_1 \wedge h(i_3) = j_2] x_{i_1}^2 x_{i_3}^2 \\
&= \sum_{j_1,j_2=1}^{k} \frac{1}{k^2} \sum_{i_1,i_3=1,i_1\neq i_3}^{n} x_{i_1}^2 x_{i_3}^2 && \text{(2-wise independence of h)} \\
&= \sum_{i_1,i_3=1,i_1\neq i_3}^{n} x_{i_1}^2 x_{i_3}^2 \\
&= \sum_{i_1,i_3=1}^{n} x_{i_1}^2 x_{i_3}^2 - \sum_{i=1}^{n} x_i^4 \\
&= |x|_2^4 - |x|_4^4
\end{aligned}$$

- Case $i_1 = i_3$ and $i_2 = i_4$, but $i_1 \neq i_2$: for the same reason as in the first case, $j_1 = j_2$, and these terms become

$$\begin{aligned}
&\sum_{j=1}^{k} \sum_{i_1,i_2=1,i_1\neq i_2}^{n} \mathbb{E}[\delta(h(i_1) = j)(h(i_2) = j) x_{i_1}^2 x_{i_2}^2 \\
&= \sum_{j=1}^{k} \sum_{i_1,i_2=1,i_1\neq i_2}^{n} \mathbf{Pr}[h(i_1) = j \wedge h(i_2) = j] x_{i_1}^2 x_{i_2}^2 \\
&= \sum_{j=1}^{k} \frac{1}{k^2} \sum_{i_1,i_2=1,i_1\neq i_2}^{n} x_{i_1}^2 x_{i_2}^2 && \text{(2-wise independence of h)} \\
&= \frac{1}{k} \sum_{i_1,i_2=1,i_1\neq i_2}^{n} x_{i_1}^2 x_{i_2}^2 \\
&\leq \frac{1}{k} \sum_{i_1,i_2=1}^{n} x_{i_1}^2 x_{i_2}^2 \\
&= \frac{1}{k} |x|_2^4
\end{aligned}$$

  Note that the case where $i_1 = i_4$ and $i_2 = i_3$, but $i_1 \neq i_2$ is equivalent to this case and hence constribute another $\leq \frac{1}{k}|x|_2^4$ towards the total sum.

Considering all four cases, we find the upperbound of $\mathbb{E}[|Sx|_2^4]$:

$$\mathbb{E}[|Sx|_2^4] \leq |x|_2^4 + \frac{2}{k}|x|_2^4 = (1 + \frac{2}{k})|x|_2^4 = 1 + \frac{2}{k}$$

Then, we proceed to upperbound $\mathbb{E}[(|Sx|_2^2 - 1)^2]$ as follows:

$$\begin{aligned}
\mathbb{E}[(|Sx|_2^2 - 1)^2] &= \mathbb{E}[|Sx|_2^4 - 2|Sx|_2^2 + 1] \\
&= \mathbb{E}[|Sx|_2^4] - 2\,\mathbb{E}[|Sx|_2^2] + 1 && \text{(linearity of expectation)} \\
&\leq 1 + \frac{2}{k} - 2 + 1 && (\mathbb{E}[|Sx|_2^2] = |x|_2^2 = 1 \text{ is an unbiased estimator}) \\
&= \frac{2}{k}
\end{aligned}$$

In order for the JL property to hold, we set $k = \frac{2}{\epsilon^2\delta}$. Thus, we've proven the JL property for $\ell = 2$ holds for CountSketch matrix $S$ and we bound the number k of rows of $S$ to be at least $\frac{2}{\epsilon^2\delta}$ ■

Recall that the setup for least squares regression is $\min_{x\in\mathbb{R}^d} |Ax - b|_2^2$, where $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$. The total runtime of CountSketch for least squares regression is $nnz(A) + poly(\frac{d}{\epsilon})$, as computing $SA$ and $Sb$ takes $nnz(A) + nnz(b)$ and sovling the projected smaller problem takes $O(kd^2)$.

## 2 Affine Embeddings

Consider the generalized regression problem

$$\min_X |AX - B|_F^2$$

where A is a tall and thin matrix with $d$ columns, and B is also a matrix with a large number of columns.

We can try minimizing each column of B separately as least squares regression problem and add the terms, but it's slow due to large number of columns of B.

Note that we also can't direclty apply subspace embeddings, because the dimension of subspace increases by the large number of columns of B.

let's try to show $|SAX - SB|_F = (1 \pm \epsilon)|AX - B|_F$ for all $X$ and see what properties we would require of $S$. Just as before, we can assume $A$ has orthonormal columns. Let $B^* = AX^* - B$, where $X^*$ is the optimum.

Let's first analyze the expression $|S(AX - B)|_F^2 - |SB^*|_F^2$ and see what we can get

$$\begin{aligned}
&|S(AX - B)|_F^2 - |SB^*|_F^2 \\
&= |SA(X - X^*) + S(AX^* - B)|_F^2 - |SB^*|_F^2 \\
&= |SA(X - X^*)|_F^2 + |S(AX^* - B)|_F^2 + 2tr[(X - X^*)^T A^T S^T S(AX^* - B)] - |SB^*|_F^2 && (1) \\
&= |SA(X - X^*)|_F^2 + 2tr[(X - X^*)^T A^T S^T SB^*] && (2) \\
&\in |SA(X - X^*)|_F^2 \pm 2|X - X^*|_F |A^T S^T SB^*|_F && (3) \\
&\in |SA(X - X^*)|_F^2 \pm 2\epsilon |X - X^*|_F |B^*|_F && (4) \\
&\in |A(X - X^*)|_F^2 \pm \epsilon(|A(X - X^*)|_F^2 + 2|X - X^*|_F |B^*|_F) && (5)
\end{aligned}$$

Facts used for steps of the above derivation:

- (1) $|C+D|_F^2 = |C|_F^2 + |D|_F^2 + 2tr(C^T D)$, which we will prove in the next lecture
- (2) $|S(AX^* - B)|_F^2 = |SB^*|_F^2$, because we defined $B^* = AX^* - B$
- (3): $tr(CD) \leq |C|_F |D|_F$, which we will prove in the next lecture
- (4) Note that $|A^T S^T SB^* - A^T B^*|_F^2 = |A^T S^T SB^* - 0|_F^2 = |A^T S^T SB^*|_F^2$, because $B^*$ is orthogonal to $A^T$. If we have approximate matrix product, then $|A^T S^T SB^*|_F^2 \leq O(\frac{1}{\#\text{ of rows of S}})|A^T|_F^2 \cdot |B|_F^2$. Since $A$ is an $n \times d$ orthonormal matrix, $|A|_F^2 = d$. Therefore, we can bound $|A^T S^T SB^*|_F^2 \leq \epsilon^2 |B|_F^2$, if the number of rows in $S \geq \frac{d}{\epsilon^2}$. Hence, under this assumption $|A^T S^T SB^*|_F \leq \epsilon |B^*|_F$
- (5) if $S$ is a subspace embedding for the column span of $A$, for which $A(X - X^*)$ is in, the Frobenius norm is preserved for a multiplicative error up to $\epsilon$.

Let's now look at the normal equation that's analogous to the one for least square regression we looked at during Lecture 1.

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

Note that in this case columns of $B^*$ are orthogonal to the column span of $A$.

Now, let's analyze the expression that subtracts the non-sketched difference between $|AX - B|_F^2$ and $|B^*|_F^2$ from the sketeched difference

$$\begin{aligned}
&|S(AX - B)|_F^2 - |SB^*|_F^2 - (|AX - B|_F^2 - |B^*|_F^2) \\
&\in |A(X - X^*)|_F^2 \pm \epsilon(|A(X - X^*|_F^2 + 2|X - X^*|_F |B^*|_F) - |A(X - X^*)|_F^2 \\
&\in \pm\epsilon(|A(X - X^*)|_F^2 + 2|X - X^*|_F |B^*|_F) \\
&\in \pm\epsilon(|A(X - X^*)|_F^2 + 2|A(X - X^*)|_F |B^*|_F + |B^*|_F^2) \qquad (1) \\
&\in \pm\epsilon(|A(X - X^*)|_F + |B^*|_F)^2 \\
&\in \pm 2\epsilon(|A(X - X^*)|_F^2 + |B|_F^2) \qquad (2) \\
&\in \pm 2\epsilon |AX - B|_F^2
\end{aligned}$$

Facts used for steps of the above derivation:

- (1) $A$ is orthonormal, preserving Frobenius norm
- (2) $(a+b)^2 \leq 2a^2 + 2b^2$, because $ab \leq \frac{a^2+b^2}{2}$

We will also assume

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

holds for our sketching matrix $S$ with constant probability. Note that $B^* = AX^* - B$, which does not depend on $X$. Hence, we just need this to hold for a fixed $B^*$, which is not much to ask.

Then, we get

$$\begin{aligned}
|S(AX-B)|_F^2 &= |AX-B|_F^2 + (|SB^*|_F^2 - |B^*|_F^2) \pm 2\epsilon |AX-B|_F^2 \\
&= (1 \pm 2\epsilon)|AX-B|_F^2 + \epsilon |B^*|_F^2 \\
&= (1 \pm 2\epsilon)|AX-B|_F^2 + \epsilon |AX^*-B|_F^2 \\
&= (1 \pm 3\epsilon)|AX-B|_F^2
\end{aligned}$$

The last step is because at optimum $X^*$, $|AX-B|_F^2$ is smaller than any other possible $X$, by definition of being optimal.

Therefore, we have shown that $S$ is an affine embedding, if it satisfies these properties:

- $S$ is a subspace embedding for columns of $A$.
- $S$ has the approximate matrix product result.
- $S$ preserves Frobenious norm up to $\epsilon$ error for a fixed matrix $B^*$ with constant probability.