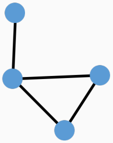
CS-GY 6763: Lecture 11
Randomized numerical linear algebra, ϵ-net arguments.

NYU Tandon School of Engineering, Prof. Christopher Musco

LAST CLASS

Represent undirected graph as symmetric matrix: $n \times n$ adjacency matrix $\mathbf{A}$ and graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ is the diagonal degree matrix.



$$\overset{\mathbf{D}}{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}} - \overset{\mathbf{A}}{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}} = \overset{\mathbf{L}}{\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}}$$

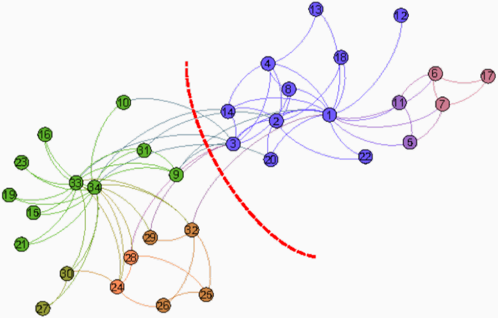
$\mathbf{L} = \mathbf{B}^T\mathbf{B}$ where B is the "edge-vertex incidence" matrix.

$$\mathbf{B} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

LAST CLASS

Balanced Cut: Partition nodes along a cut that:

- Has few crossing edges: $|\{(u, v) \in E : u \in B, v \in C\}|$ is small.
- Separates large partitions: $|B|, |C|$ are not too small.



(a) Zachary Karate Club Graph

We observed that $\mathbf{x}^T L \mathbf{x} = \sum_{(i,j)\in E} (\mathbf{x}(i) - \mathbf{x}(j))^2$. If $\mathbf{c}$ is a "cut indicator vector" for a cut between node set B and C – i.e. $\mathbf{c}[i] = 1$ for all $i \in B$ and -1 elsewhere, then it followed that:

$$\mathbf{c}^T \mathbf{L} \mathbf{c} = 4 \cdot cut(B, C).$$

We used this basic fact to argue heuristically that the smallest eigenvectors of $\mathbf{L}$ can be used to find balanced cuts in a graph.

Note: $\mathbf{c}$ often denote by $\chi_{B,C}$.

"Relax and round" algorithm:

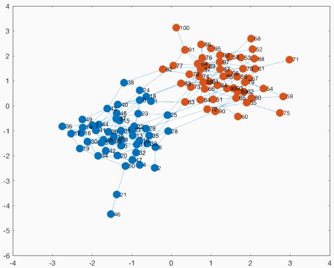
- Relax problem $\min \mathbf{c}^T\mathbf{L}\mathbf{c}$ by not requiring $\mathbf{c}$ to be a binary cut-indicator vector.
- Showed that second smallest eigenvector $\mathbf{v}_{n-1}$ of $\mathbf{L}$ solved the relaxed problem.
- Round this vector to be a cut indicator vector: all negative entries rounded to -1, all positive entries rounded to 1.

Main theoretical result: This approach is hard to analyze in general, but can be proven to work well on random graphs drawn from the stochastic block model!.

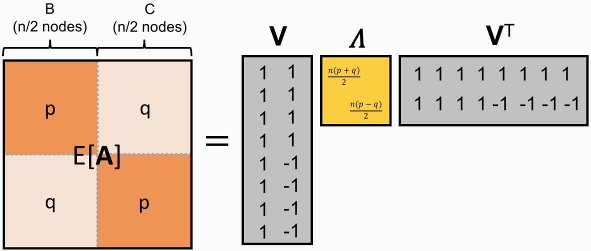
Stochastic Block Model (Planted Partition Model):

Let $G_n(p, q)$ be a distribution over graphs on n nodes, split equally into two groups B and C, each with $n/2$ nodes.

- Any two nodes in the same group are connected with probability p (including self-loops).
- Any two nodes in different groups are connected with prob. $q < p$.



$\mathbb{E}[\mathbf{A}] = p \cdot \mathbf{I} - \mathbb{E}[\mathbf{L}]$, so smallest eigenvectors of $\mathbb{E}[\mathbf{L}]$ are equal to largest of $\mathbb{E}[\mathbf{A}]$.



- $\mathbf{v}_1 = \mathbf{1}$ with eigenvalue $\lambda_1 = \frac{(p+q)n}{2}$.
- $\mathbf{v}_2 = \boldsymbol{\chi}_{B,C}$ with eigenvalue $\lambda_2 = \frac{(p-q)n}{2}$.
- $\chi_{B,C}(i) = 1$ if $i \in B$ and $\chi_{B,C}(i) = -1$ for $i \in C$.

If we compute $\mathbf{v}_2$ then we recover the communities B and C.

Upshot: The second small eigenvector of $\mathbb{E}[\mathbf{L}]$ (i.e. the second largest of $\mathbb{E}[\mathbf{A}]$) is $\chi_{B,C}$ – the indicator vector for the cut between the communities.

- If the random graph G (equivilantly $\mathbf{A}$ and $\mathbf{L}$) were exactly equal to its expectation, partitioning using this eigenvector would exactly recover communities B and C.

How do we show that a matrix (e.g., $\mathbf{A}$) is close to its expectation? **Matrix concentration inequalities.**

Matrix Concentration Inequality: If $p \geq O\left(\frac{\log^4 n}{n}\right)$, then with high probability

$$\|\mathbf{A} - \mathbb{E}[\mathbf{A}]\|_2 \leq O(\sqrt{pn}).$$

where $\|\cdot\|_2$ is the matrix spectral norm (operator norm).

For $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\|\mathbf{X}\|_2 = \max_{z \in \mathbb{R}^d : \|z\|_2 = 1} \|\mathbf{X}z\|_2 = \sigma_1(\mathbf{X})$.

For the stochastic block model application, we want to show that the second eigenvectors of $\mathbf{A}$ and $\mathbb{E}[\mathbf{A}]$ are close. How does this relate to their difference in spectral norm?

Davis-Kahan Eigenvector Perturbation Theorem: Suppose $\mathbf{A}, \overline{\mathbf{A}} \in \mathbb{R}^{d \times d}$ are symmetric with $\|\mathbf{A} - \overline{\mathbf{A}}\|_2 \leq \epsilon$ and eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ and $\bar{\mathbf{v}}_1, \bar{\mathbf{v}}_2, \ldots, \bar{\mathbf{v}}_n$. Letting $\theta(\mathbf{v}_i, \bar{\mathbf{v}}_i)$ denote the angle between $\mathbf{v}_i$ and $\bar{\mathbf{v}}_i$, for all i:

$$\sin[\theta(\mathbf{v}_i, \bar{\mathbf{v}}_i)] \leq \frac{\epsilon}{\min_{j \neq i} |\lambda_i - \lambda_j|}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\overline{\mathbf{A}}$.

Claim 1 (Matrix Concentration): For $p \geq O\left(\frac{\log^4 n}{n}\right)$,

$$\|\mathbf{A} - \mathbb{E}[\mathbf{A}]\|_2 \leq O(\sqrt{pn}).$$

Claim 2 (Davis-Kahan): For $p \geq O\left(\frac{\log^4 n}{n}\right)$,

$$\sin\theta(\mathbf{v}_2, \bar{\mathbf{v}}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq i} |\lambda_i - \lambda_j|} \leq \frac{O(\sqrt{pn})}{(p-q)n/2} = O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)$$

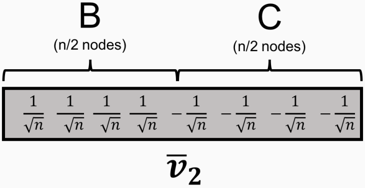
Recall: $\mathbb{E}[\mathbf{A}]$, has eigenvalues $\lambda_1 = \frac{(p+q)n}{2}$, $\lambda_2 = \frac{(p-q)n}{2}$, $\lambda_i = 0$ for $i \geq 3$.

$$\min_{j \neq i} |\lambda_i - \lambda_j| = \min\left(qn, \frac{(p-q)n}{2}\right).$$

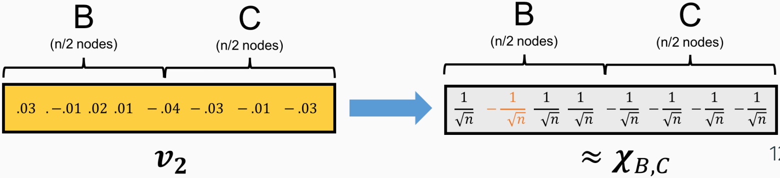
Assume $\frac{(p-q)n}{2}$ will be the minimum of these two gaps.

So far: $\sin\theta(\mathbf{v}_2, \bar{\mathbf{v}}_2) \leq O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)$. What does this give us?

- Can show that this implies $\|\mathbf{v}_2 - \bar{\mathbf{v}}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$ (exercise).
- $\bar{\mathbf{v}}_2$ is $\frac{1}{\sqrt{n}}\chi_{B,C}$: the community indicator vector.



- To understand how well rounding recovers $\bar{\mathbf{v}}_2$, need to understand how many locations $\mathbf{v}_2$ and $\bar{\mathbf{v}}_2$ can differ in sign.



Main argument:

- Every i where $v_2(i), \bar{v}_2(i)$ differ in sign contributes $\geq \frac{1}{n}$ to $\|v_2 - \bar{v}_2\|_2^2$.
- We know that $\|\mathbf{v}_2 - \bar{\mathbf{v}}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$.
- So $\mathbf{v}_2$ and $\bar{\mathbf{v}}_2$ differ in sign in at most $O\left(\frac{p}{(p-q)^2}\right)$ positions.

Upshot: If G is a stochastic block model graph with adjacency matrix $\mathbf{A}$, if we compute its second large eigenvector v_2 and assign nodes to communities according to the sign pattern of this vector, we will correctly assign all but $O\left(\frac{p}{(p-q)^2}\right)$ nodes.

- **Hard case:** $p = c/n$ for some factor c. Even when $p - q = O(1/n)$, assign all but an $O(n)$ fraction of nodes correctly. E.g., assign 99% of nodes to the right cluster.

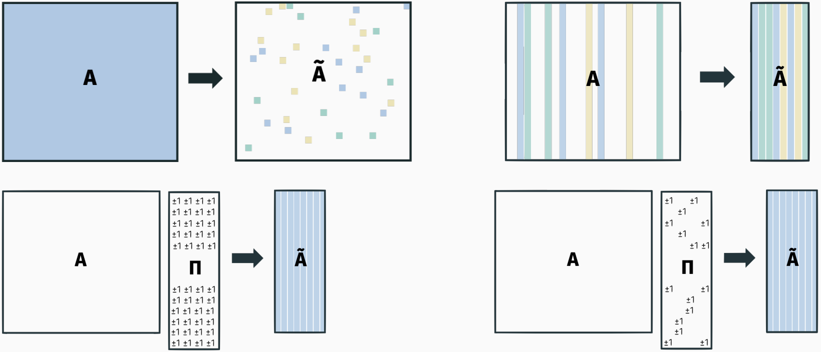
Forget about the previous problem, but still consider the matrix $\mathbf{M} = \mathbb{E}[\mathbf{A}]$.

- Dense $n \times n$ matrix.
- Computing top eigenvectors takes $\approx O(n^2/\sqrt{\epsilon})$ time.

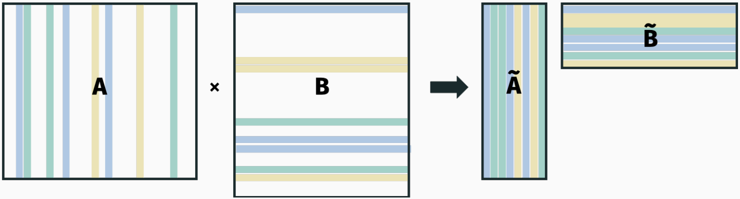
If someone asked you to speed this up and return approximate top eigenvectors, what could you do?

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

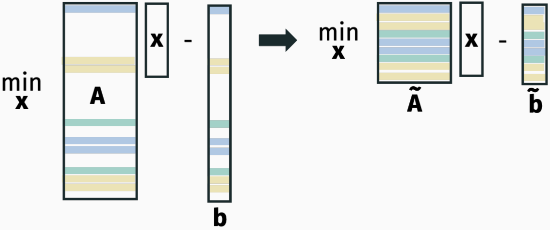
1. Compress your matrices using a randomized method (e.g. subsampling).
2. Solve the problem on the smaller or sparser matrix.
   - $\tilde{\mathbf{A}}$ called a "sketch" or "coreset" for $\mathbf{A}$.



Approximate matrix multiplication:

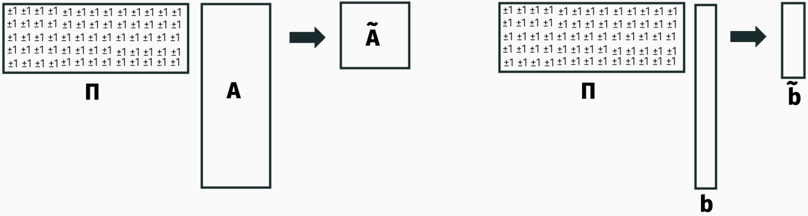


Approximate regression:



SKETCHED REGRESSION

Randomized approximate regression using a Johnson-Lindenstrauss Matrix:



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

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

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

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* $m = O\left(\frac{d}{\epsilon^2}\right)$ *rows*[1]*. Then with probability* 9/10*, for any* $\mathbf{A} \in \mathbb{R}^{n \times d}$ *and* $\mathbf{b} \in \mathbb{R}^n$*,*

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

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

[1] This can be improved to $O(d/\epsilon)$ with a tighter analysis

PLAN

- Prove this theorem using an ϵ-net argument, which is a popular technique for applying our standard concentration inequality + union bound argument to an infinite number of events.
- These sort of arguments appear all the time in theoretical algorithms and ML research, so this lecture is as much about the technique as the final result.
- You will use and ϵ-net argument to prove a matrix concentration inequality on your problem set.

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 can't be tackled directly with a union bound argument.

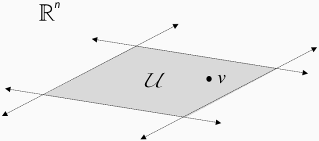
Note that all vectors of the form $(\mathbf{A}\mathbf{x} - \mathbf{b})$ lie in a low dimensional subspace: spanned by $d+1$ vectors, where d is the width of $\mathbf{A}$. **So even though the set is infinite, it is "simple" in some way. Parameterized by just $d+1$ numbers.**

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 $m = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$[2].



[2] 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,

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

for all $\mathbf{x}$ and thus

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

I.e., our main theorem is proven.

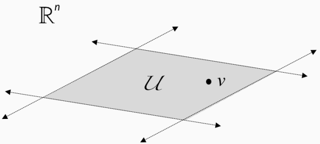
Proof: Apply Subspace Embedding Thm. to the $(d+1)$ dimensional subspace spanned by $\mathbf{A}$'s d columns and $\mathbf{b}$. Every vector $\mathbf{A}\mathbf{x}-\mathbf{b}$ lies in this subspace.

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



Subspace embeddings have tons of other applications!

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

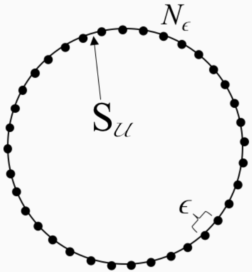
First Observation: The theorem holds as long as (2) holds for all $\mathbf{w}$ on the unit sphere in $\mathcal{U}$. Denote the sphere $S_{\mathcal{U}}$:

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

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

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

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 $\forall \mathbf{v} \in S_{\mathcal{U}}$,

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

Take this claim to be true for now: we will prove later.

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 $\mathbf{w} \in N_\epsilon$,

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

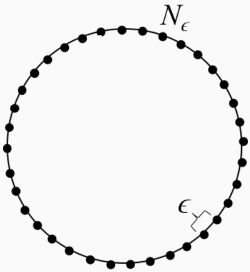
as long as Π has $O\left(\frac{\log(1/\delta')}{\epsilon^2}\right) = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon^2}\right)$ rows.

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

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

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

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

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

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

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 $1 - \delta$,

$$(1-\epsilon)\|\mathbf{v}\|_2^2 \leq \|\mathbf{\Pi}\mathbf{v}\|_2^2 \leq (1+\epsilon)\|\mathbf{v}\|_2^2 \tag{3}$$

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

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 $\forall \mathbf{v} \in S_{\mathcal{U}}$,

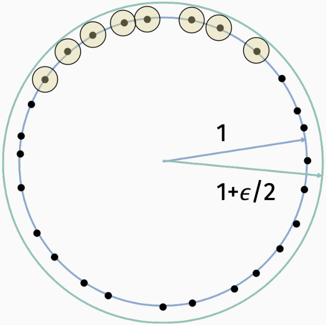
$$\min_{\mathbf{w} \in N_\epsilon} \|\mathbf{v} - \mathbf{w}\| \leq \epsilon.$$

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 $1 + \epsilon/2$.

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:

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

You can actually show that $m = O\left(\frac{d+\log(1/\delta)}{\epsilon}\right)$ suffices to be a d dimensional subspace embedding, instead of the bound we proved of $m = O\left(\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon}\right)$.

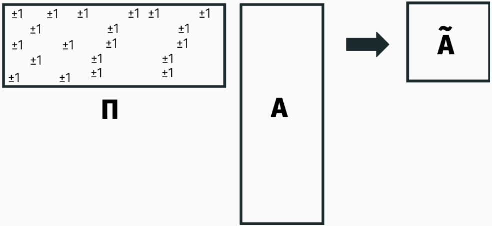
The trick is to show that a constant factor net is actually all that you need instead of an ϵ factor.

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 $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$.
  - $O(nd) \cdot$ (# 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.
  - $O(d^2) \cdot$ (# 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.

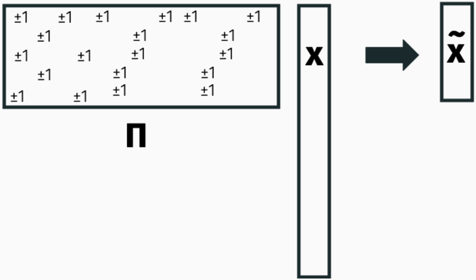


Typically using sparse and structured matrices.

We will describe a construction where $\mathbf{\Pi A}$ can be computed in $O(nd \log n)$ time.

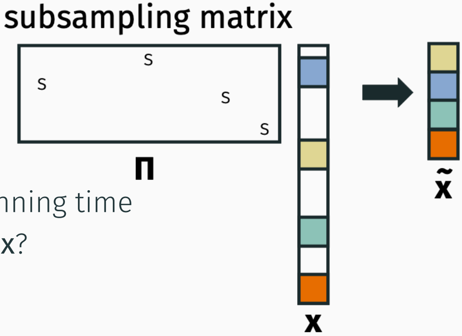
Goal: Develop methods that reduce a vector $\mathbf{x} \in \mathbb{R}^n$ down to $m \approx \frac{\log(1/\delta)}{\epsilon^2}$ dimensions in $o(mn)$ time and guarantee:

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



We will learn about a truly brilliant method that runs in $O(n \log n)$ time. **Preview:** Will involve Fast Fourier Transform in disguise.

Let $\mathbf{\Pi}$ be a random sampling matrix. Every row contains a value of $s = \sqrt{n/m}$ in a single location, and is zero elsewhere.

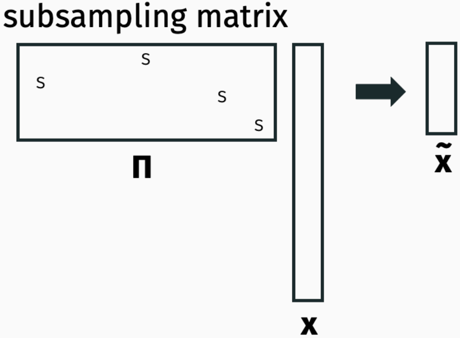


What's the running time to compute $\mathbf{\Pi}\mathbf{x}$?

$\|\mathbf{\Pi}\mathbf{x}\|_2^2 =$

$\mathbb{E}[\|\mathbf{\Pi}\mathbf{x}\|_2^2] =$

So $\mathbb{E}\|\mathbf{\Pi x}\|_2^2 = \|\mathbf{x}\|_2^2$ in expectation. To show it is close with high probability we would need to apply a concentration inequality. How do you think this will work out?



$\|\mathbf{\Pi x}\|_2^2 =$

$\sigma^2 = \text{Var}[\|\mathbf{\Pi x}\|_2^2] =$

Recall Chebyshev's Inequality:

$$\Pr[\left|\|\mathbf{\Pi x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \leq 10 \cdot \sigma] \leq \frac{1}{100}$$

We want additive error $\left|\|\mathbf{\Pi x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \leq \epsilon\|\mathbf{x}\|_2^2$

We need to choose m so that:

$$10\sqrt{\frac{n}{m}}\|\mathbf{x}\|_4^2 \leq \epsilon\|\mathbf{x}\|_2^2.$$

How do these two two norms compare?

$$\|\mathbf{x}\|_4^2 = \left(\sum_{i=1}^{n} x_i^4\right)^{1/2} \qquad\qquad \|\mathbf{x}\|_2^2 = \sum_{i=1}^{n} x_i^2$$

Consider 2 extreme cases:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad\qquad \mathbf{x} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

We need to choose m so that:

$$\frac{1}{10}\sqrt{\frac{n}{m}}\|\mathbf{x}\|_4^2 \leq \epsilon\|\mathbf{x}\|_2^2.$$

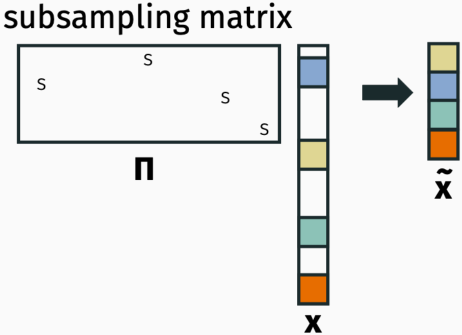
Suppose $\mathbf{x}$ is very evenly distributed. I.e., for all $i \in 1, \ldots, n$,

$$x_i^2 \leq \frac{c}{n}\sum_{i=1}^{n} x_i^2 = \frac{c}{n}\|\mathbf{x}\|_2^2$$

Claim: $\|\mathbf{x}\|_4^2 \leq \frac{c}{\sqrt{n}}\|\mathbf{x}\|_2^2$. So $m = O(c/\epsilon^2)$ samples suffices.[3]

[3] Using the right Bernstein bound we can prove $m = O(c\log(1/\delta)/\epsilon^2)$ suffices for failure probability δ.

So sampling does work to preserve the norm of $\mathbf{x}$, but only when the vector is relatively "smooth" (not concentrated). Do we expect to see such vectors in the wild?



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

Key idea: First multiply $\mathbf{x}$ by a "mixing matrix" $\mathbf{M}$ which ensures it cannot be too concentrated in one place.

$\mathbf{M}$ should have the property that $\|\mathbf{M}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$ exactly, or is very close. Then we will multiply by a subsampling matrix $\mathbf{S}$ to do the actual dimensionality reduction:

$$\mathbf{\Pi}\mathbf{x} = \mathbf{S}\mathbf{M}\mathbf{x}$$

Oh… and $\mathbf{M}$ needs to be fast to multiply by!

Good mixing matrices should look random:

$$\underbrace{\begin{bmatrix} +1 & -1 & +1 & +1 & +1 & -1 & +1 & -1 \\ -1 & -1 & -1 & +1 & +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & +1 & +1 & -1 & -1 & -1 \\ +1 & +1 & +1 & +1 & -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 & -1 & +1 & +1 & -1 \\ -1 & +1 & -1 & -1 & -1 & +1 & -1 & -1 \\ -1 & +1 & -1 & +1 & -1 & -1 & -1 & +1 \end{bmatrix}}_{\mathbf{M}} \quad \mathbf{x}$$

For this approach to work, we need to be able to compute $\mathbf{Mx}$ very quickly. So we will use a pseudorandom matrix instead.

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

$\mathbf{\Pi} = \mathbf{S}\mathbf{M}$ where $\mathbf{M} = \mathbf{H}\mathbf{D}$:

- $\mathbf{D} \in n \times n$ is a diagonal matrix with each entry uniform ± 1.
- $\mathbf{H} \in n \times n$ is a Hadamard matrix.

The Hadarmard matrix is an othogonal matrix closely related to the discrete Fourier matrix. It has two critical properties:

1. $\|\mathbf{H}\mathbf{v}\|_2^2 = \|\mathbf{v}\|_2^2$ exactly. Thus $\|\mathbf{H}\mathbf{D}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$
2. $\|\mathbf{H}\mathbf{v}\|_2^2$ can be computed in $O(n \log n)$ time.

Assume that n **is a power of** 2. For $k = 0, 1, \ldots,$ the k^{th} Hadamard matrix $\mathbf{H}_k$ is a $2^k \times 2^k$ matrix defined by:

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

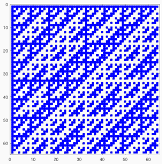
$$H_k = \frac{1}{\sqrt{2}} \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix}$$

The $n \times n$ Hadamard matrix has all entries as $\pm\frac{1}{\sqrt{n}}$.

Property 1: For any $k = 0, 1, \ldots$, we have $\|\mathbf{H}_k\mathbf{v}\|_2^2 = \|\mathbf{v}\|_2^2$ for all $\mathbf{v}$. I.e., $\mathbf{H}_k$ is orthogonal.

Property 2: Can compute $\mathbf{\Pi x} = \mathbf{SHDx}$ in $O(n \log n)$ time.

Property 3: The randomized Hadamard matrix is a good "mixing matrix" for smoothing out vectors.



Deterministic Hadamard matrix.

Randomized Hadamard **PHD**.

Fully random sign matrix.

Blue squares are $1/\sqrt{n}$'s, white squares are $-1/\sqrt{n}$'s.

Lemma (SHRT mixing lemma)

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

$$(z_i)^2 \leq \frac{c \log(n/\delta)}{n} \|\mathbf{z}\|_2^2$$

for some fixed constant c.

The vector is very close to uniform with high probability. As we saw earlier, we can thus argue that $\|\mathbf{Sz}\|_2^2 \approx \|\mathbf{z}\|_2^2$. I.e. that:

$$\|\mathbf{\Pi x}\|_2^2 = \|\mathbf{SHDx}\|_2^2 \approx \|\mathbf{x}\|_2^2$$

Theorem (The Fast JL Lemma)

Let $\mathbf{\Pi} = \mathbf{SHD} \in \mathbb{R}^{m \times n}$ *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* $\mathbf{x}$*,*

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

with probability $(1-\delta)$*.*

Very little loss in embedding dimension compared to full random matrix, and $\mathbf{\Pi}$ can be multiplied by $\mathbf{x}$ in $O(n \log n)$ (nearly linear) time.

SHRT mixing lemma proof: Need to prove $(z_i)^2 \leq \frac{c\log(n/\delta)}{n}\|\mathbf{z}\|_2^2$ for all i.

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

$$\mathbf{h}_i^T\mathbf{D} = \frac{1}{\sqrt{n}}\begin{bmatrix}1 & 1 & \ldots & -1 & -1\end{bmatrix}\begin{bmatrix}D_1 & & & \\ & D_2 & & \\ & & \ddots & \\ & & & D_n\end{bmatrix}$$

where $D_1, \ldots, D_n$ are random ± 1's.

This is equivalent to

$$\mathbf{h}_i^T\mathbf{D} = \frac{1}{\sqrt{n}}\begin{bmatrix}R_1 & R_2 & \ldots & R_n\end{bmatrix},$$

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

So we have, for all i, $\mathbf{z}_i = \mathbf{h}_i^T \mathbf{D}\mathbf{x} = \frac{1}{\sqrt{n}} \sum_{i=1}^n R_i x_i$.

- $\mathbf{z}_i$ is a random variable with mean 0 and variance $\frac{1}{n}\|\mathbf{x}\|_2^2$, and is a sum of independent random variables.
- By Central Limit Theorem, we expect that:

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

- Setting $t = \sqrt{\log(n/\delta)}$, we have for constant c,

$$\Pr\left[|\mathbf{z}_i| \geq c\sqrt{\frac{\log(n/\delta)}{n}}\|\mathbf{y}\|_2\right] \leq \frac{\delta}{n}$$

.

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

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

This is call the Khintchine Inequality. It is specialized to sums of scaled ± 1's, and is a bit tighter and easier to apply than using a generic Bernstein bound.

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

As shown earlier, we can thus guarantee that:

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

as long as $\mathbf{S} \in \mathbb{R}^{m\times n}$ is a random sampling matrix with

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

$\|\mathbf{S}\mathbf{z}\|_2^2 = \|\mathbf{S}\mathbf{H}\mathbf{D}\mathbf{x}\|_2^2 = \|\mathbf{\Pi}\mathbf{x}\|_2^2$ and $\|\mathbf{z}\|_2^2 = \|\mathbf{x}\|_2^2$, so we are done.

Theorem (The Fast JL Lemma)

Let $\mathbf{\Pi} = \mathbf{SHD} \in \mathbb{R}^{m \times n}$ *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* $\mathbf{x}$*,*

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

with probability $(1-\delta)$*.*

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 $\text{poly}(d)$ rows in:

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

$\boldsymbol{\Pi}$ 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:

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

Simple, inspired algorithm that has been used for accelerating:

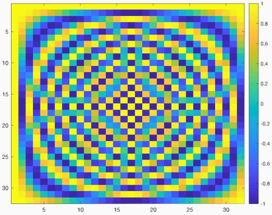
- Vector dimensionality reduction
- Linear algebra
- Locality sensitive hashing (SimHash)
- Randomized kernel learning methods (we will discuss after Thanksgiving)

```
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.

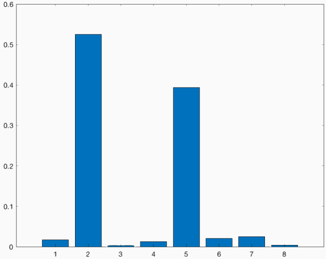
$$\mathbf{F}_{j,k} = e^{-2\pi i \frac{j \cdot k}{n}}, \qquad\qquad \mathbf{F}^* \mathbf{F} = \mathbf{I}.$$



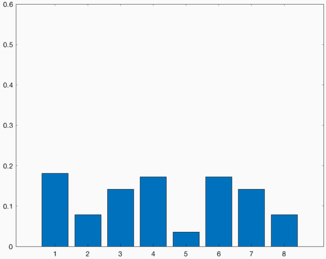
Real part of $\mathbf{F}_{j,k}$.

$\mathbf{F}\mathbf{y}$ computes the Discrete 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 $\mathbf{y}$.



Fourier transform $\mathbf{F}\mathbf{y}$.

What do we know?

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.

One of the central tools in the field of **sparse recovery** aka **compressed sensing.**