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

Lecture 2 — 9/13/2019

Prof. David Woodruff | *Scribe: Shicong Cen*

Note: details differing from the original slides are marked red.

We have already shown that by using Subsampled Randomized Hadamard Transform, one can improve the time for regression to $O(nd \log n) + poly((d \log n)/\varepsilon)$. When A is sparse, we can do even better by exploiting the sparsity of A.

Definition (CountSketch matrix)**.** A $k \times n$ matrix S is a CountSketch matrix when it is constructed in the following way:

1. For every column, one entry is uniformly selected with random, and is then assigned a random variable that takes value ± 1 with equal probability.
2. All the other entries are assigned 0.

When A is sparse, we are able to compute $S \cdot A$ in $O(\mathrm{nnz}(A))$ with appropriate data structure. It remains to show that it is indeed a subspace embedding.

Theorem 1. *When $k = 18d^2/(\delta\varepsilon^2)$, the CountSketch matrix is a subspace embedding. To be specific, the following statement holds with probability exceeding $1 - \delta$:*

$$(1-\varepsilon)\|Ax\| \le \|SAx\| \le (1+\varepsilon)\|Ax\|, \forall x \in \mathbb{R}^d.$$

where $\delta \in (0, 1/2)$, $\varepsilon \in (0, 1)$.

Proof. We outline the proof idea as follows:

1. We can still assume that columns of A are orthonormal, and it is sufficient to prove the theorem for all unit vectors x.
2. We show that with probability exceeding $1 - \delta$, we have

$$\left\|A^\top S^\top SA - I\right\|_2 \le \varepsilon$$

However, we are not going to use the matrix Chernoff bounds as we did when proving SRHT. Instead, we are going to use the following matrix product result:

$$\mathbf{Pr}\Big[\left\|CS^\top SD - CD\right\|_F^2 \le 18/\left(\delta(\#\text{ rows of } S)\right) \cdot \|C\|_F^2 \|D\|_F^2\Big] \ge 1 - \delta$$

Plugging $C = A^\top$, $D = A$ into the result along with the fact $\|\cdot\|_2 \le \|\cdot\|_F$ finishes the proof.

∎

Remark 1. While the slides keep the notation ε throughout the proof, we use $\hat{\varepsilon}$ for JL property to prevent confusing.

We will then make the proof complete by focusing on the matrix product result. To proceed, we will introduce JL property:

Definition (JL Property)**.** A distribution on matrices $S \in \mathbb{R}^{k\times n}$ has the $(\hat{\varepsilon}, \delta, \ell)$ −JL moment property if $\forall x \in \mathbb{R}^n$ with $\|x\| = 1$,

$$\mathbb{E}\Big[|\,\|Sx\|_2^2 - 1|^\ell\Big] \leq \hat{\varepsilon}^\ell \cdot \delta.$$

We then claim that with appropriate JL property, the matrix product result holds. We introduce the following notion:

Definition (p-norm of random variables)**.** For a random variable X and $p \geq 1$, the p-norm is defined as

$$\|X\|_p = \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}$$

The only non-trivial detail about the notion is the triangle inequality, a.k.a. Minkowski inequality for p-norm. The details are included in the Appendix A. We show that to prove matrix product result, it suffices to verify the JL property of S.

Theorem 2. *For $\hat{\varepsilon}, \delta \in (0, 1/2)$, let the distribution of S satisfies the $(\hat{\varepsilon}, \delta, \ell)$-JL moment property for some $\ell \geq 2$. Then we have*

$$\mathbf{Pr}\Big[\,\Big\|A^\top S^\top SB - A^\top B\Big\|_F \geq 3\hat{\varepsilon}\,\|A\|_F\,\|B\|_F\,\Big] \leq \delta$$

for all matrices A, B over the randomness of S.

Proof. First, we show that JL-property implies that S is almost inner product preserving: for unit vectors x, y,

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

Hence for any vector x, y, we have $\|\langle Sx, Sy\rangle - \langle x, y\rangle\|_\ell \leq 3\hat{\varepsilon}\cdot\delta^{1/\ell}\,\|x\|_2\,\|y\|_2$. Let columns of A be $A_1, \cdots, A_d$ and columns of B be $B_1, \cdots, B_e$, the (i, j)-th entry of $A^\top S^\top SB - A^\top B$ can be written as

$$\langle SA_i, SB_j\rangle - \langle A_i, B_j\rangle$$

Hence

$$\begin{aligned}
\left\| \left\| A^\top S^\top SB - A^\top B \right\|_F^2 \right\|_{\ell/2} &= \left\| \sum_i \sum_j \left(\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle \right)^2 \right\|_{\ell/2} \\
&\leq \sum_i \sum_j \left\| \left(\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle \right)^2 \right\|_{\ell/2} \\
&= \sum_i \sum_j \left\| \left(\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle \right) \right\|_\ell^2 \\
&\leq \left(3\hat{\varepsilon} \cdot \delta^{1/\ell} \right)^2 \sum_i \sum_j \|A_i\|_2^2 \|B_j\|_2^2 \\
&= \left(3\hat{\varepsilon} \cdot \delta^{1/\ell} \right)^2 \|A\|_F^2 \|B\|_F^2 .
\end{aligned}$$

Note that the first inequality relies on Minkowski's inequality, which requires $\ell \geq 2$. With $\mathbb{E}\left[\left\| A^\top S^\top SB - A^\top B \right\|_F^\ell \right] = \left\| \left\| A^\top S^\top SB - A^\top B \right\|_F^2 \right\|_{\ell/2}^{\ell/2}$, we may complete the proof with Markov's inequality:

$$\begin{aligned}
\mathbf{Pr}\left[\left\| A^\top S^\top SB - A^\top B \right\|_F > 3\hat{\varepsilon} \|A\|_F \|B\|_F \right] \leq & \left(\frac{1}{3\hat{\varepsilon} \|A\|_F \|B\|_F} \right)^\ell \mathbb{E}\left[\left\| A^\top S^\top SB - A^\top B \right\|_F^\ell \right] \\
\leq & \delta.
\end{aligned}$$

■

We then proceed to the JL property of CountSketch matrix.

Theorem 3 (JL Property of CountSketch)**.** *When $k \geq 2/(\hat{\varepsilon}^2 \delta)$, the distribution of CountSketch matrix $S \in \mathbb{R}^{k \times n}$ satisfies the following $(\hat{\varepsilon}, \delta, \ell)$-JL moment property with $\ell = 2$ for all $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$:*

$$\mathbb{E}\left[\left| \|Sx\|_2^2 - 1 \right|^\ell \right] \leq \hat{\varepsilon}^\ell \cdot \delta$$

Proof. We first show that $\|Sx\|_2^2$ is an unbiased estimator of $\|x\|_2^2$. We will use the so-called hash function to characterize matrix S throughout this proof:

- $h : [n] \to [k]$ characterizes the position of non-zero entries of S. That is, the $f(i)$-th entry of the i-th column of S is non-zero.
- $\sigma : [n] \to \{-1, 1\}$ characterizes the values of the non-zero entries: $[S]_{f(i),i} = \sigma(i)$.
- h is 2-wise independent: $\forall i \neq j \in [n]$ and $\forall z_1, z_2 \in [k]$, $\mathbf{Pr}[h(i) = z_1, h(j) = z_2] = 1/k^2$. σ is 4-wise independent, where the definition is similar.

Let $\delta(E)$ be the indicator function of event E. We have

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

The rest of the proof will be covered by the next lecture. ■

By combining theorem 3 and 2 and setting $k = 2/(\hat{\varepsilon}^2\delta)$, we have

$$\begin{aligned}
\delta \geq & \mathbf{Pr}\Big[\left\| A^\top S^\top SB - A^\top B \right\|_F \geq 3\hat{\varepsilon} \, \|A\|_F \, \|B\|_F \Big] \\
= & \mathbf{Pr}\Big[\left\| A^\top S^\top SB - A^\top B \right\|_F^2 \geq 9\hat{\varepsilon}^2 \, \|A\|_F^2 \, \|B\|_F^2 \Big] \\
= & \mathbf{Pr}\Big[\left\| A^\top S^\top SB - A^\top B \right\|_F^2 = 18/(k\delta) \, \|A\|_F^2 \, \|B\|_F^2 \Big]
\end{aligned}$$

with $\hat{\varepsilon}, \delta \in (0, 1/2)$, which is exactly the result we needed.

Remark 2. When k is set to $18d^2/(\delta\varepsilon^2)$, we have $\hat{\varepsilon} = \varepsilon/(3d)$, so $\hat{\varepsilon} \in (0, 1/2)$ holds.

A Minkowski's Inequality for Random Variables

First, we show that when $\|X\|_p$ and $\|Y\|_p$ are finite, then so is $\|X+Y\|_p$. Note that $f(x) = x^p$ is convex for $p \geq 1$, so we have

$$\|(x+y)/2\|^p \leq (\|x\|^p + \|y\|^p)/2$$

or $\|(x+y)\|^p \leq 2^{p-1}(\|x\|^p + \|y\|^p)$. By taking the expectation, we have

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

so $\|X+Y\|_p$ is finite.

We are now ready to prove the Minkowski's inequality: let μ be the probability measure of (x, y) and we have

$$\begin{aligned}
\|X+Y\|_p^p &= \int_{x,y} \|x+y\|^p \,\mathrm{d}\mu \\
&\leq \int_{x,y} (\|x\| + \|y\|)\, \|x+y\|^{p-1} \,\mathrm{d}\mu \\
&= \int_{x,y} \|x\|\, \|x+y\|^{p-1} \,\mathrm{d}\mu + \int_{x,y} \|y\|\, \|x+y\|^{p-1} \,\mathrm{d}\mu \\
&\leq \left(\left(\int_x \|x\|^p \,\mathrm{d}\mu\right)^{1/p} + \left(\int_x \|y\|^p \,\mathrm{d}\mu\right)^{1/p}\right) \left(\int_{x,y} \|x+y\|^{(p-1)\cdot p/(p-1)} \,\mathrm{d}\mu\right)^{(p-1)/p} \\
&= \left(\|X\|_p + \|Y\|_p\right) \|X+Y\|_p^{p-1}.
\end{aligned}$$

where the last inequality is due to Hölder's inequality. Hence $\|X+Y\|_p \leq \|X\|_p + \|Y\|_p$.