We have already shown that by using Subsampled Randomized Hadamard Transform, one can improve the time for regression to $O(nd \log n) + \text{poly}((d \log n)/\varepsilon)$. 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 $\pm 1$ with equal probability.
2. All the other entries are assigned 0.

When $A$ is sparse, we are able to compute $S \cdot A$ in $O(\text{nnz}(A))$ 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 $1 - \delta$:

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

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

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

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:

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

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

**Remark 1.** While the slides keep the notation $\varepsilon$ 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 $\|x\| = 1$,

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

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

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

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

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

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 $x, y$,

$$\|\langle Sx, Sy \rangle - \langle x, y \rangle\|_\ell$$

$$= \frac{1}{2} \left\| (\|Sx\|_2^2 - 1) + (\|Sy\|_2^2 - 1) - (\|S(x - y)\|_2^2 - \|x - y\|_2^2) \right\|_\ell$$

$$\leq \frac{1}{2} \left[ \left\| (\|Sx\|_2^2 - 1) \right\|_\ell + \left\| (\|Sy\|_2^2 - 1) \right\|_\ell + \left\| (\|S(x - y)\|_2^2 - \|x - y\|_2^2) \right\|_\ell \right]$$

$$\leq \frac{1}{2} \left( \hat{\varepsilon} \cdot \delta^{1/\ell} + \hat{\varepsilon} \cdot \delta^{1/\ell} + \|x - y\|_2^2 \cdot \hat{\varepsilon} \cdot \delta^{1/\ell} \right)$$

$$\leq 3\hat{\varepsilon} \cdot \delta^{1/\ell}$$

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
\[
\|A^\top S^\top SB - A^\top B\|_F^2 = \sum_{i,j} (\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle)^2 \|/2 \\
\leq \sum_{i,j} (\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle)^2 \\
= \sum_{i,j} (\langle SA_i, SB_j \rangle - \langle A_i, B_j \rangle)^2 _\ell \\
\leq (3\varepsilon \cdot \delta^{1/\ell})^2 \sum_{i,j} A_i^2 B_j^2 \\
= (3\varepsilon \cdot \delta^{1/\ell})^2 \|A\|_F^2 \|B\|_F^2.
\]

Note that the first inequality relies on Minkowski’s inequality, which requires \( \ell \geq 2 \). With
\[
\mathbb{E}\left[\|A^\top S^\top SB - A^\top B\|_F^\ell\right] = \mathbb{E}\left[\|A^\top S^\top SB - A^\top B\|_F^2 \right]^{\ell/2},
\]
we may complete the proof with Markov’s inequality:
\[
\mathbb{P}\left[\|A^\top S^\top SB - A^\top B\|_F > 3\varepsilon \|A\|_F \|B\|_F \right] \leq \left(\frac{1}{3\varepsilon \|A\|_F \|B\|_F}\right)^\ell \mathbb{E}\left[\|A^\top S^\top SB - A^\top B\|_F^\ell\right] \\
\leq \delta.
\]

We then proceed to the JL property of CountSketch matrix.

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

**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: \( [S]_{f(i),j} = \sigma(i) \).
- \( h \) is 2-wise independent: \( \forall i \neq j \in [n] \) and \( \forall z_1, z_2 \in [k] \), \( \mathbb{P}[h(i) = z_1, h(j) = z_2] = 1/k^2 \). \( \sigma \) is 4-wise independent, where the definition is similar.
Let \( \delta(E) \) be the indicator function of event \( E \). We have

\[
\mathbb{E}
\left[
\|Sx\|^2_2
\right]
= \sum_{j \in [k]} \mathbb{E}
\left[
\left( \sum_{i \in [n]} \delta(h(i) = j)\sigma(i)x_i \right)^2
\right]
\]

\[
= \sum_{j \in [k]} \sum_{i_1, i_2 \in [n]} \mathbb{E}
\left[
\delta(h(i_1) = j)\delta(h(i_2) = j)\sigma(i_1)\sigma(i_2)x_{i_1}x_{i_2}
\right]
\]

\[
= \sum_{j \in [k]} \sum_{i \in [n]} \mathbb{E}
\left[
(\delta(h(i) = j))^2
\right] x_i^2
\]

\[
= \frac{1}{k} \sum_{j \in [k]} \sum_{i \in [n]} x_i^2 = \|x\|^2_2.
\]

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

\[
\delta \geq \Pr
\left[
\|A^T S^\top S B - A^T B\|^2_F \geq 3\hat{\varepsilon} \|A\|^2_F \|B\|^2_F
\right]
\]

\[
= \Pr
\left[
\|A^T S^\top S B - A^T B\|^2_F \geq 9\hat{\varepsilon}^2 \|A\|^2_F \|B\|^2_F
\right]
\]

\[
= \Pr
\left[
\|A^T S^\top S B - A^T B\|^2_F = 18/(k\delta) \|A\|^2_F \|B\|^2_F
\right]
\]

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

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

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

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

$$\|X + Y\|_p^p = \int_{x,y} \|x + y\|^p d\mu$$

$$\leq \int_{x,y} (\|x\| + \|y\|) \|x + y\|^{p-1} d\mu$$

$$= \int_{x,y} \|x\| \|x + y\|^{p-1} d\mu + \int_{x,y} \|y\| \|x + y\|^{p-1} d\mu$$

$$\leq \left((\int_x \|x\|^p d\mu)^{1/p} + (\int_y \|y\|^p d\mu)^{1/p}\right) \left(\int_{x,y} \|x + y\|^{(p-1)p/(p-1)} d\mu\right)^{(p-1)/p}$$

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