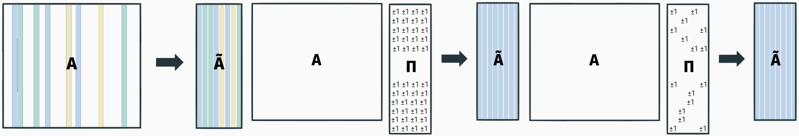
CS-GY 9223 I: Lecture 12
Randomized numerical linear algebra, fast Johnson-Lindenstrauss Transform

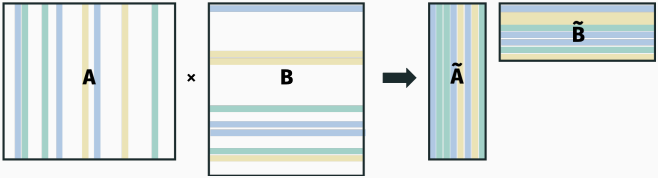
NYU Tandon School of Engineering, Prof. Christopher Musco

Main idea: If you want to compute singular vectors or eigenvectors, multiply two matrices, solve a regression problem, etc.:

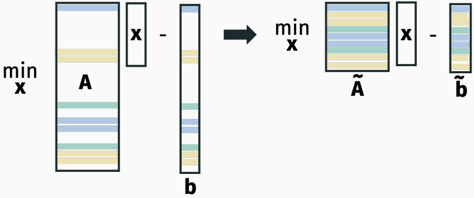
1. Compress your matrices using a randomized method.
2. Solve the problem on the smaller or sparser matrix.
    - $\tilde{\mathbf{A}}$ called a "sketch" or "coreset" for $\mathbf{A}$.



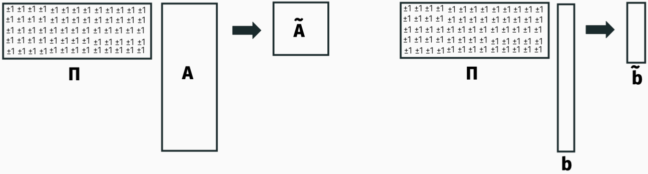
Approximate matrix multiplication:



Approximate regression:



Randomized approximate regression using a Johnson-Lindenstrauss Matrix:



Input: $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^{n}$.

Algorithm: Let $\tilde{\mathbf{x}}^* = \arg\min_{\mathbf{x}} \|\mathbf{\Pi}\mathbf{A}\mathbf{x} - \mathbf{\Pi}\mathbf{b}\|_2^2$.

Goal: Want $\|\mathbf{A}\tilde{\mathbf{x}}^* - \mathbf{b}\|_2^2 \leq (1+\epsilon)\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$

If $\mathbf{\Pi} \in \mathbb{R}^{m\times n}$, how large does m need to be? Is it even clear this should work as $m \to \infty$?

Theorem (Randomized Linear Regression)

Let $\mathbf{\Pi}$ *be a properly scaled JL matrix (random Gaussian, sign, sparse random, etc.) with* $m = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$ *rows. Then with probability* $(1-\delta)$*, for any* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and* $\mathbf{b} \in \mathbb{R}^n$*,*

$$\|\mathbf{A}\tilde{\mathbf{x}}^* - \mathbf{b}\|_2^2 \leq (1+\epsilon)\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

where $\tilde{\mathbf{x}}^* = \arg\min_{\mathbf{x}} \|\mathbf{\Pi}\mathbf{A}\mathbf{x} - \mathbf{\Pi}\mathbf{b}\|_2^2$.

Claim: Suffices to prove that for all $\mathbf{x} \in \mathbb{R}^d$,

$$(1-\epsilon)\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \leq \|\mathbf{\Pi}\mathbf{A}\mathbf{x}-\mathbf{\Pi}\mathbf{b}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$$

Lemma (Distributional JL)

If $\mathbf{\Pi}$ *is chosen to a properly scaled random Gaussian matrix, sign matrix, sparse random matrix, etc., with* $O\left(\frac{\log(1/\delta}{\epsilon^2}\right)$ *rows then for any fixed* $\mathbf{y}$,

$$(1-\epsilon)\|\mathbf{y}\|_2^2 \leq \|\mathbf{\Pi y}\|_2^2 \leq (1+\epsilon)\|\mathbf{y}\|_2^2$$

with probability $(1-\delta)$.

Corollary: For any fixed $\mathbf{x}$, with probability $(1-\delta)$,

$$(1-\epsilon)\|\mathbf{Ax}-\mathbf{b}\|_2^2 \leq \|\mathbf{\Pi Ax}-\mathbf{\Pi b}\|_2^2 \leq (1+\epsilon)\|\mathbf{Ax}-\mathbf{b}\|_2^2.$$

How do we go from "for any fixed $\mathbf{x}$" to "for all $\mathbf{x} \in \mathbb{R}^d$".

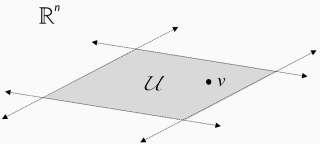
This statement requires establishing a Johnson-Lindenstrauss type bound for an infinity of possible vectors $(\mathbf{A}\mathbf{x} - \mathbf{b})$, which obviously can't be tackled with a union bound argument.

Theorem (Subspace Embedding from JL)

Let $\mathcal{U} \subset \mathbb{R}^n$ *be a d-dimensional linear subspace in* $\mathbb{R}^n$. *If* $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ *is chosen from any distribution* $\mathcal{D}$ *satisfying the Distributional JL Lemma, then with probability* $1 - \delta$,

$$(1-\epsilon)\|\mathbf{v}\|_2^2 \leq \|\mathbf{\Pi}\mathbf{v}\|_2^2 \leq (1+\epsilon)\|\mathbf{v}\|_2^2$$

for <u>*all*</u> $\mathbf{v} \in \mathcal{U}$, *as long as* $m = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$[1].



[1] It's possible to obtain a slightly tighter bound of $O\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)$. It's a nice challenge to try proving this.

Corollary: If we choose $\mathbf{\Pi}$ and properly scale, then with $O\left(d/\epsilon^2\right)$ rows,

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \leq \|\mathbf{\Pi}\mathbf{A}\mathbf{x} - \mathbf{\Pi}\mathbf{b}\|_2^2 \leq (1+\epsilon)\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

<u>for all $\mathbf{x}$</u> and thus

$$\|\mathbf{A}\tilde{\mathbf{x}}^* - \mathbf{b}\|_2^2 \leq (1+\epsilon)\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2.$$

I.e., our main theorem is proven.

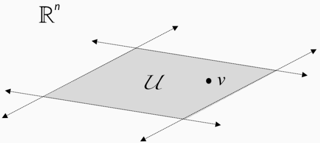
Proof: Apply Subspace Embedding Thm. to the $(d+1)$ dimensional subspace spanned by $\mathbf{A}$'s d columns and $\mathbf{b}$. Every vector $\mathbf{A}\mathbf{x} - \mathbf{b}$ lies in this subspace.

Theorem (Subspace Embedding from JL)

Let $\mathcal{U} \subset \mathbb{R}^n$ be a d-dimensional linear subspace in $\mathbb{R}^n$. If $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ is chosen from any distribution $\mathcal{D}$ satisfying the Distributional JL Lemma, then with probability $1 - \delta$,

$$(1-\epsilon)\|\mathbf{v}\|_2^2 \leq \|\mathbf{\Pi}\mathbf{v}\|_2^2 \leq (1+\epsilon)\|\mathbf{v}\|_2^2 \tag{1}$$

for all $\mathbf{v} \in \mathcal{U}$, as long as $m = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$



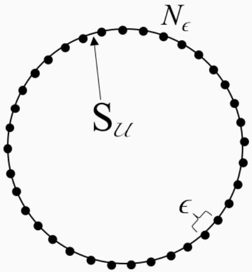
Observation: The theorem holds as long as (1) holds for all $\mathbf{w}$ on the unit sphere in $\mathcal{U}$. Denote the sphere $S_{\mathcal{U}}$:

$$S_{\mathcal{U}} = \{\mathbf{w} \mid \mathbf{w} \in \mathcal{U} \text{ and } \|\mathbf{w}\|_2 = 1\}.$$

Follows from linearity: Any point $\mathbf{v} \in \mathcal{U}$ can be written as $c\mathbf{w}$ for some scalar c and some point $\mathbf{w} \in S_{\mathcal{U}}$.

- If $(1-\epsilon)\|\mathbf{w}\|_2 \leq \|\mathbf{\Pi}\mathbf{w}\|_2 \leq (1+\epsilon)\|\mathbf{w}\|_2$.
- then $c(1-\epsilon)\|\mathbf{w}\|_2 \leq c\|\mathbf{\Pi}\mathbf{w}\|_2 \leq c(1+\epsilon)\|\mathbf{w}\|_2$,
- and thus $(1-\epsilon)\|c\mathbf{w}\|_2 \leq \|\mathbf{\Pi}c\mathbf{w}\|_2 \leq (1+\epsilon)\|c\mathbf{w}\|_2$.

Intuition: There are not too many "different" points on a d-dimensional sphere:



N_ϵ is called an "ϵ"-net.

If we can prove

$$(1-\epsilon) \leq \|\Pi \mathbf{w}\|_2 \leq (1+\epsilon)$$

for all points $\mathbf{w} \in N_\epsilon$, we can hopefully extend to all of $S_\mathcal{U}$.

Lemma (ϵ-net for the sphere)

For any $\epsilon \leq 1$, there exists a set $N_\epsilon \subset S_{\mathcal{U}}$ with $|N_\epsilon| = \left(\frac{4}{\epsilon}\right)^d$ such that $\forall \mathbf{v} \in S_{\mathcal{U}}$,

$$\min_{\mathbf{w} \in N_\epsilon} \|\mathbf{v} - \mathbf{w}\| \leq \epsilon.$$

1. Preserving norms of all points in net N_ϵ.

Set $\delta' = \left(\frac{\epsilon}{4}\right)^d \cdot \delta$. By a union bound, with probability $1-\delta$, for all $\mathbf{w} \in N_\epsilon$,

$$(1-\epsilon) \leq \|\Pi \mathbf{w}\|_2 \leq (1+\epsilon).$$

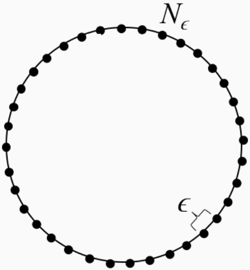
as long as Π has $O\left(\frac{\log(1/\delta')}{\epsilon^2}\right) = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$ rows.

2. Writing any point in sphere as linear comb. of points in N_ϵ.

For some $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2 \ldots \in N_\epsilon$, any $\mathbf{v} \in S_{\mathcal{U}}$. can be written:

$$\mathbf{v} = \mathbf{w}_0 + c_1\mathbf{w}_1 + c_2\mathbf{w}_2 + \ldots$$

for constants $c_1, c_2, \ldots$ where $|c_i| \leq \epsilon^i$.



3. Preserving norm of v.

Applying triangle inequality, we have

$$\begin{aligned}
\|\mathbf{\Pi v}\|_2 &= \|\mathbf{\Pi w}_0 + c_1\mathbf{\Pi w}_1 + c_2\mathbf{\Pi w}_2 + \ldots\| \\
&\leq \|\mathbf{\Pi w}_0\| + \epsilon\|\mathbf{\Pi w}_1\| + \epsilon^2\|\mathbf{\Pi w}_2\| + \ldots \\
&\leq (1+\epsilon) + \epsilon(1+\epsilon) + \epsilon^2(1+\epsilon) + \ldots \\
&\leq 1 + O(\epsilon).
\end{aligned}$$

3. Preserving norm of v.

Similarly,

$$\begin{aligned}\|\mathbf{\Pi v}\|_2 &= \|\mathbf{\Pi w}_0 + c_1\mathbf{\Pi w}_1 + c_2\mathbf{\Pi w}_2 + \ldots\| \\ &\geq \|\mathbf{\Pi w}_0\| - \epsilon\|\mathbf{\Pi w}_1\| - \epsilon^2\|\mathbf{\Pi w}_2\| - \ldots \\ &\geq (1-\epsilon) - \epsilon(1+\epsilon) - \epsilon^2(1+\epsilon) - \ldots \\ &\geq 1 - O(\epsilon).\end{aligned}$$

So we have proven

$$1 - O(\epsilon) \leq \|\mathbf{\Pi v}\|_2 \leq 1 + O(\epsilon)$$

for all $\mathbf{v} \in S_{\mathcal{U}}$, which in turn implies for small ϵ,

$$1 - O(\epsilon) \leq \|\mathbf{\Pi v}\|_2^2 \leq 1 + O(\epsilon)$$

Adjusting ϵ proves the Subspace Embedding theorem.

Theorem (Subspace Embedding from JL)

Let $\mathcal{U} \subset \mathbb{R}^n$ be a d-dimensional linear subspace in $\mathbb{R}^n$. If $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ is chosen from any distribution $\mathcal{D}$ satisfying the Distributional JL Lemma, then with probability $1 - \delta$,

$$(1-\epsilon)\|\mathbf{v}\|_2 \leq \|\mathbf{\Pi}\mathbf{v}\|_2 \leq (1+\epsilon)\|\mathbf{v}\|_2 \tag{2}$$

for all $\mathbf{v} \in \mathcal{U}$, as long as $m = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$

Theorem (Randomized Linear Regression)

Let $\mathbf{\Pi}$ *be a properly scaled JL matrix (random Gaussian, sign, sparse random, etc.) with* $m = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$ *rows. Then with probability* $(1-\delta)$, *for any* $\mathbf{A} \in \mathbb{R}^{n\times d}$ *and* $\mathbf{b} \in \mathbb{R}^n$,

$$\|\mathbf{A}\tilde{\mathbf{x}}^* - \mathbf{b}\|_2^2 \leq (1+\epsilon)\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$$

where $\tilde{\mathbf{x}}^* = \arg\min_{\mathbf{x}} \|\mathbf{\Pi}\mathbf{A}\mathbf{x} - \mathbf{\Pi}\mathbf{b}\|_2^2$.

Subspace embeddings have many other applications!

For example, if $m = O(k/\epsilon)$, $\mathbf{\Pi}\mathbf{A}$ can be used to compute an approximate partial SVD, which leads to a $(1+\epsilon)$ approximate low-rank approximation for $\mathbf{A}$.

Lemma (ϵ-net for the sphere)

For any $\epsilon \leq 1$, there exists a set $N_\epsilon \subset S_\mathcal{U}$ with $|N_\epsilon| = \left(\frac{4}{\epsilon}\right)^d$ such that $\forall \mathbf{v} \in S_\mathcal{U}$,

$$\min_{\mathbf{w} \in N_\epsilon} \|\mathbf{v} - \mathbf{w}\| \leq \epsilon.$$

Imaginary algorithm for constructing N_ϵ:

- Set $N_\epsilon = \{\}$
- While such a point exists, choose an arbitrary point $\mathbf{v} \in S_\mathcal{U}$ where $\nexists \mathbf{w} \in N_\epsilon$ with $\|\mathbf{v} - \mathbf{w}\| \leq \epsilon$. Set $N_\epsilon = N_\epsilon \cup \{\mathbf{w}\}$.

After running this procedure, we have $N_\epsilon = \{\mathbf{w}_1, \ldots, \mathbf{w}_{|N_\epsilon|}\}$ and $\min_{\mathbf{w} \in N_\epsilon} \|\mathbf{v} - \mathbf{w}\| \leq \epsilon$ for all $\mathbf{v} \in S_\mathcal{U}$ as desired.

How many steps does this procedure take?

Can place a ball of radius $\epsilon/2$ around each $\mathbf{w}_i$ without intersecting any other balls. All of these balls live in a ball of radius $1 + \epsilon/2$.

Volume of d dimensional ball of radius r is

$$\mathrm{vol}(d, r) = c \cdot r^d,$$

where c is a constant that depends on d, but not r. From previous slide we have:

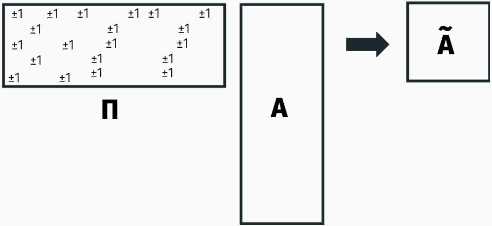
$$\begin{aligned} \mathrm{vol}(d, \epsilon/2) \cdot |N_\epsilon| &\leq \mathrm{vol}(d, 1 + \epsilon/2) \\ |N_\epsilon| &\leq \frac{\mathrm{vol}(d, 1 + \epsilon/2)}{\mathrm{vol}(d, \epsilon/2)} \\ &\leq \left(\frac{4}{\epsilon}\right)^d \end{aligned}$$

For $\epsilon, \delta = O(1)$, we need $\mathbf{\Pi}$ to have $m = O(d)$ rows.

- Cost to solve $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$:
  - $O(nd^2)$ time for direct method. Need to compute $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$.
  - $O(nd) \cdot (\text{\# of iterations})$ time for iterative method (GD, AGD, conjugate gradient method).
- Cost to solve $\|\mathbf{\Pi}\mathbf{A}\mathbf{x} - \mathbf{\Pi}\mathbf{b}\|_2^2$:
  - $O(d^3)$ time for direct method.
  - $O(d^2) \cdot (\text{\# of iterations})$ time for iterative method.

But time to compute $\mathbf{\Pi A}$ is an $(m \times n) \times (n \times d)$ matrix multiply: $O(mnd) = O(nd^2)$ time.

Goal: Develop faster Johnson-Lindenstrauss projections.



Typically using sparse and structured matrices.

Subsampled Randomized Hadamard Transform (SHRT) (Ailon-Chazelle, 2006):

Construct $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$ as follows:

$$\mathbf{\Pi} = \sqrt{\frac{n}{m}} \cdot \mathbf{SHD}, \text{ where}$$

- $\mathbf{S} \in \mathbb{R}^{m \times n}$ is a row subsampling matrix. Each row has a single 1 in a random column, all other entries 0.
- $\mathbf{D} \in n \times n$ is a diagonal matrix with each entry uniform ± 1.
- $\mathbf{H} \in n \times n$ is a Hadamard matrix.

Assume for now that n is a power of 2. For $i = 0, 1, \ldots,$ H_i is a Hadamard matrix with dimension $2^i \times 2^i$.

$$H_0 = 1 \quad H_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad H_2 = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

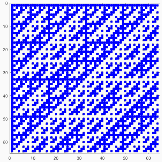
$$H_k = \frac{1}{\sqrt{2}} \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix}$$

How long does it take to compute $\mathbf{Hx}$ for a vector $\mathbf{x} \in \mathbb{R}^n$?

Property 1: Can compute $\mathbf{\Pi x} = \mathbf{SHDx}$ in $O(n \log n)$ time.

Compare to $O(nm)$ time for random Gaussian or ± 1 $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$.

RANDOMIZED HADAMARD TRANSFORM



Deterministic Hadamard matrix.

Randomized Hadamard **PHD**.

Fully random sign matrix.

Theorem (JL from SRHT)

Let $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$ *be a subsampled randomized Hadamard transform with* $m = O\left(\frac{\log(n/\delta)^2 \log(1/\delta)}{\epsilon^2}\right)$ *rows. Then for any fixed* $\mathbf{y}$,

$$(1-\epsilon)\|\mathbf{y}\|_2^2 \leq \|\mathbf{\Pi}\mathbf{y}\|_2^2 \leq (1+\epsilon)\|\mathbf{y}\|_2^2$$

with probability $(1-\delta)$.

Property 2: For any $k = 0, 1, \ldots$, we have $\mathbf{H}_k^T \mathbf{H}_k = \mathbf{I}$.

We want to show that $\|\sqrt{\frac{1}{m}}\mathbf{SHDy}\|_2^2 \approx \|\mathbf{y}\|_2^2$.

Let $\mathbf{z} \in \mathbb{R}^n = \mathbf{HDy}$.

- **Claim**: $\|\mathbf{z}\|_2^2 = \|\mathbf{y}\|_2^2$, exactly.
- $\|\mathbf{SHDy}\|_2^2 = \frac{n}{m}\|\mathbf{Sz}\|_2^2 =$ subsample of $\mathbf{z}$.
- $\mathbb{E}\left[\frac{n}{m}\|\mathbf{Sz}\|_2^2\right] = \|\mathbf{z}\|_2^2$.

What would z have to look like for $\|\mathbf{Sz}\|_2^2$ to look very different from $\|\mathbf{z}\|_2^2$ with high probability? I.e. when does subsampling fail. When does subsampling work?

Lemma (SHRT mixing lemma)

Let $\mathbf{H}$ *be an* $(n \times n)$ *Hadamard matrix and* $\mathbf{D}$ *a random* ± 1 *diagonal matrix. Let* $\mathbf{z} = \mathbf{HDy}$ *for some* $\mathbf{y} \in \mathbb{R}^n$*. With probability* $1 - \delta$,

$$|\mathbf{z}_i| \leq c \cdot \sqrt{\frac{\log(n/\delta)}{n}} \|\mathbf{y}\|_2$$

for some fixed constant c.

If all entries in $\mathbf{z}$ were uniform magnitude, we would have $|\mathbf{z}_i| = \frac{1}{\sqrt{n}} \|\mathbf{y}\|_2$. So we are very close to uniform with high probability.

SHRT mixing lemma proof:

Let $\mathbf{h}_i^T$ be the i^{th} row of $\mathbf{H}$. $\mathbf{z}_i = \mathbf{h}_i^T \mathbf{D} \mathbf{y}$ where:

$$\mathbf{h}_i^T \mathbf{D} = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} R_1 & & & \\ & R_2 & & \\ & & R_3 & \\ & & & R_4 \end{bmatrix}$$

where $R_1, \ldots, R_n$ are random ± 1's.

This is equivalent to

$$\mathbf{h}_i^T \mathbf{D} = \frac{1}{\sqrt{n}} \begin{bmatrix} R_1 & R_2 & R_3 & R_4 \end{bmatrix}.$$

SHRT mixing lemma proof:

So we have, for all i,

$$\mathbf{z}_i = \mathbf{h}_i^T \mathbf{D}\mathbf{y} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} R_i y_i.$$

- $\sqrt{n} \cdot \mathbf{z}_i$ is a random variable with mean 0 and variance $\|\mathbf{y}\|_2^2$, which is a sum of independent random variables.
- By Central Limit Theorem, we expect that:

$$\Pr[|\sqrt{n} \cdot \mathbf{z}_i| \geq t\|\mathbf{y}\|_2] \leq e^{-O(t^2)}.$$

- Setting t gives $\Pr\left[|\mathbf{z}_i| \geq O\left(\sqrt{\frac{\log(n/\delta)}{n}}\|\mathbf{y}\|_2\right)\right] \leq \frac{\delta}{n}$.
- Applying a union bound to all n entries of $\mathbf{z}$ gives the SHRT mixing lemma.

Formally, need to use Bernstein type concentration inequality to prove the bound:

Lemma (Rademacher Concentration)

Let $R_1, \ldots, R_n$ be Rademacher random variables (i.e. uniform ± 1's). Then for any vector $\mathbf{a} \in \mathbb{R}^n$,

$$\Pr\left[\sum_{i=1}^{n} R_i a_i \geq t\|\mathbf{a}\|_2\right] \leq e^{-t^2/2}.$$

With probability $1-\delta$, we have that all $\mathbf{z}_i \leq O\left(\sqrt{\frac{\log(n/\delta)}{n}}\|\mathbf{y}\|_2\right)$.

We want to analyze:

$$L = \|\sqrt{\frac{n}{m}}\mathbf{SHD}\|_2^2 = \frac{1}{m}\|\sqrt{\mathbf{n}}\mathbf{Sz}\|_2^2 = \frac{1}{m}\sum_{i=1}^{m}(\sqrt{n}\mathbf{z}_{j_i})^2$$

where j_i is a random index in $1, \ldots, n$.

We have that $\mathbb{E}L = \|\mathbf{z}\|_2^2 = \|\mathbf{y}\|_2^2$ and L is a sum of random variables, each bounded by $O(\log(n/\delta))$, which means they have bounded variance.

Apply a Chernoff/Hoeffding bound to get that $\left|L = \|\mathbf{y}\|_2^2\right| \leq \epsilon\|\mathbf{y}\|_2^2$ with probability $1-\delta$ as long as:

$$m \geq O\left(\frac{\log^2(n/\delta)\log(1/\delta)}{\epsilon^2}\right).$$

Theorem (JL from SRHT)

Let $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$ *be a subsampled randomized Hadamard transform with* $m = O\left(\frac{\log(n/\delta)^2 \log(1/\delta)}{\epsilon^2}\right)$ *rows. Then for any fixed* $\mathbf{y}$,

$$(1-\epsilon)\|\mathbf{y}\|_2^2 \leq \|\mathbf{\Pi y}\|_2^2 \leq (1+\epsilon)\|\mathbf{y}\|_2^2$$

with probability $(1-\delta)$.

Can be improved to $m = O\left(\frac{\log(n/\delta)\log(1/\delta)}{\epsilon^2}\right)$.

Upshot for regression: Compute $\mathbf{\Pi A}$ in $O(nd\log n)$ time instead of $O(nd^2)$ time. Compress problem down to $\tilde{\mathbf{A}}$ with $O(d^2)$ dimensions.

$O(nd \log n)$ is nearly linear in the size of $\mathbf{A}$ when $\mathbf{A}$ is dense.

Clarkson-Woodruff 2013, STOC Best Paper: Possible to compute $\tilde{\mathbf{A}}$ with poly(d) rows in:

$$O(\text{nnz}(\mathbf{A})) \text{ time.}$$

$\mathbf{\Pi}$ is chosen to be an ultra-sparse random matrix. Uses totally different techniques (you can't do JL + ϵ-net).

Lead to a whole close of matrix algorithms (for regression, SVD, etc.) which run in time:

$$O(\text{nnz}(\mathbf{A})) + \text{poly}(d, \epsilon).$$

WHAT WERE AILON AND CHAZELLE THINKING?

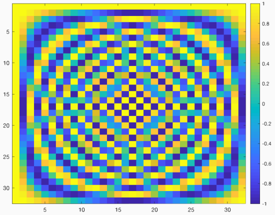
Simple, inspired algorithm that has been used for accelerating:

- Vector dimensionality reduction
- Linear algebra
- Locality sensitive hashing (SimHash)
- Randomized kernel learning methods (we will discuss after Thanksgiving)

```
m = 20;
c1 = (2*randi(2,1,n)-3).*y;
c2 = sqrt(n)*fwht(dy);
c3 = c2(randperm(n));
z = sqrt(n/m)*c3(1:m);
```

The Hadamard Transform is closely related to the Discrete Fourier Transform.

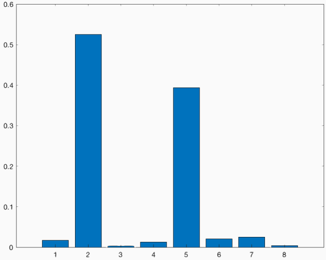
$$\mathsf{F}_{j,k} = e^{-2\pi i \frac{j \cdot k}{n}}, \qquad \mathsf{F}^*\mathsf{F} = \mathsf{I}.$$



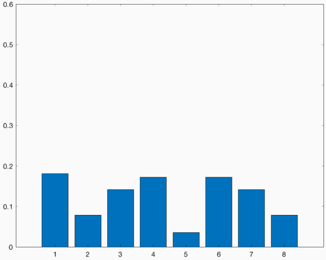
Real part of $\mathsf{F}_{j,k}$.

$\mathsf{F}\mathbf{y}$ computes the Fourier-transform of the vector $\mathbf{y}$. Can be computed in $O(n \log n)$ time using a divide and conquer algorithm (the Fast Fourier Transform).

The Uncertainty Principal (informal): A function and it's Fourier transform cannot both be concentrated.



Vector **y**.



Fourier transform **Fy**.

Sampling does not preserve norms, i.e. $\|\mathbf{S}\mathbf{y}\|_2 \not\approx \|\mathbf{y}\|_2$ when $\mathbf{y}$ has a few large entries.

Taking a Fourier transform exactly eliminates this hard case, without changing $\mathbf{y}$'s norm.