Lecture 2-2 — September 14, 2017

Prof. David Woodruff *Scribe: Taisuke Yasuda*

1 Sketching with the CountSketch matrix

Note that sketching with Subsampled Randomized Hadamard Transform matrices is not optimal for sparse matrices, since its runtime is $O\left(\left(nd\log\left(\frac{d\log n}{\varepsilon}\right)\right)\right)$ which depends on nd, the size of the whole matrix. We introduce a new sketch matrix, the CountSketch matrix but we want dependence only on $\mathrm{nnz}(\mathbf{A})$, the number of nonzero entries of $\mathbf{A}$. We define a $k\times n$ matrix $\mathbf{S}$, for $k = O(d^2/\varepsilon^2)$. Our matrix $\mathbf{S}$ is very sparse: for every column, we choose a single randomly chosen nonzero entry, which takes value ± 1 with equal probability. Note then that we may compute $\mathbf{S}\cdot\mathbf{A}$ in time proportional to the number of nonzero entries of $\mathbf{A}$.

Remark 1. There is a tradeoff between the number of nonzero entries you choose in $\mathbf{S}$ and the dependence of k on d, e.g. if you increase the number of nonzero entries of $\mathbf{S}$ then you can get $k = O(d^{1.1}/\varepsilon)$ or something like this.

Theorem 1. *The* CountSketch *matrix is a subspace embedding.*

Proof. We again assume that $\mathbf{A}$ is orthonormal. We first show that $\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2 = 1\pm\varepsilon$ for all unit vectors $\mathbf{x}$. It suffices to show that

$$\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{A} - \mathbf{I}_d\right\|_2 \le \left\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{A} - \mathbf{I}_d\right\|_F \le \varepsilon$$

since the Frobenius norm upper bounds the operator norm. To prove this, we will show that the CountSketch matrix satisfies a property known as the JL-property and that this property implies an inequality that is equivalent to the one above.

Definition (Johnson-Lindenstrauss Property)**.** A distribution on matrices $\mathbf{S}\in\mathbb{R}^{k\times n}$ satisfies the $(\varepsilon,\delta,\ell)$-*JL moment property* if for all $\mathbf{x}\in\mathbb{R}^n$ with $\|\mathbf{x}\|_2 = 1$,

$$\mathop{\mathbb{E}}_{\mathbf{S}}\left|\|\mathbf{S}\mathbf{x}\|_2^2 - 1\right|^\ell \le \varepsilon^\ell\cdot\delta$$

where the expectation is taken over the distribution over the matrices.

Lemma 1 (Kane, Nelson)**.** *Fix* $\mathbf{C},\mathbf{D}$ *matrices with the right dimensions,* $\mathbf{S}\in\mathbb{R}^{k\times d}$ *a distribution over matrices satisfying the JL property, and fix* δ*. Then*

$$\mathbb{P}\left[\left\|\mathbf{C}\mathbf{S}^T\mathbf{S}\mathbf{D} - \mathbf{C}\mathbf{D}\right\|_F^2 \le \frac{6}{\delta\cdot k}\|\mathbf{C}\|_F^2\|\mathbf{D}\|_F^2\right] \ge 1-\delta.$$

Proof. We first prove Minkowski's inequality for random scalar variables, which we will need in the proof.

Definition (p-norm of Random Variables)**.** For a random scalar variable X, we define its *p-norm* to be

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

Remark 2. We will also often consider the p-norm of the Frobenius norm of a random matrix $\mathbf{T}$, which is

$$\left|\left\|\mathbf{T}\right\|_F\right|_p = \left(\mathbb{E}\left[\left\|\mathbf{T}\right\|_F^p\right]\right)^{1/p}.$$

The word "norm" in the p-norm is justified for $p \geq 1$, and the triangle inequality in this case is known as Minkowski's inequality.

Lemma 2 (Minkowski's Inequality)**.** *Let $p \geq 1$. Then*

$$\left|X+Y\right|_p \leq \left|X\right|_p + \left|Y\right|_p.$$

Proof. We first show that if $\left|X\right|_p, \left|Y\right|_p$ are finite, then so is $\left|X+Y\right|_p$. Note that $f(x) = x^p$ is convex for $p \geq 1$ and $x \geq 0$, so for any fixed x, y, we have that

$$\left|\frac{1}{2}x + \frac{1}{2}y\right|^p \leq \left(\frac{1}{2}\left|x\right| + \frac{1}{2}\left|y\right|\right)^p \leq \frac{1}{2}\left|x\right|^p + \frac{1}{2}\left|y\right|^p$$

so

$$\left|x+y\right|^p \leq 2^{p-1}\left(\left|x\right|^p + \left|y\right|^p\right)$$

and thus integrating the above inequality over the probability measure (taking the expectation) gives

$$\left|X+Y\right|_p^p \leq 2^{p-1}\left(\left|X\right|_p^p + \left|Y\right|_p^p\right)$$

and thus $\left|X+Y\right|_p^p$ is finite. Now note that

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

Applying Hölder's inequality with $\frac{1}{p} + \frac{p-1}{p} = 1$ applied to each term, we bound the above by

$$\left(\left(\int \left|x\right|^p \ d\mu\right)^{1/p} + \left(\int \left|y\right|^p \ d\mu\right)^{1/p}\right)\left(\int \left|x+y\right|^{(p-1)(p/(p-1))} \ d\mu\right)^{(p-1)/p} = \left(\left|X\right|_p + \left|Y\right|_p\right)\left|X+Y\right|_p^{p-1}.$$

If $\left|X+Y\right|_p = 0$, then the inequality is trivially true and otherwise, since we showed that $\left|X+Y\right|_p$ is finite, we may divide through by $\left|X+Y\right|_p^{p-1}$ to conclude. ■

Theorem 2. *Let $\mathbf{S} \in \mathbb{R}^{k \times n}$ be a distribution of matrices that satisfies the $(\varepsilon, \delta, \ell)$-JL property with $\varepsilon, \delta \in (0, 1/2)$ and $\ell \geq 2$. Then for matrices $\mathbf{A}, \mathbf{B}$ with n rows,*

$$\mathbb{P}_{\mathbf{S}}\left[\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{B}\right\|_F \geq 3\varepsilon\left\|\mathbf{A}\right\|_F\left\|\mathbf{B}\right\|_F\right] \leq \delta.$$

Proof. Let $\mathbf{x}, \mathbf{y}$ be unit vectors. Then

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

Now by linearity, we conclude that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\frac{|\langle \mathbf{Sx}, \mathbf{Sy}\rangle - \langle \mathbf{x}, \mathbf{y}\rangle|_\ell}{\|\mathbf{x}\|_2\,\|\mathbf{y}\|_2} \le 3\varepsilon\delta^{1/\ell}.$$

Now let $\mathbf{A}$ have columns $\mathbf{A}_1, \ldots, \mathbf{A}_d$ and let B have columns $\mathbf{B}_1, \ldots, \mathbf{B}_e$. Then define $\mathbf{X}$ entrywise by defining for each $i \in [d], j \in [e]$,

$$\mathbf{X}_{i,j} := \frac{\langle \mathbf{SA}_i, \mathbf{SB}_j\rangle - \langle \mathbf{A}_i, \mathbf{B}_j\rangle}{\|\mathbf{A}_i\|_2\,\|\mathbf{B}_j\|_2}.$$

Then,

$$\begin{aligned}
\left|\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{SB} - \mathbf{A}^T\mathbf{B}\right\|_F^2\right|_{\ell/2} &= \left|\sum_{i=1}^d\sum_{j=1}^e \|\mathbf{A}\|_2^2\,\|\mathbf{B}\|_2^2\,\mathbf{X}_{i,j}^2\right|_{\ell/2} \le \sum_{i=1}^d\sum_{j=1}^e \|\mathbf{A}\|_2^2\,\|\mathbf{B}\|_2^2\left|\mathbf{X}_{i,j}^2\right|_{\ell/2} \\
&= \sum_{i=1}^d\sum_{j=1}^e \|\mathbf{A}\|_2^2\,\|\mathbf{B}\|_2^2\,|\mathbf{X}_{i,j}|_\ell^2 \le \left(3\varepsilon\delta^{1/\ell}\right)^2\sum_{i=1}^d\sum_{j=1}^e \|\mathbf{A}\|_2^2\,\|\mathbf{B}\|_2^2 \\
&= \left(3\varepsilon\delta^{1/\ell}\right)^2\|\mathbf{A}\|_F^2\,\|\mathbf{B}\|_F^2.
\end{aligned}$$

Finally, since we have

$$\mathbb{E}\left[\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{SB} - \mathbf{A}^T\mathbf{B}\right\|_F^\ell\right] = \left|\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{SB} - \mathbf{A}^T\mathbf{B}\right\|_F^2\right|_{\ell/2}^{\ell/2},$$

Markov's inequality allows us to conclude by

$$\begin{aligned}
\mathbb{P}\left[\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{SB} - \mathbf{A}^T\mathbf{B}\right\|_F > 3\varepsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F\right] &= \mathbb{P}\left[\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{SB} - \mathbf{A}^T\mathbf{B}\right\|_F^\ell > (3\varepsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F)^\ell\right] \\
&\le \frac{1}{(3\varepsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F)^\ell}\mathbb{E}\left[\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{SB} - \mathbf{A}^T\mathbf{B}\right\|_F^\ell\right] \\
&\le \frac{1}{(3\varepsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F)^\ell}\left(\left(3\varepsilon\delta^{1/\ell}\right)^2\|\mathbf{A}\|_F^2\,\|\mathbf{B}\|_F^2\right)^{\ell/2} \\
&\le \frac{\delta\,(3\varepsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F)^\ell}{(3\varepsilon\|\mathbf{A}\|_F\|\mathbf{B}\|_F)^\ell} = \delta.
\end{aligned}$$

■

■

With the lemma by Kane and Nelson in hand, it remains to show that CountSketch satisfies the JL property. Note that $\|\mathbf{A}\|_F = \left\|\mathbf{A}^T\right\|_F = d$ since $\mathbf{A}$ was assumed to be orthonormal. Then, after proving the CountSketch is JL, we may let $\mathbf{SA}$ be a $k \times d$ matrix with $k = 6d^2/(\delta\varepsilon^2)$, $\mathbf{C} := \mathbf{A}^T$, and $\mathbf{D} := \mathbf{A}$ to conclude that

$$\mathbb{P}\left[\left\|\mathbf{A}^T\mathbf{S}^T\mathbf{S}\mathbf{A} - \mathbf{I}_d\right\|_F \leq \varepsilon\right] \geq 1 - \delta$$

as desired.

Theorem 3. *The* CountSketch *matrix satisfies the* $(\varepsilon, \delta, 2)$*-JL property.*

Proof. Next lecture. ■

■