CS 15-859: Algorithms for Big Data Fall 2020

Lecture 2.2 — 9/17/2020

Prof. David Woodruff *Scribe: Tommy Jiang*

1 Recap: Subsampled Randomized Hadamard Transform

Using a Subsampled Randomized Hadamard Transform (SRHT) allows us to reduce the time complexity of approximating least squares from $O(nd^2)$ to $O(nd \log n)$. We previously proved the Flattening Lemma and a consequence of it:

Lemma 1. *(Flattening Lemma) For any fixed unit vector y and some constant $C > 0$,*

$$\mathbf{Pr}[\|HDy\|_\infty \geq C\frac{\sqrt{\log(\frac{nd}{\delta})}}{\sqrt{n}}] \leq \frac{\delta}{2d} \tag{1}$$

Corollary 1. *For all $j \in [n]$,*

$$\|e_j HDA\|_2 \leq C\frac{\sqrt{d\log(nd/\delta)}}{\sqrt{n}} \tag{2}$$

Our goal is to prove that the SRHT is a subspace embedding; i.e., $\|SAx\|_2 = \|PHDAx\|_2^2 = 1 \pm \epsilon$ for all unit vectors x. We will proceed by conditioning on the consequence of the Flattening Lemma being true, with probability at least $1 - \delta/2$.

2 Matrix Chernoff Bound

Theorem 1. (Matrix Chernoff Bound) Let $X_1, \ldots, X_s$ be independent copies of a symmetric random matrix $X \in \mathbb{R}^{d\times d}$ with $\mathbb{E}[X] = 0$, $\|X\|_2 \leq \gamma$ with probability 1, and $\|\mathbb{E}[X^\top X]\|_2 \leq \sigma^2$. Let $W = \frac{1}{s}\sum_{i\in[s]} X_i$. For any $\epsilon > 0$,

$$\mathbf{Pr}[\|W\|_2 > \epsilon] \leq 2d \cdot e^{-s\epsilon^2/(\sigma^2 + \frac{\gamma\epsilon}{3})} \tag{3}$$

Before we can apply the Matrix Chernoff Bound, we need to define our random matrix X. Let $V = HDA$, and recall that V has orthonormal columns. Furthermore, suppose the matrix P in our SRHT samples s rows uniformly and with replacement, scaling each row by a factor of $\sqrt{n/s}$. In other words, if row j is sampled in the ith sample, $P_{i,j} = \sqrt{n/s}$.

Now, let Y_i be the ith sampled row of V and let $X_i = I_d - nY_i^\top Y_i$.

Remark 1. Each X_i is symmetric, since I_d is symmetric, the outer product of a vector with itself is symmetric, and the linear combination of two symmetric matrices is symmetric.

Remark 2. Each Y_i was sampled uniformly with replacement, so each Y_i is independent, making each X_i independent as well.

Claim 1. Each matrix X_i satisfies the conditions for X_i in the Matrix Chernoff Bound, namely that they are independent and $\mathbb{E}[X_i] = 0$.

Proof. By Remark 2, each X_i is independent. Now we just need to show that $\mathbb{E}[X_i] = 0$. Recall that Y_i is the ith sampled row of V. Since Y_i was sampled uniformly, we have

$$\mathbb{E}[Y_i^\top Y_i] = \sum_{j=1}^{n} \mathbf{Pr}[Y_i = v_j] \cdot v_j^\top v_j = \sum_{j=1}^{n} \frac{1}{n} \cdot v_j^\top v_j = \frac{1}{n} V^\top V \tag{4}$$

Since V has orthogonal columns, $V^\top V = I_d$, meaning

$$\mathbb{E}[X_i] = \mathbb{E}[I_d - nY_i^\top Y_i] = I_d - n\mathbb{E}[Y_i^\top Y_i] = I_d - n \cdot \frac{1}{n} V^\top V = I_d - I_d = 0 \qquad \blacksquare$$

Claim 2. Each row vector Y_i of HDA satisfies $\left\| nY_i^\top Y_i \right\|_2 \leq n \cdot \max_j \|e_j HDA\|_2^2$.

Proof: Rewriting $nY_i^\top Y_i$, we have

$$nY_i^\top Y_i = Y_i^\top nY_i = \left(\frac{Y_i^\top}{\|Y_i\|_2} \right) n \|Y_i\|_2^2 \left(\frac{Y_i}{\|Y_i\|_2} \right) \tag{5}$$

It follows that $\left\| nY_i^\top Y_i \right\|_2 = n \|Y_i\|_2^2$. Also, Y_i is a row vector of HDA, which means for some $j \in [n]$, $Y_i = e_j HDA$. So, we can conclude that $\|Y_i\|_2 \leq \max_j \|e_j HDA\|_2$. Thus,

$$\left\| nY_i^\top Y_i \right\|_2 \leq n \cdot \max_j \|e_j HDA\|_2^2 \qquad \blacksquare$$

Claim 3. The matrices X_i satisfy $\|X_i\|_2 \leq \gamma$ for $\gamma = \Theta(d \log(nd/\delta))$.

Proof: The operator norm is a norm, which means it satisfies the triangle inequality.

$$\begin{aligned}
\|X_i\|_2 &= \left\| I_d - n \cdot Y_i^\top Y_i \right\|_2 &(6)\\
&\leq \|I_d\|_2 + \left\| nY_i^\top Y_i \right\|_2 &(7)\\
&\leq \|I_d\|_2 + n \cdot \max_j \|e_j HDA\|_2^2 &(8)\\
&\leq 1 + n \cdot \left(C\sqrt{d \log(nd/\delta)}/\sqrt{n} \right)^2 &(9)\\
&= 1 + C^2 d \log(nd/\delta) &(10)\\
&= \Theta(d \log(nd/\delta)) &(11)
\end{aligned}$$

(7) to (8) follows from Claim 2, and (8) to (9) follows from Corollary 1. $\blacksquare$

Claim 4. Letting X be the random matrix that $X_1, \ldots, X_s$ are independent copies of, we have $\left\| \mathbb{E}[X^\top X] \right\|_2 \leq \sigma^2$ where $\sigma^2 = O(d \log(nd/\delta))$.

Proof: We will come up with an expression for $\mathbb{E}[X^\top X + I_d]$. To do so, we will first come up with an expression for $\mathbb{E}[X^\top X]$. Recall each X_i is symmetric, so $X_i = X_i^\top$.

$$\begin{aligned}
\mathbb{E}[X^\top X] &= \mathbb{E}_i[X_i \cdot X_i] && (12)\\
&= \mathbb{E}_i[(I_d - nY_i^\top Y_i)^2] && (13)\\
&= \mathbb{E}_i[I_d - 2nY_i^\top Y_i + n^2 Y_i^\top Y_i Y_i^\top Y_i] && (14)\\
&= I_d - 2n\mathbb{E}_i[Y_i^\top Y_i] + n^2\mathbb{E}_i[Y_i^\top Y_i Y_i^\top Y_i] && (15)
\end{aligned}$$

We can solve for $\mathbb{E}_i[Y_i^\top Y_i]$ and $\mathbb{E}_i[Y_i^\top Y_i Y_i^\top Y_i]$. Recall that v_i is the ith row vector of matrix V, so $v_i^\top$ is a column vector and $v_i v_i^\top = \|v_i\|_2^2$.

$$\mathbb{E}_i[Y_i^\top Y_i] = \sum_{i=1}^{n} \frac{1}{n} v_i^\top v_i = \frac{1}{n} V^\top V = \frac{1}{n} \cdot I_d \tag{16}$$

$$\mathbb{E}_i[Y_i^\top Y_i Y_i^\top Y_i] = \sum_{i=1}^{n} \frac{1}{n} \cdot v_i^\top v_i v_i^\top v_i = \sum_{i=1}^{n} \frac{1}{n} \cdot v_i^\top (v_i v_i^\top) v_i = \frac{1}{n} \sum_{i=1}^{n} v_i^\top v_i \cdot \|v_i\|_2^2 \tag{17}$$

Now we can get an expression for $\mathbb{E}[X^\top X + I_d]$:

$$\begin{aligned}
\mathbb{E}[X^\top X + I_d] &= I_d + \mathbb{E}[X^\top X] && (18)\\
&= I_d + I_d - 2n\mathbb{E}_i[Y_i^\top Y_i] + n^2\mathbb{E}_i[Y_i^\top Y_i Y_i^\top Y_i] && (19)\\
&= 2I_d - 2n\left(\frac{1}{n} \cdot I_d\right) + n^2\left(\frac{1}{n}\sum_{i=1}^{n} v_i^\top v_i\right) \cdot \|v_i\|_2^2 && (20)\\
&= n\sum_{i=1}^{n} v_i^\top v_i \cdot \|v_i\|_2^2 && (21)
\end{aligned}$$

Now, we will define Z to be $Z = n\sum_i v_i^\top v_i C^2 \log(\frac{nd}{\delta}) \cdot \frac{d}{n}$.

Remark 3. We can rewrite Z to get $Z = C^2 d \log(\frac{nd}{\delta}) \sum_i v_i^\top v_i = C^2 d\log(\frac{nd}{\delta}) I_d$. From this, we can tell that $\|Z\|_2 = \left\|C^2 d\log(\frac{nd}{\delta}) I_d\right\|_2 = C^2 d\log(\frac{nd}{\delta}) \|I_d\|_2 = C^2 d \log(\frac{nd}{\delta})$.

We will use Loewner's ordering on positive semi-definite matrices to help us reach the desired bound for $\left\|\mathbb{E}[X^\top X]\right\|_2$.

Definition. (Loewner order) If A, B are positive semi-definite matrices, matrices whose eigenvalues are all non-negative, then $A \leq B$ if and only if for all vectors x, $x^\top A x \leq x^\top B x$.

Theorem 2. If $A \leq B$ in Loewner's ordering, then $\|A\|_2 \leq \|B\|_2$.

Noting that $\mathbb{E}[X^\top X + I_d]$ and Z are both real symmetric matrices with non-negative eigenvalues, if we can show that $\mathbb{E}[X^\top X + I_d] \leq Z$ in Loewner's ordering, then we can use Theorem 2 to get an upper bound on $\left\|\mathbb{E}[X^\top X + I_d]\right\|_2$ and in turn get an upper bound on $\left\|\mathbb{E}[X^\top X]\right\|_2$.

Claim 5. For all vectors y, $y^\top \mathbb{E}[X^\top X + I_d] y \leq y^\top Z y$; i.e., $\mathbb{E}[X^\top X + I_d] \leq Z$.

Proof: Using the result of (21) and the definition of Z, we have

$$y^\top \mathbb{E}[X^\top X + I_d]y = y^\top \left(n \sum_{i=1}^{n} v_i^\top v_i \cdot \|v_i\|_2^2 \right) y \tag{22}$$

$$= n \sum_{i=1}^{n} y^\top v_i^\top v_i y \cdot \|v_i\|_2^2 \tag{23}$$

$$= n \sum_{i=1}^{n} \langle v_i, y \rangle^2 \cdot \|v_i\|_2^2 \tag{24}$$

$$y^\top Z y = y^\top \left(n \sum_i v_i^\top v_i C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} \right) y \tag{25}$$

$$= n \sum_i y^\top v_i^\top v_i y \cdot C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} \tag{26}$$

$$= n \sum_i \langle v_i, y \rangle^2 \cdot C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} \tag{27}$$

$v_i = e_i V = e_i HDA$, so by Corollary 1,

$$\|v_i\|_2 \leq C \frac{\sqrt{d \log(nd/\delta)}}{\sqrt{n}} \tag{28}$$

$$\implies \|v_i\|_2^2 \leq C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} \tag{29}$$

So, we can conclude that

$$y^\top \mathbb{E}[X^\top X + I_d]y = n \sum_{i=1}^{n} \langle v_i, y \rangle^2 \cdot \|v_i\|_2^2 \tag{30}$$

$$\leq n \sum_{i=1}^{n} \langle v_i, y \rangle^2 \cdot C^2 \log\left(\frac{nd}{\delta}\right) \cdot \frac{d}{n} \tag{31}$$

$$= y^\top Z y \qquad \blacksquare$$

Now that we proved Claim 5, we can finish the proof for Claim 4. By Claim 5, Theorem 2 and Remark 3, $\left\|\mathbb{E}[X^\top X + I_d]\right\|_2 \leq \|Z\|_2 = C^2 d \log(\frac{nd}{\delta})$. We have

$$\left\|\mathbb{E}[X^\top X]\right\|_2 = \left\|\mathbb{E}[X^\top X] + I_d - I_d\right\|_2 \tag{32}$$

$$\leq \left\|\mathbb{E}[X^\top X] + I_d\right\|_2 + \|I_d\|_2 \tag{33}$$

$$= \left\|\mathbb{E}[X^\top X + I_d]\right\|_2 + 1 \tag{34}$$

$$\leq C^2 d \log\left(\frac{nd}{\delta}\right) + 1 \tag{35}$$

$$= O\left(d \log \frac{nd}{\delta}\right) \qquad \blacksquare$$

Now, we're finally ready to apply the Matrix Chernoff Bound. The matrix W can be expressed as

$$
\begin{aligned}
W &= \frac{1}{s}\sum_{i\in[s]} X_i && (36)\\
&= \frac{1}{s}\sum_{i\in[s]} I_d - nY_i^\top Y_i && (37)\\
&= \frac{1}{s}\left(sI_d - n\sum_{i\in[s]} Y_i^\top Y_i\right) && (38)\\
&= I_d - \sum_{i\in[s]}\left(Y_i^\top\sqrt{\frac{n}{s}}\right)\left(Y_i\sqrt{\frac{n}{s}}\right) && (39)
\end{aligned}
$$

Notice that by definition, each i represents the ith randomly sampled random matrix X_i, which corresponds to the ith randomly sampled row of $V = PHD$. Furthermore, $\sqrt{n/s}$ is equivalent to the scaling factor used in our SRHT matrix P. This means $Y_i\sqrt{n/s}$ corresponds exactly to the ith row of the sketch, $(PHDA)_i$. Thus,

$$W = I_d - (PHDA)^\top(PHDA) \tag{40}$$

By the Matrix Chernoff Bound, we get

$$\mathbf{Pr}[\left\|I_d - (PHDA)^\top(PHDA)\right\|_2 > \epsilon] \le 2d\cdot e^{-s\epsilon^2/(\sigma^2+\frac{\gamma\epsilon}{3})} = 2d\cdot e^{-s\epsilon^2/\Theta(d\log(nd/\delta))} \tag{41}$$

Set $s = d\log(nd/\delta)\frac{\log(d/\delta)}{\epsilon^2}$ to make this probability less than $\delta/2$.

3 SRHT Wrap Up

We have shown that with $s = d\log(nd/\delta)\frac{\log(d/\delta)}{\epsilon^2}$, we can achieve $\left\|I_d - (PHDA)^\top(PHDA)\right\|_2 < \epsilon$ with probability at least $1-\delta/2$. So, for every unit vector x, if we left and right multiply $I_d - (PHDA)^\top(PHDA)$ by x, we can get

$$|1 - \|PHDAx\|_2^2| = |x^\top x - x^\top(PHDA)^\top(PHDA)x| < \epsilon, \tag{42}$$

so $\|PHDAx\|_2^2 \in 1\pm\epsilon$ for all unit vectors x, proving that SRHT is a subspace embedding. We can then solve the regression problem in the same way we did last lecture, by considering the column span of A adjoined with b.

The time needed is $O(n\log n)$ to calculate Sb and $O(nd\log n)$ to calculate SA, plus an additional $\mathrm{poly}(d\log(n)/\epsilon)$ to compute the least squares approximation. The total time complexity is $O(nd\log n) + \mathrm{poly}(d\log(n)/\epsilon)$, which is nearly optimal in the matrix dimensions when $n >> d$.

4 Faster Subspace Embeddings

Using SRHT, we've managed to find a nearly optimal runtime with tight bounds for approximating linear regression on dense matrices A. So, a natural follow-up is whether or not we can further improve the time complexity on sparse matrices.

Definition. (CountSketch) The CountSketch Matrix is a $k \times n$ matrix S for $k = O(d^2/\epsilon^2)$, such that there is only a single randomly chosen non-zero entry for each column of S.

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 1: Example of a 4×8 CountSketch matrix

Claim 6. If we let $\text{nnz}(A)$ be the number of non-zero entries in A, then we can compute SA in $\text{nnz}(A)$ time.

A simple algorithm for doing this is to use a sparse representation of A (e.g., keep a list of non-zero entries of A with the positions of said entries), and then iterate over the non-zero entries in A, multiplying each entry by the corresponding column in S. Since each column in S only has one non-zero entry, this can be done in constant time for each entry in A, for a total of $\text{nnz}(A)$ time.

4.1 CountSketch matrix S is a subspace embedding

As with our previous proofs of subspace embeddings, in order to show S is a subspace embedding, we can assume the columns of A are orthonormal, and it suffices to show that $\|SAx\|_2 = 1 \pm \epsilon$ for all unit x. We can then apply S to the matrix with b adjoined to the columns of A for regression. Let $k = 6d^2/(\delta\epsilon^2)$, so SA is a $6d^2/(\delta\epsilon^2) \times d$ matrix.

Claim 7. To show that S is a subspace embedding, it suffices to show $\left\|A^\top S^\top SA - I\right\|_F \leq \epsilon$.

Proof: Suppose we showed that $\left\|A^\top S^\top SA - I\right\|_F \leq \epsilon$. Since $\left\|A^\top S^\top SA - I\right\|_2 \leq \left\|A^\top S^\top SA - I\right\|_F$, we get

$$\begin{aligned} &\left\|A^\top S^\top SA - I\right\|_2 \leq \epsilon && (43) \\ \implies &|x^\top A^\top S^\top SAx - x^\top x| \leq \epsilon && (44) \\ \implies &|\, \|SAx\|_2^2 - 1| \leq \epsilon && (45) \\ \implies &\|SAx\|_2^2 = 1 \pm \epsilon && (46) \\ \implies &\|SAx\|_2 = 1 \pm O(\epsilon) && (47) \end{aligned}$$

as desired.

Lemma 2. *(Matrix Product Result) For matrices C, D, and S,*

$$\mathbf{Pr}[\left\|CS^\top SD - CD\right\|_F^2 \leq [6/(\delta(\#\text{ rows of } S))] \cdot \|C\|_F^2 \|D\|_F^2] \geq 1 - \delta \tag{48}$$

We will use the matrix product result first, and then prove it later. Let $C = A^\top$ and $D = A$. Notice that since A has orthonormal columns, the norm of each column of A is 1, so the squared Frobenius norm of A is just the number of columns; i.e., $\|A\|_F^2 = d$. Also, A is an orthogonal matrix, so $A^\top A = I$. We use the CountSketch matrix for S, so (# rows of S) $= 6d^2/(\delta\epsilon^2)$. By the matrix product result, we get

$$\mathbf{Pr}[\left\|A^\top S^\top SA - A^\top A\right\|_F^2 \leq \left[6/(\delta(6d^2/(\delta\epsilon^2)))\right] \cdot \left\|A^\top\right\|_F^2 \|A\|_F^2] \tag{49}$$

$$= \mathbf{Pr}[\left\|A^\top S^\top SA - I\right\|_F^2 \leq (\epsilon^2/d^2) \cdot d \cdot d] \tag{50}$$

$$= \mathbf{Pr}[\left\|A^\top S^\top SA - I\right\|_F^2 \leq \epsilon^2] \tag{51}$$

$$= \mathbf{Pr}[\left\|A^\top S^\top SA - I\right\|_F \leq \epsilon] \geq 1 - \delta \tag{52}$$

So, by Claim 7, S is a subspace embedding w.p. at least $1 - \delta$.

4.2 Matrix Product Result

We now show that we can use the matrix product result for the CountSketch matrix. Recall the matrix product result

$$\mathbf{Pr}[\left\|CS^\top SD - CD\right\|_F^2 \leq [6/(\delta(\text{\# rows of } S))] \cdot \|C\|_F^2 \|D\|_F^2] \geq 1 - \delta \tag{53}$$

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

$$\mathbb{E}_S[\|Sx\|_2^2 - 1]^\ell \leq \epsilon^\ell \cdot \delta \tag{54}$$

The goal is to first show that the JL Property implies the matrix product result, and then show that CountSketch satisfies the JL Property.

Claim 8. (From vectors to matrices) For $\epsilon, \delta \in (0, 1/2)$, let 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,

$$\mathbf{Pr}_S[\left\|A^\top S^\top SB - A^\top B\right\|_F \geq 3\epsilon \|A\|_F \|B\|_F] \leq \delta \tag{55}$$

Before we prove this, we will introduce and prove Minkowski's Inequality.

Definition. For a random scalar X, define the norm $\|\cdot\|_p$ as $(\mathbb{E}[|X|^p])^{1/p}$.

Lemma 3. *(Minkowski's Inequality) For any matrices X and Y,*

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p \tag{56}$$

Proof: Suppose we have matrices X, Y, where $\|X\|_p$ and $\|Y\|_p$ are both finite. The function $f(x) = |x|^p$ is convex for $p \geq 1$, which means $f(\frac{x+y}{2}) \leq \frac{1}{2}f(x) + \frac{1}{2}f(y)$. So, for any fixed x and y,

$$\left|\frac{1}{2}x + \frac{1}{2}y\right|^p \leq \left|\frac{1}{2}|x| + \frac{1}{2}|y|\right|^p \leq \frac{1}{2}|x|^p + \frac{1}{2}|y|^p \tag{57}$$

$$2^p\left|\frac{1}{2}x + \frac{1}{2}y\right|^p \leq 2^p\left(\frac{1}{2}|x|^p + \frac{1}{2}|y|^p\right) \tag{58}$$

$$|x + y|^p \leq 2^{p-1}(|x|^p + |y|^p) \tag{59}$$

So, $\mathbb{E}[|X+Y|^p] \leq \mathbb{E}[2^{p-1}(|X|^p + |Y|^p)]$. By definition, $(\mathbb{E}[|X+Y|^p])^{1/p} = \|X+Y\|_p \implies \mathbb{E}[|X+Y|^p] = \|X+Y\|_p^p$. It follows that since $\mathbb{E}[|X+Y|^p]$ is finite, $\|X+Y\|_p$ is finite. Now, we can get an upper bound for $\|X+Y\|_p^p$:

$$\begin{aligned}
\|X+Y\|_p^p &= \int |x+y|^p d\mu && (60)\\
&= \int |x+y| \cdot |x+y|^{p-1} d\mu && (61)\\
&\leq (|x| + |y|)|x+y|^{p-1} d\mu && (62)\\
&= \int |x||x+y|^{p-1} d\mu + \int |y||x+y|^{p-1} d\mu && (63)
\end{aligned}$$

Theorem 3. (Hölder's Inequality) For vectors u, v, and scalars p, q such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\langle u, v \rangle \leq \|u\|_p \|v\|_q = \left(\sum |u_i|^p\right)^{1/p} \left(\sum |v_i|^q\right)^{1/q} \tag{64}$$

Applying Hölder's Inequality, with the norm of the first vector being p and the norm of the second vector being $\frac{p}{p-1}$, we get

$$\begin{aligned}
\int |x||x+y|^{p-1} d\mu &\leq \left(\int |x|^p d\mu\right)^{1/p} \left(\int \left(|x+y|^{p-1}\right)^{\frac{p}{p-1}} d\mu\right)^{(p-1)/p} && (65)\\
\int |y||x+y|^{p-1} d\mu &\leq \left(\int |y|^p d\mu\right)^{1/p} \left(\int \left(|x+y|^{p-1}\right)^{\frac{p}{p-1}} d\mu\right)^{(p-1)/p} && (66)
\end{aligned}$$

So,

$$\begin{aligned}
\|X+Y\|_p^p &\leq \left(\left(\int |x|^p d\mu\right)^{1/p} + \left(\int |y|^p d\mu\right)^{1/p}\right) \left(\int |x+y|^p d\mu\right)^{(p-1)/p} && (67)\\
&= \left((\mathbb{E}[|X|^p])^{1/p} + (\mathbb{E}[|Y|^p])^{1/p}\right) \left(\mathbb{E}[|X+Y|^p]^{1/p}\right)^{p-1} && (68)\\
&= (\|X\|_p + \|Y\|_p) \|X+Y\|_p^{p-1} && (69)
\end{aligned}$$

$$\|X+Y\|_p \leq \|X\|_p + \|Y\|_p \qquad \blacksquare$$

Now that we proved Minkowski's inequality, we can proceed to prove the matrix product result in the next lecture.