CS-GY 9223 I: Lecture 12

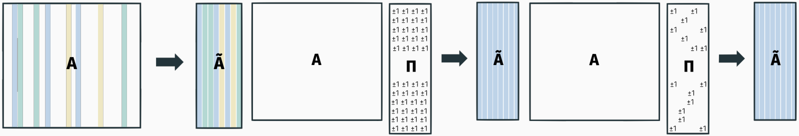
Randomized numerical linear algebra, fast Johnson-Lindenstrauss Transform

NYU Tandon School of Engineering, Prof. Christopher Musco

RANDOMIZED NUMERICAL LINEAR ALGEBRA

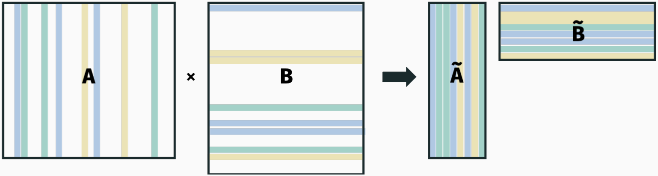
Main idea: If you want to compute singular vectors or eigenvectors, multiply two matrices, solve a regression problem, etc.:

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 $\mathbf{A}$.

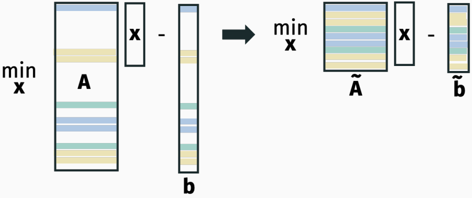


RANDOMIZED NUMERICAL LINEAR ALGEBRA

Approximate matrix multiplication:



Approximate regression:

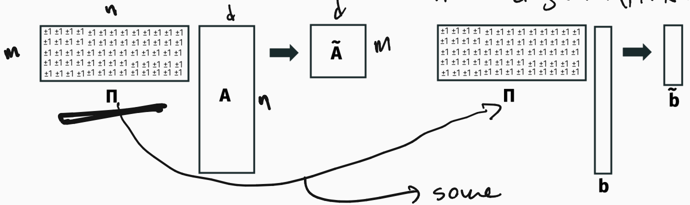


SKETCHED REGRESSION

Randomized approximate regression using a Johnson-Lindenstrauss Matrix:

$\underline{\|\tilde{x}^* - x^*\|_2^2}$

$x^* = \arg\min \|Ax - b\|_2^2$



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

$\underline{\|Ax - b\|_2^2}$

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

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

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

TARGET RESULT

Theorem (Randomized Linear Regression)

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

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

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

$m = O\left(\frac{d}{\varepsilon^2}\right)$

When $\varepsilon = O(1)$,

$m = O(d)$

SKETCHED REGRESSION

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

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

$$x^* = \arg\min_x \|Ax-b\|_2^2 .$$

Want to prove: $\|A\tilde{x}^*-b\|_2^2 \le (1+\epsilon)\|Ax^*-b\|_2^2 .$

$$\|A\tilde{x}^*-b\|_2^2 \le \frac{1}{1-\epsilon}\|\Pi A\tilde{x}^*-\Pi b\|_2^2 \le \frac{1}{1-\epsilon}\|\Pi A x^* - \Pi b\|_2^2$$

by fact that $\tilde{x}^*$ is opt for ΠA, Πb

$$\le \frac{(1+\epsilon)}{(1-\epsilon)}\|Ax^*-b\|_2^2 .$$

For small ϵ, $\frac{1+\epsilon}{1-\epsilon} = 1+O(\epsilon)$

$\epsilon' = \frac{\epsilon}{\text{constant}} \rightarrow$ gives result

DISTRIBUTIONAL JOHNSON-LINDENSTRAUSS REVIEW

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* $\mathbf{y}$,

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

with probability $(1-\delta)$.

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

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

$y = Ax - b$ $\quad \|y\|_2^2 \quad \|\Pi y\|_2^2$

FOR ANY TO FOR ALL

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.

false

Union bound: prove a statement with
prob. δ/# events

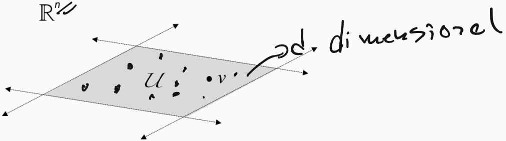
SUBSPACE EMBEDDINGS

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



---

[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.

SUBSPACE EMBEDDING TO APPROXIMATE REGRESSION

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

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

for all $\mathbf{x}$ and thus

$$m = O\left(\frac{d+1}{\epsilon^2}\right)$$

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

I.e., our main theorem is proven.

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

↳ n dimension vector, but always in $d+1$ dimensional subspace of $\mathbb{R}^n$

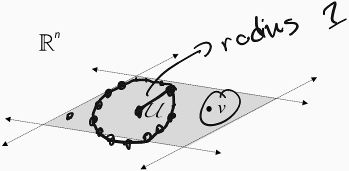
SUBSPACE EMBEDDINGS

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



SUBSPACE EMBEDDING PROOF

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

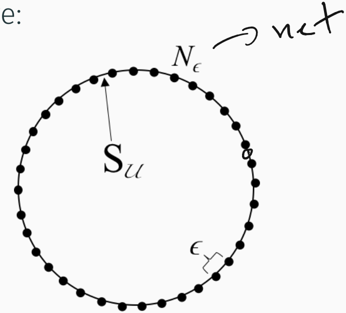
$$\boxed{S_{\mathcal{U}}} = \{\mathbf{w} \mid \mathbf{w} \in \mathcal{U} \text{ and } \|\mathbf{w}\|_2 = 1\}.$$

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

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

SUBSPACE EMBEDDING PROOF

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



N_ϵ is called an "ϵ"-net.

If we can prove

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

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

ϵ-NET FOR THE SPHERE

Lemma (ϵ-net for the sphere)

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

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

SUBSPACE EMBEDDING PROOF

1. Preserving norms of all points in net N_ϵ.

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

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

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

$$\log(1/\delta') = \log\left(\frac{1}{(\varepsilon/4)^d \cdot \delta}\right) = \log(1/\delta) + \log\left(\left(\frac{4}{\varepsilon}\right)^d\right)$$

$$= \underline{\log(1/\delta) + d\log 4/\varepsilon}$$

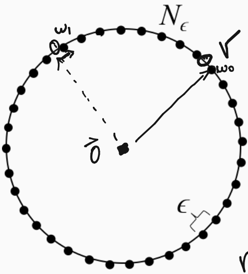
SUBSPACE EMBEDDING PROOF

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:

$$\mathbf{v} = \underline{\overset{\to 1}{\mathbf{w}_0} + \overset{\le \epsilon}{c_1}\mathbf{w}_1 + \overset{\le \epsilon^2}{c_2}\mathbf{w}_2 + \overset{\le \epsilon^3}{\ldots}} \cdot \cdot \cdot \cdot \cdot$$

for constants $c_1, c_2, \ldots$ where $|c_i| \le \underline{\underline{\epsilon^i}}$.



$$v = w_0 + \overset{\to \text{remainder}}{r_0}$$

where $\|r_0\|_2 \le \epsilon$

$$\frac{r_0}{\|r_0\|_2} = w_1 + r_1$$

$$r_0 = \underbrace{\|r_0\|_2}_{\le \epsilon} w_1 + \underbrace{\|r_0\|_2 r_1}_{\text{norm} \le \epsilon^2}$$

SUBSPACE EMBEDDING PROOF

$(1-\varepsilon) \le \|\Pi w_0\|_2 \le (1+\varepsilon)$

for $\underline{\underline{\text{all}}}$ $w_0 \in N_\varepsilon$

3. Preserving norm of v.

Applying triangle inequality, we have

$$
\begin{aligned}
\|\mathbf{\Pi v}\|_2 &= \|\mathbf{\Pi w}_0 + c_1\mathbf{\Pi w}_1 + c_2\mathbf{\Pi w}_2 + \dots\| \\
&\le \underline{\|\mathbf{\Pi w}_0\|} + \epsilon\underline{\|\mathbf{\Pi w}_1\|} + \epsilon^2\underline{\|\mathbf{\Pi w}_2\|} + \dots \\
&\le (1+\epsilon) + \underbrace{\boxed{\epsilon(1+\epsilon)}}_{\le 2\varepsilon} + \underbrace{\epsilon^2(1+\epsilon)}_{\le 2\varepsilon^2} + \dots \\
&\le \underline{\underline{1 + O(\epsilon)}}.
\end{aligned}
$$

$\|\Pi w_0 + \underline{\underline{\Pi r}}\|_2$

$\le \|\Pi w_0\|_2$

$+ \underbrace{\|\Pi r\|_2}$

huge?

even if

$\|r\|_2 \le \varepsilon$

$\|\Pi v\|_2 \le (1+O(\varepsilon))\|v\|_2$

SUBSPACE EMBEDDING PROOF

3. Preserving norm of v.

Similarly,

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

SUBSPACE EMBEDDING PROOF

So we have proven

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

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

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

Adjusting ϵ proves the Subspace Embedding theorem.

SUBSPACE EMBEDDINGS

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

FINAL RESULT

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 $\mathbf{A}$.

ϵ-NET FOR THE SPHERE

Lemma (ϵ-net for the sphere)

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

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

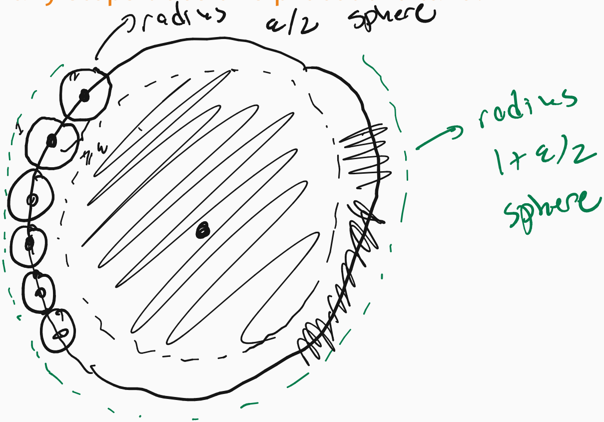
Imaginary algorithm for constructing N_ϵ:

- 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.

ϵ-NET FOR THE SPHERE

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 $1 + \epsilon/2$.

ϵ-NET FOR THE SPHERE

Volume of d dimensional ball of radius r is

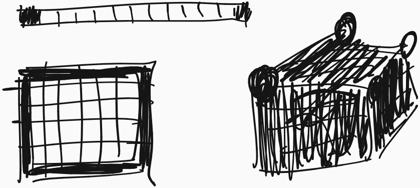
$$\mathrm{vol}(d, r) = c \cdot r^d,$$

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

$$\mathrm{vol}(d, \epsilon/2) \cdot |N_\epsilon| \leq \mathrm{vol}(d, 1 + \epsilon/2)$$

$$|N_\epsilon| \leq \frac{\mathrm{vol}(d, 1 + \epsilon/2)}{\mathrm{vol}(d, \epsilon/2)} = \frac{c \cdot \overset{\leq 2}{(1+\epsilon/2)}^d}{c \cdot (\epsilon/2)^d}$$

$$\leq \left(\frac{4}{\epsilon}\right)^d \qquad \leq \left(\frac{4}{\epsilon}\right)^d$$



RUNTIME CONSIDERATION

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 $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$.

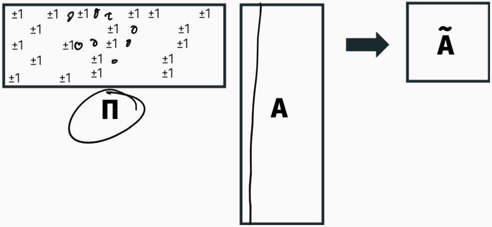
      $A^TA \to O(nd^2)$
    - $O(nd) \cdot (\text{\# of iterations})$ time for iterative method (GD, AGD, conjugate gradient method).

      $\underline{2A^TAx - 2A^Tb} \to O(nd)$ time
- 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.

RUNTIME CONSIDERATION

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

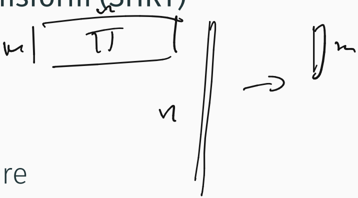
Goal: Develop faster Johnson-Lindenstrauss projections.



Typically using $\underbrace{\text{sparse}}$ and $\underbrace{\text{structured}}$ matrices.

THE FAST JOHNSON-LINDENSTRAUSS TRANSFORM

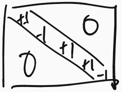
Subsampled Randomized Hadamard Transform (SHRT) (Ailon-Chazelle, 2006):

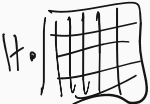


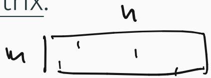
Construct $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$ as follows:

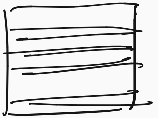
$$\mathbf{\Pi} = \underline{\sqrt{\frac{n}{m}}} \cdot \underline{\hat{\mathbf{S}}\mathbf{H}\mathbf{D}}, \text{ where}$$

- $\mathbf{S} \in \mathbb{R}^{m \times n}$ 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.
- $\mathbf{H} \in \boxed{n \times n}$ is a Hadamard matrix.









HADAMARD MATRICES

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 **Hx** for a vector $\mathbf{x} \in \mathbb{R}^n$?

$\Pi y = S\underbrace{HD}y$

HADAMARD MATRICES

n time   O(n log n) time

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

[illegible]

n

$Dx = y$

$Hy \rightarrow H_k$ where $k = \log_2(n)$

$$\underline{H_k y} = \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix} \begin{bmatrix} y_1 \\ \cdots \\ y_2 \end{bmatrix} = \begin{bmatrix} H_{k-1} y_1 + H_{k-1} y_2 \\ \cdots \\ H_{k-1} y_1 - H_{k-1} y_2 \end{bmatrix}$$

compute: $H_{k-1} y_1$, $H_{k-1} y_2$

$$\boxed{T_k = 2T_{k-1} + n}$$

$T_n : c n \log_2 n$

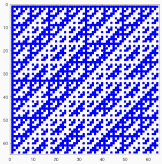
Compare to $\boxed{O(nm)}$ time for random Gaussian or ±1 $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$.

$T_{k-1} = \frac{n}{2} \log_2(n/2)$

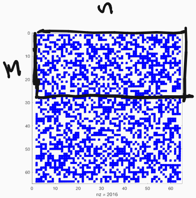
$T_n = n \log_2(n/2) + n \rightarrow n \log_2(2)$

$= n \log_2(n)$

RANDOMIZED HADAMARD TRANSFORM



Deterministic Hadamard matrix.



Randomized Hadamard **PHD**.

Fully random sign matrix.

JOHNSON-LINDENSTRAUSS WITH SHRTs

Theorem (JL from SRHT)

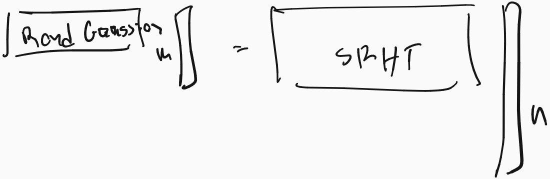
Let $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$ *be a subsampled randomized Hadamard transform with* $m = O\left(\frac{\underline{\log(n/\delta)^2} \log(1/\delta)}{\epsilon^2}\right)$ *rows. Then for any fixed* $\mathbf{y}$*,*

↓ small cost in embedding dimension

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

with probability $(1-\delta)$.

$$\underline{\underline{m = O\left(\frac{\log 1/\delta}{\epsilon^2}\right)}}$$



HADAMARD MATRICES ARE ORTHOGONAL

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

$$\frac{1}{\sqrt{2}} \begin{bmatrix} H_{n-1}^T & H_{n-1}^T \\ H_{n-1}^T & -H_{n-1}^T \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} H_{n-1}^T H_{n-1} + H_{n-1}^T H_{n-1} & \overbrace{H_{n-1}^T H_{n-1} - H_{n-1}^T H_{n-1}}^{0} \\ \underbrace{H_{n-1}^T H_{k-1} - H_{n-1}^T H_{n-1}}_{0} & H_{n-1}^T H_{n-1} + H_{n-1}^T H_{n-1} \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} 2I & 0 \\ 0 & 2I \end{bmatrix} = \boxed{I}$$

RANDOMIZED HADAMARD ANALYSIS

We want to show that $\|\underbrace{\sqrt{\tfrac{n}{m}}\mathbf{SH}}_{\Pi}\mathbf{Dy}\|_2^2 \approx \|\mathbf{y}\|_2^2$.

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

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

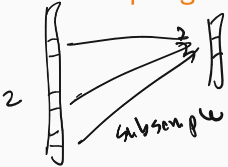
$\|Dy\|_2^2 = y^T D D y = \|y\|_2^2$

$\|HDy\|_2^2$

$y^T D \underbrace{H^T H}_{I} D y = \|Dy\|_2^2$

$= \|y\|_2^2$

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?



$\frac{n}{m} Sz$

$\frac{n}{m}\|Sz\|_2^2 \approx \|z\|_2^2$

$= \|\Pi y\|_2^2$ $\quad = \|y\|_2^2$

RANDOMIZED HADAMARD ANALYSIS

Lemma (SHRT mixing lemma)

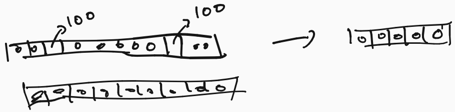
Let $\underline{\mathbf{H}}$ *be an* $(n \times n)$ *Hadamard matrix and* $\underline{\mathbf{D}}$ *a random* ± 1 *diagonal matrix. Let* $\underline{\underline{\mathbf{z}}} = \underline{\mathbf{HD}}\mathbf{y}$ *for some* $\mathbf{y} \in \mathbb{R}^n$*. With probability* $1 - \delta$*,*

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

for some *<u>fixed constant</u>* *c.*

$\|z\|_2^2 = \|y\|_2^2$

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.**



$z_i^2 = \frac{\|z\|_2^2}{n} = \frac{\|y\|_2^2}{n}$

$\sum z_i^2 = \|z\|_2^2 = \|y\|_2^2$

$|z_i| = \frac{1}{\sqrt{n}}\|y\|_2$

RANDOMIZED HADAMARD ANALYSIS

SHRT mixing lemma proof:

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

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

where $R_1, \ldots, \underline{R_n}$ are random ±1's. → "Rademacher random variable"

This is equivalent to

$$\mathbf{h}_i^T \mathbf{D} = \frac{1}{\sqrt{n}} \underline{\begin{bmatrix} R_1 & R_2 & R_3 & R_4 \end{bmatrix}} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

RANDOMIZED HADAMARD ANALYSIS

SHRT mixing lemma proof:

$$\mathbb{E}[\sqrt{n}\, z_i] = \mathbb{E} \sum B_i y_i = \sum \underbrace{\mathbb{E}[B_i]}_{=0} y_i$$

So we have, for all i,

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

$$\text{Var}[\sqrt{n}\, z_i] = \sum \text{Var}[B_i y_i] = \sum y_i^2 \text{Var}[B_i] = \sum y_i^2 = \|y\|_2^2$$

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

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

$$e^{-\sqrt{\log(n/\delta)}^2} = \frac{\delta}{n}$$

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

holds $\underline{\text{for all } i}$ w/ prob. $1-\delta$

RADEMACHER CONCENTRATION

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 $\mathbf{a} \in \mathbb{R}^n$,

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

y from previous page

FINISHING UP

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: want $L \approx \|y\|_2^2 = \|z\|_2^2$

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

random index

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.

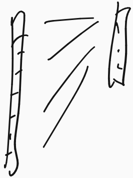
$|z_{j_i}| \leq \sqrt{\log(n/\delta)} \cdot \|y\|_2$

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:

$\downarrow$

$\mathbb{E}[L]$

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



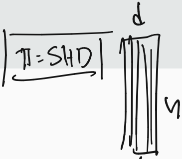
JOHNSON-LINDENSTRAUSS WITH SHRTs

Theorem (JL from SRHT)

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

$$\delta = O(\varepsilon^d)$$

$$\log(n/\varepsilon^d)\cdot\log(1/\varepsilon^d)$$

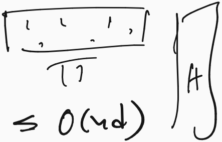
$$\textcircled{d}\log(n/\varepsilon)\,\textcircled{d}\log(1/\varepsilon)$$

BRIEF COMMENT ON OTHER METHODS

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

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

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



$\leq O(nd)$

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

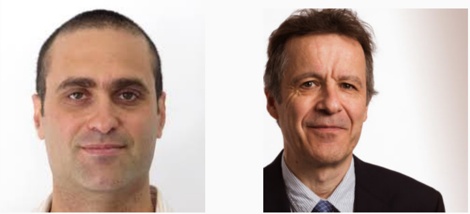
Lead to a whole close of matrix algorithms (for regression, SVD, etc.) which run in time:

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

$\text{nnz}(A) \cdot (\#\text{ iterations})$

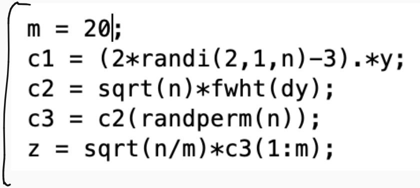
$\rightarrow f(\kappa, \epsilon)$

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)



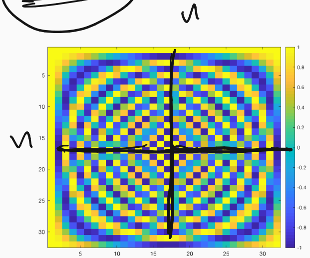
```
m = 20;
c1 = (2*randi(2,1,n)-3).*y;
c2 = sqrt(n)*fwht(dy);
c3 = c2(randperm(n));
z = sqrt(n/m)*c3(1:m);
```

WHAT WERE AILON AND CHAZELLE THINKING?

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

$i = \sqrt{-1}$

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



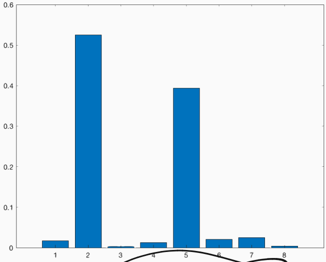
$$\left| e^{-2\pi i \frac{jk}{n}} \right| = 1$$

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

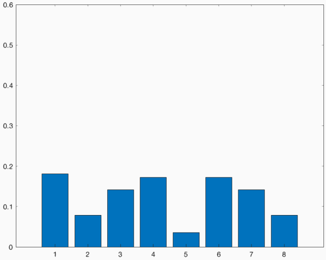
$\mathbf{F}\mathbf{y}$ computes the Fourier-transform of the vector $\mathbf{y}$. Can be computed in $O(n \log n)$ time using a divide and conquer algorithm (the Fast Fourier Transform).

THE UNCERTAINTY PRINCIPAL

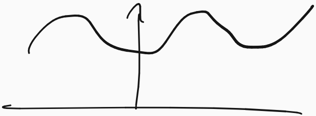
The Uncertainty Principal (informal): A function and it's Fourier transform cannot both be concentrated.

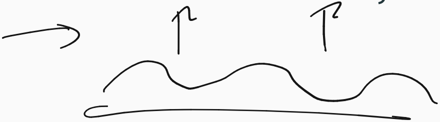


Vector **y**.



Fourier transform **Fy**.





THE UNCERTAINTY PRINCIPAL

$$SFy$$

$$\rightarrow \boxed{SHy} \rightarrow SHDy$$

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.