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 $C > 0$,

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

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

$$
\|e_jHDA\|_2 \leq C \frac{\sqrt{d\log(\frac{nd}{\delta})}}{\sqrt{n}}
$$

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 $\|\mathbb{E}[X^T X]\|_2 \leq \sigma^2$. Let $W = \frac{1}{s} \sum_{i \in [s]} X_i$. For any $\epsilon > 0$,

$$
\Pr[\|W\|_2 > \epsilon] \leq 2d \cdot e^{-\frac{\epsilon^2}{2(\sigma^2 + \frac{\gamma^2}{4})}}
$$

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 $i$th sample, $P_{i,j} = \sqrt{n/s}$.

Now, let $Y_i$ be the $i$th sampled row of $V$ and let $X_i = I_d - nY_i^T Y_i$.

**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 $i$th sampled row of $V$. Since $Y_i$ was sampled uniformly, we have

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

(4)

Since $V$ has orthogonal columns, $V^T V = I_d$, meaning

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

Claim 2. Each row vector $Y_i$ of HDA satisfies $\|nY_i^T Y_i\|_2 \leq n \cdot \max_i \|e_j HDA\|_2^2$.

Proof: Rewriting $nY_i^T Y_i$, we have

$$nY_i^T Y_i = Y_i^T nY_i = \left( \frac{Y_i^T}{\|Y_i\|_2} \right) n \cdot \|Y_i\|_2^2 \left( \frac{Y_i}{\|Y_i\|_2} \right)$$

(5)

It follows that $\|nY_i^T Y_i\|_2 = n \cdot \|Y_i\|_2^2$. 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 $\|Y_i\|_2 \leq \max_i \|e_j HDA\|_2$. Thus,

$$\|nY_i^T Y_i\|_2 \leq n \cdot \max_j \|e_j HDA\|_2^2$$

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.

$$\|X_i\|_2 = \left\| I_d - n \cdot Y_i^T Y_i \right\|_2$$

(6)

$$\leq \|I_d\|_2 + \|nY_i^T Y_i\|_2$$

(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)

(7) to (8) follows from Claim 2 and (8) to (9) follows from Corollary 1.

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

2
Proof: We will come up with an expression for \( \mathbb{E}[X^TX + I_d] \). To do so, we will first come up with an expression for \( \mathbb{E}[X^TX] \). Recall each \( X_i \) is symmetric, so \( X_i = X_i^\top \).

\[
\mathbb{E}[X^TX] = \mathbb{E}_i[X_i\cdot X_i]
\]

\[
= \mathbb{E}_i[(I_d - nY_i^\top Y_i)^2] 
\]

\[
= \mathbb{E}_i[I_d - 2nY_i^\top Y_i + n^2Y_i^\top Y_iY_i^\top Y_i] 
\]

\[
= I_d - 2n\mathbb{E}_i[Y_i^\top Y_i] + n^2\mathbb{E}_i[Y_i^\top Y_iY_i^\top Y_i] 
\]

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

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

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

Now we can get an expression for \( \mathbb{E}[X^TX + I_d] \):

\[
\mathbb{E}[X^TX + I_d] = I_d + \mathbb{E}[X^TX] 
\]

\[
= I_d + I_d - 2n\mathbb{E}_i[Y_i^\top Y_i] + n^2\mathbb{E}_i[Y_i^\top Y_iY_i^\top Y_i] 
\]

\[
= 2I_d - 2n \left( \frac{1}{n} \cdot I_d \right) + n^2 \left( \frac{1}{n} \sum_{i=1}^{n} v_i^\top v_i \right) \cdot \|v_i\|^2 
\]

\[
= n \sum_{i=1}^{n} v_i^\top v_i \cdot \|v_i\|^2 
\]

Now, we will define \( Z \) to be \( Z = n \sum_{i=1}^{n} v_i^\top v_i C^2 \log(\frac{nd}{n}) \cdot \frac{d}{n} \).

Remark 3. We can rewrite \( Z \) to get \( Z = C^2d \log(\frac{nd}{n}) \sum_{i=1}^{n} v_i^\top v_i = C^2d \log(\frac{nd}{n}) I_d \). From this, we can tell that \( \|Z\|_2 = \left\| C^2d \log(\frac{nd}{n}) I_d \right\|_2 = C^2d \log(\frac{nd}{n}) \|I_d\|_2 = C^2d \log(\frac{nd}{n}) \).

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

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 \( x \), \( x^\top Ax \leq x^\top Bx \).

Theorem 2. If \( A \leq B \) in Loewner’s ordering, then \( \|A\|_2 \leq \|B\|_2 \).

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

Claim 5. For all vectors \( y \), \( y^\top \mathbb{E}[X^TX + I_d] y \leq y^\top Z y \); i.e., \( \mathbb{E}[X^TX + I_d] \leq Z \).
Proof: Using the result of (21) and the definition of $Z$, we have

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

(22)

\[ y^\top Zy = y^\top \left( n \sum_{i} v_i^\top v_i C^2 \log \left( \frac{nd}{\delta} \right) \cdot \frac{d}{n} \right) y \]  
\[ = n \sum_{i} y^\top v_i y \cdot C^2 \log \left( \frac{nd}{\delta} \right) \cdot \frac{d}{n} \]  
\[ = n \sum_{i} (v_i, y)^2 \cdot C^2 \log \left( \frac{nd}{\delta} \right) \cdot \frac{d}{n} \]  

(23)

$v_i = e_i V = e_i HDA$, so by Corollary 3

\[ \|v_i\|_2 \leq C \frac{\sqrt{\log(n)\log(nd/\delta)}}{\sqrt{n}} \]  
\[ \implies \|v_i\|_2^2 \leq C^2 \log \left( \frac{nd}{\delta} \right) \cdot \frac{d}{n} \]  

(24)

So, we can conclude that

\[ y^\top \mathbb{E}[X^\top X + I_d]y = n \sum_{i=1}^n (v_i, y)^2 \cdot \|v_i\|_2^2 \]  
\[ \leq n \sum_{i=1}^n (v_i, y)^2 \cdot C^2 \log \left( \frac{nd}{\delta} \right) \cdot \frac{d}{n} \]  
\[ = y^\top Zy \]  

(25)

(26)

(27)

(28)

(29)

(30)

(31)

(32)

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

\[ \mathbb{E}[X^\top X]_2 = \|\mathbb{E}[X^\top X] + I_d - I_d\|_2 \]  
\[ \leq \|\mathbb{E}[X^\top X] + I_d\|_2 + \|I_d\|_2 \]  
\[ = \|\mathbb{E}[X^\top X] + I_d\|_2 + 1 \]  
\[ \leq C^2 d \log \left( \frac{nd}{\delta} \right) + 1 \]  
\[ = O \left( d \log \frac{nd}{\delta} \right) \]  

(33)

(34)

(35)

4
Now, we’re finally ready to apply the Matrix Chernoff Bound. The matrix $W$ can be expressed as

$$W = \frac{1}{s} \sum_{i \in [s]} X_i \quad (36)$$

$$= \frac{1}{s} \sum_{i \in [s]} I_d - nY_i^\top Y_i \quad (37)$$

$$= \frac{1}{s} \left( sI_d - n \sum_{i \in [s]} Y_i^\top Y_i \right) \quad (38)$$

$$= I_d - \sum_{i \in [s]} \left( Y_i^\top \sqrt{\frac{n}{s}} \right) \left( Y_i \sqrt{\frac{n}{s}} \right) \quad (39)$$

Notice that by definition, each $i$ represents the $i$th randomly sampled random matrix $X_i$, which corresponds to the $i$th 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 $i$th row of the sketch, $(PHDA)_i$. Thus,

$$W = I_d - (PHDA)^\top (PHDA) \quad (40)$$

By the Matrix Chernoff Bound, we get

$$\Pr[\left\| I_d - (PHDA)^\top (PHDA) \right\|_2 > \epsilon] \leq 2d \cdot e^{-se^2/(\sigma^2 + \epsilon^2)} = 2d \cdot e^{-se^2/\Theta(d \log(n/d))} \quad (41)$$

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

### 3 SRHT Wrap Up

We have shown that with $s = d \log(nd/\delta) \log(d/\epsilon^2)$, 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

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

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 $\text{poly}(d \log(n)/\epsilon)$ to compute the least squares approximation. The total time complexity is $O(nd \log n) + \text{poly}(d \log(n)/\epsilon)$, 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 $S$.

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

Figure 1: Example of a $4 \times 8$ 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 $\|A^\top S^\top SA - I\|_F \leq \epsilon$.

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

$$
\|A^\top S^\top SA - I\|_2 \leq \epsilon \quad \Rightarrow \quad |x^\top A^\top S^\top SAx - x^\top x| \leq \epsilon
$$

$$
\Rightarrow \quad \|SAx\|_2^2 - 1 \leq \epsilon
$$

$$
\Rightarrow \quad \|SAx\|_2^2 = 1 \pm \epsilon
$$

$$
\Rightarrow \quad \|SAx\|_2 = 1 \pm O(\epsilon)
$$

as desired.

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

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

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

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

(49)

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

(50)

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

(51)

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

(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

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

**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 $\|x\|_2 = 1$,

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

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 $(\epsilon, \delta, \ell)$-JL moment property for some $\ell \geq 2$. Then, for matrices $A, B$ with $n$ rows,

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

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 $f\left( \frac{x + y}{2} \right) \leq \frac{1}{2} f(x) + \frac{1}{2} f(y)$. So, for any fixed $x$ and $y$,

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

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

$$|x + y|^p \leq 2^{p-1} (|x|^p + |y|^p) \tag{59}$$
So, $E[|X + Y|^p] \leq E[2^{p-1}(|X|^p + |Y|^p)]$. By definition, $(E[|X + Y|^p])^{1/p} = \|X + Y\|_p \implies E[|X + Y|^p] = \|X + Y\|_p^p$. It follows that since $E[|X + Y|^p]$ is finite, $\|X + Y\|_p$ is finite. Now, we can get an upper bound for $\|X + Y\|_p$.

$$\|X + Y\|_p^p = \int |x + y|^pd\mu$$  (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)

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

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

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

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

So,

$$\|X + Y\|_p^p \leq \left(\left(\int |x|^p d\mu\right)^{1/p} + \left(\int |y|^p d\mu\right)^{1/p}\right) \left(\int |x + y|^p d\mu\right)^{(p-1)/p}$$  (67)

$$= \left(E[|X|^p]\right)^{1/p} + \left(E[|Y|^p]\right)^{1/p} \left(E[|X + Y|^p]\right)^{(p-1)/p}$$  (68)

$$= (\|X\|_p^p + \|Y\|_p^p) \|X + Y\|_p^{p-1}$$  (69)

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

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