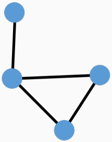
CS-GY 9223 I: Lecture 11
Randomized numerical linear algebra, ϵ-net arguments.

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.

n columns, where n = # of nodes

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

m rows, 1 for each edge

- L is positive semidefinite: $\mathbf{x}^T \mathbf{L}\mathbf{x} \geq 0$ for all $\mathbf{x}$.
- For any vector $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^T L \mathbf{x} = \sum_{(i,j) \in E} (\mathbf{x}(i) - \mathbf{x}(j))^2.$$

$\mathbf{x}^T L \mathbf{x}$ is small if $\mathbf{x}$ is a “smooth” function with respect to the graph.

Courant–Fischer min-max principle

Let $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ be the eigenvectors of $\mathbf{L}$.

smoothness

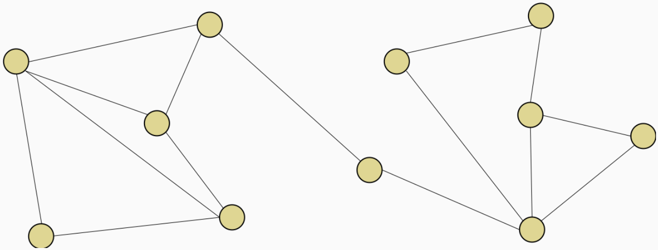
$$\mathbf{v}_n = \arg\min_{\|\mathbf{v}\|=1} \mathbf{v}^T\mathbf{L}\mathbf{v}$$

$$\mathbf{v}_{n-1} = \arg\min_{\|\mathbf{v}\|=1, \mathbf{v}\perp\mathbf{v}_n} \mathbf{v}^T\mathbf{L}\mathbf{v}$$

$$\vdots$$

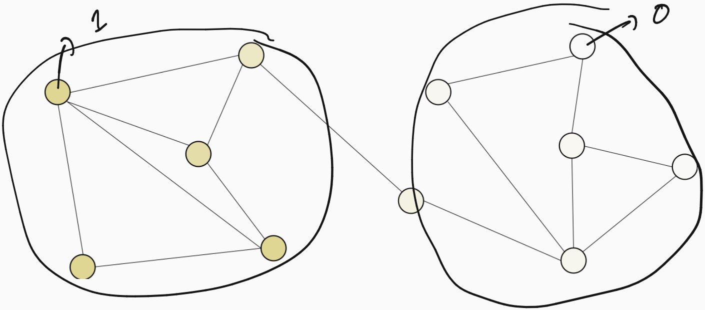
$$\mathbf{v}_1 = \arg\min_{\|\mathbf{v}\|=1, \mathbf{v}\perp\mathbf{v}_n,\ldots,\mathbf{v}_2} \mathbf{v}^T\mathbf{L}\mathbf{v}$$

Eigenvectors of the Laplacian with small eigenvalues correspond to smooth functions over the graph.



Smoothest function is constant. $\mathbf{v}_n = \mathbf{1}$ for any Laplacian $\mathbf{L}$

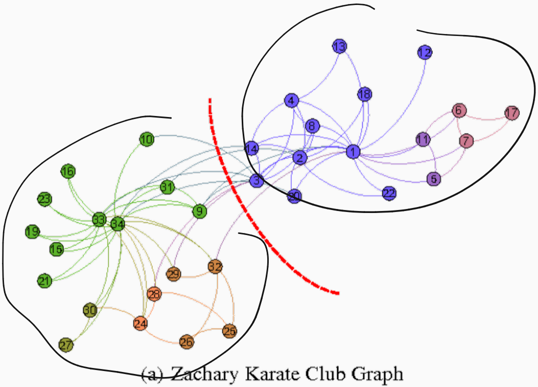
Eigenvectors of the Laplacian with small eigenvalues correspond to smooth functions over the graph.



Other small eigenvectors are not constant, but change slowly in well-connected components.

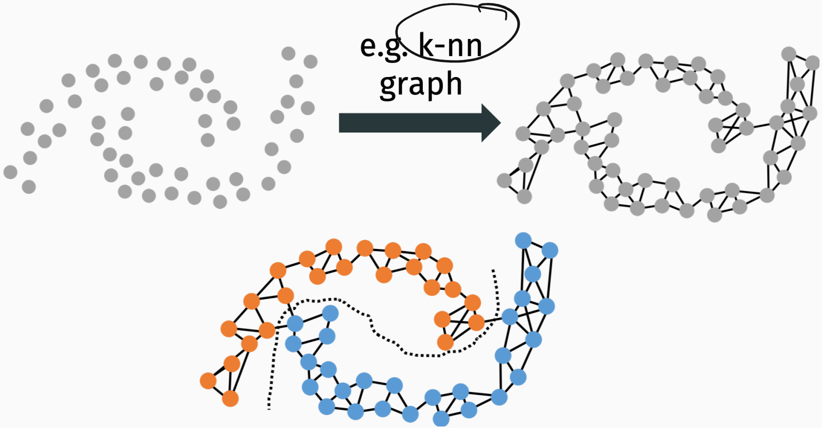
Balanced Cut: Partition nodes along a cut that:

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



(a) Zachary Karate Club Graph

Idea: Construct synthetic graph for data that is hard to cluster.



Spectral Clustering, Laplacian Eigenmaps, Locally linear embedding, Isomap, etc.

- The balanced cut problem is a combinatorial optimization problem: difficult to solve in general.
- Obtain a satisfactory approximate solution through a relax and round approach.
- The problem we relax to is that of computing the second smallest eigenvector of the Laplacian.
- Can be analyzed rigorously for certain classes of random graphs.

SECOND SMALLEST LAPLACIAN EIGENVECTOR

By Courant-Fischer, $\mathbf{v}_{n-1}$ is given by:

$$\mathbf{v}_{n-1} = \arg\min_{\|\mathbf{v}\|=1,\ \mathbf{v}_n^T\mathbf{v}=0} \mathbf{v}^T L \mathbf{v}$$

$\vec{1}$

$\sum_{i=1}^{n} v_{n-1}(i) = 0.$

If $\mathbf{v}_{n-1}$ were binary, i.e. $\in \{-1, 1\}^n$, scaled by $\frac{1}{\sqrt{n}}$, it would have:

- $\mathbf{v}_{n-1}^T L \mathbf{v}_{n-1} = 4 \cdot cut(S, T)$ as small as possible given that $\mathbf{v}_{n-1}^T \mathbf{1} = |T| - |S| = 0$.
- $\mathbf{v}_{n-1}$ would indicate the smallest perfectly balanced cut.

In reality, $\mathbf{v}_{n-1} \in \mathbb{R}^n$ has fractional entries, but we can round these to obtain a good balanced cut.

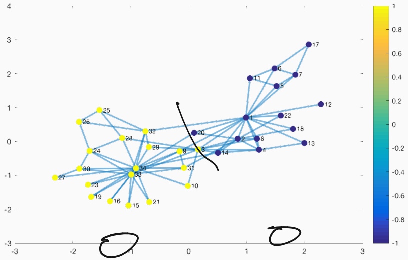
- Compute

$v^T v_n$

$$\mathbf{v}_{n-1} = \underset{v \in \mathbb{R}^n \text{ with } \|\mathbf{v}\|=1,\ \mathbf{v}^T\mathbf{1}=0}{\arg\min} \mathbf{v}^T L \mathbf{v}$$

↳ all ones vector

- Set S to be all nodes with $\mathbf{v}_{n-1}(i) < 0$, and T to be all with $\mathbf{v}_{n-1}(i) \geq 0$.



Stochastic Block Model (Planted Partition Model): $m = \frac{n}{2}$

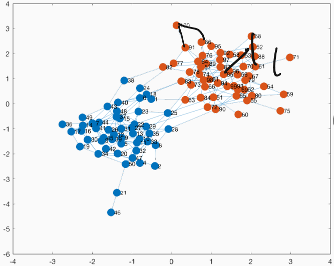
Let $G_n(p, q)$ be a distribution over graphs on n nodes, split equally into two groups B and C, each with $n/2$ nodes.

- 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. $q < p$.

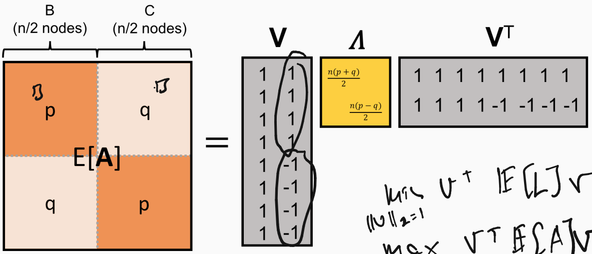
$\binom{m}{2} = \frac{m^2}{2}$

$p m^2 / 2$

$q m^2$



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



$\min_{\|v\|_2=1} v^T \mathbb{E}[L] v$

$\max_{\|v\|_2=1} v^T \mathbb{E}[A] v$

- $\mathbf{v}_1 = \mathbf{1}$ with eigenvalue $\lambda_1 = \frac{(p+q)n}{2}$.
- $\mathbf{v}_2 = \underline{\chi_{B,C}}$ 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 $\mathbb{E}[\mathbf{L}]$ 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.**

- Analogous to scalar concentration inequalities like Markovs, Chebyshevs, Bernsteins.

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 $\mathbf{A}, \bar{\mathbf{A}} \in \mathbb{R}^{d \times d}$ are symmetric with $\|\mathbf{A} - \bar{\mathbf{A}}\|_2 \leq \epsilon$ and eigenvectors $v_1, v_2, \ldots, v_d$ and $\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_d$. Letting $\theta(v_i, \bar{v}_i)$ denote the angle between v_i and $\bar{v}_i$, for all i:

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

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\bar{\mathbf{A}}$.

The error gets larger if there are eigenvalues with similar magnitudes.

$$\overset{\mathbf{A}}{\begin{bmatrix}1+\varepsilon & 0\\ 0 & 1\end{bmatrix}} - \overset{\bar{\mathbf{A}}}{\begin{bmatrix}1 & 0\\ 0 & 1+\varepsilon\end{bmatrix}} = \overset{\|\mathbf{A}\text{-}\bar{\mathbf{A}}\|_2 = \varepsilon}{\begin{bmatrix}\varepsilon & 0\\ 0 & -\varepsilon\end{bmatrix}}$$

$$\begin{bmatrix}1\\ 0\end{bmatrix} \qquad\qquad \begin{bmatrix}0\\ 1\end{bmatrix}$$

v_2 second eigenvector of A

$\bar{v}_2$ is eigenvector of $\mathbb{E}[A]$

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

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

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

$$\sin\theta(\underset{A}{v_2}, \underset{\mathbb{E}[A]}{\bar{v}_2}) \leq \frac{O(\sqrt{pn})}{\min_{j\neq i}|\lambda_i - \lambda_j|} \leq \frac{O(\sqrt{pn})}{(p-q)n/2} = O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)$$

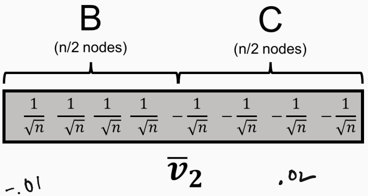
Recall: $\mathbb{E}[A]$, has eigenvalues $\lambda_1 = \frac{(p+q)n}{2}$, $\lambda_2 = \frac{(p-q)n}{2}$, $\lambda_i = 0$ for $i \geq 3$.

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

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

So Far: $\sin\theta(v_2, \bar{v}_2) \leq O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)$. What does this give us?

- Can show that this implies $\|v_2 - \bar{v}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$ (exercise).
- $\bar{v}_2$ is $\frac{1}{\sqrt{n}}\chi_{B,C}$: the community indicator vector.

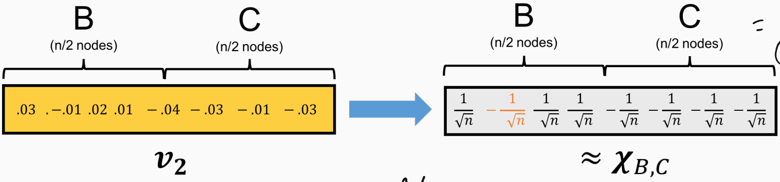


- Every i where $v_2(i)$, $\bar{v}_2(i)$ differ in sign contributes $\geq \frac{1}{n}$ to $\|v_2 - \bar{v}_2\|_2^2$.
- So they differ in sign in at most $O\left(\frac{p}{(p-q)^2}\right)$ positions.

APPLICATION TO STOCHASTIC BLOCK MODEL

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.

c/n $\quad ((c-c')/n)^2$ $\quad = \frac{c}{(c-c')^2 n}$



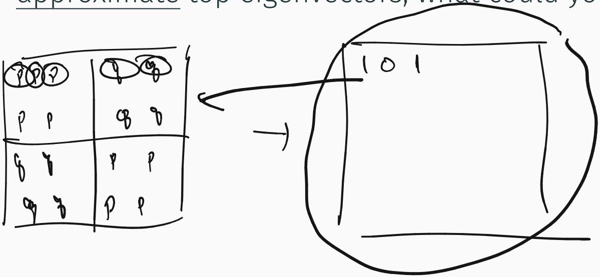
$q = c'/n$

- Think of $p = c/n$ 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 correctly.

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

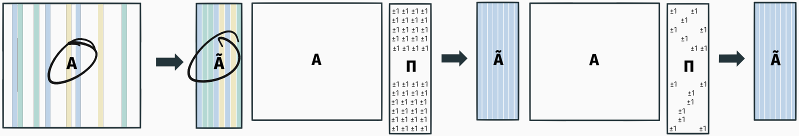
- Dense $n \times n$ matrix.
- Computing top eigenvectors takes $\approx O(n^2/\sqrt{\epsilon})$ time.

If someone asked you to speed this up and return <u>approximate</u> top eigenvectors, what could you do?.

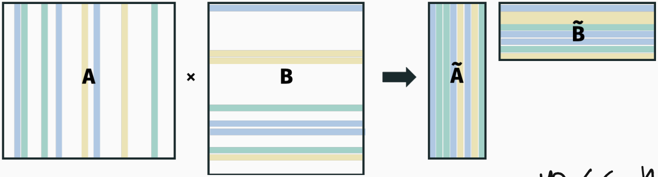


Main idea: If you want to compute singular vectors or eigenvectors, multiply two matrices, solve a regression problem, etc.:

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 $\mathbf{A}$.



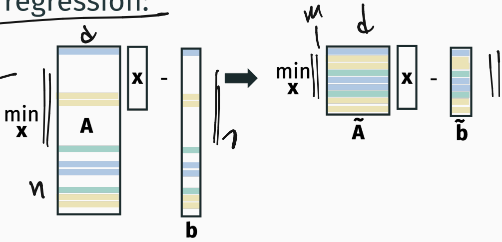
Approximate matrix multiplication:



$m << n$

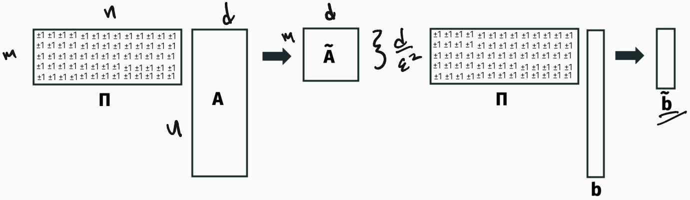
Approximate regression:

$O(nd^2)$



$O(md^2)$

Randomized approximate regression using a Johnson-Lindenstrauss Matrix:



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

$m \to \tilde{A}$ $\quad m \to \tilde{b}$

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

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

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

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

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

$$\|A\tilde{x}-b\|_2^2 \leq \frac{1}{(1-\varepsilon)}\|\Pi A\tilde{x}-\Pi b\|_2^2 \leq \frac{1}{(1-\varepsilon)}\|\Pi Ax^*-\Pi b\|_2^2$$

$$\leq \frac{(1+\varepsilon)}{(1-\varepsilon)}\|Ax^*-b\|_2^2$$

$$= (1+O(\varepsilon))\|Ax^*-b\|_2^2$$

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* $\mathbf{y}$*,*

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

with probability $(1-\delta)$*.* $y = Ax - b$

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

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

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

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

This statement requires establishing a Johnson-Lindenstrauss type bound for an <u>infinity</u> 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 $\mathbf{A}$.

$\leq d+1$

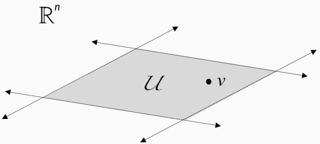
Theorem (Subspace Embedding from JL)

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

$\in \mathbb{B}^n$

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

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



[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.

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)

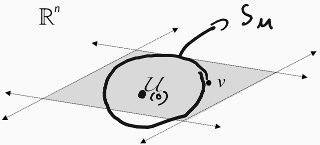
Let $\mathcal{U} \subset \mathbb{R}^n$ be a d-dimensional linear subspace in $\mathbb{R}^n$. If $\mathbf{\Pi} \in \mathbb{R}^{m \times d}$ is chosen from any distribution $\mathcal{D}$ satisfying the Distributional JL Lemma, then with probability $1 - \delta$,

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

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

$\frac{d}{\epsilon^2} < n$

$d \ll n$



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

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

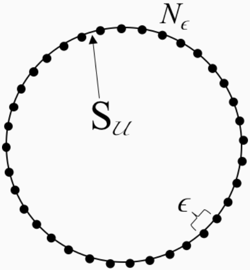
Follows from linearity: Any point $\mathbf{v} \in \mathcal{U}$ can be written as $c\mathbf{w}$ for some scalar c and some point $\mathbf{w} \in S_{\mathcal{U}}$.

$v = c\,w$

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

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

$N_\varepsilon \subset S_{\mathcal{U}}$



For all $v \in S_{\mathcal{U}}$ there exist a $w \in N_\varepsilon$ s.t. $\|w - v\|_2 \le \varepsilon$.

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

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

$\frac{1}{|N_\epsilon|}$

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

with prob $1-\delta$ we preserve norm for all $w \in N_\epsilon$

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

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.

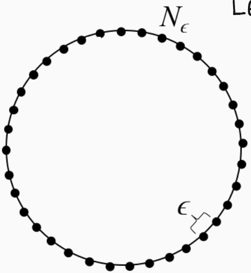
$$\frac{\log\left(\left(\frac{4}{\epsilon}\right)^d \cdot \frac{1}{\delta}\right)}{\epsilon^2} = \frac{d\log(4/\epsilon)+\log(1/\delta)}{\epsilon^2} = O\left(\frac{d\log(1/\epsilon)}{\epsilon^2}\right)$$

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 $|c_i| \leq \epsilon^i$.



Let $w_0 = \arg\min_{w \in N_\epsilon} \|w - v\|_2$. Let $r_0 = v - w_0$

For $i = 1, 2, \ldots$

- Let $w_i = \arg\min_{w \in N_\epsilon} \left\| w - \frac{r_{i-1}}{\|r_{i-1}\|} \right\|_2$

- $r_i = \left(\frac{r_{i-1}}{\|r_{i-1}\|} - w_i \right)$

Then notice: $v = w_0 + \|r_1\| w_1 + \|r_1\| \|r\|_2 w_2 + \ldots + \left(\prod_{i=1}^{q} \|r_i\| \right) \cdot w_i + \ldots$

This sum satisfies the claim.

$\|w_i\| = 1$ for all i

3. Preserving norm of v.

$\|\Pi w_i\| \leq (1+\epsilon) \cdot 1$

$= (1+\epsilon)$

Applying triangle inequality, we have

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

$= (1+\epsilon)\underbrace{(1+\epsilon+\epsilon^2+\ldots)}$

$\leq (1+2\epsilon) = \boxed{1+O(\epsilon)}$

3. Preserving norm of v.

$\|\Pi w_i\|_2 \geq (1-\epsilon)\|w_i\|_2$
for all i
$= (1-\epsilon)\cdot 1$
$= (1-\epsilon)$

Similarly,

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

$$= (1-\epsilon) - \underbrace{(\epsilon + \epsilon^2 + \ldots)(1+\epsilon)}_{\leq 2\epsilon}$$

$$\geq (1-3\epsilon)$$

So we have proven

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

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

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

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 $1 - \delta$,

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

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

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

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

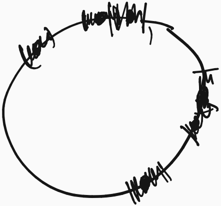
Imaginary algorithm for constructing N_ϵ:

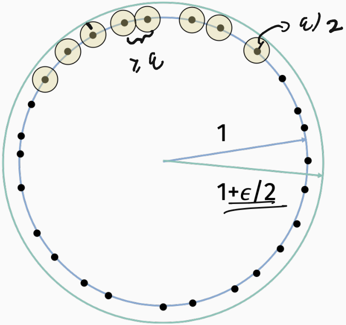
- Set $N_\epsilon = \{\}$
- While such a point exists, choose an arbitrary point $\mathbf{v} \in S_{\mathcal{U}}$ where $\nexists \mathbf{w} \in N_\epsilon$ with $\|\mathbf{v} - \mathbf{w}\| \leq \epsilon$. Set $N_\epsilon = N_\epsilon \cup \{\mathbf{w}\}$.

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.

$O(\frac{1}{\epsilon})^d$

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:

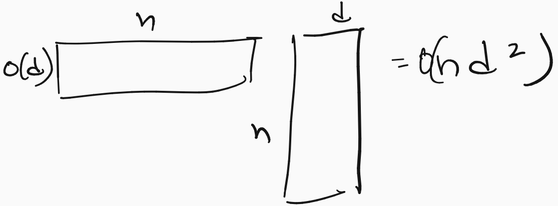
$$\begin{aligned}
\mathrm{vol}(d, \epsilon/2) \cdot |N_\epsilon| &\leq \mathrm{vol}(d, 1+\epsilon/2) \\
|N_\epsilon| &\leq \frac{\mathrm{vol}(d, 1+\epsilon/2)}{\mathrm{vol}(d, \epsilon/2)} \quad \frac{c\,(1+\epsilon/2)^d}{c\,(\epsilon/2)^d} \\
\leq 2 \qquad &\leq \left(\frac{1+\epsilon/2}{\epsilon/2}\right)^d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

RUNTIME CONSIDERATION

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

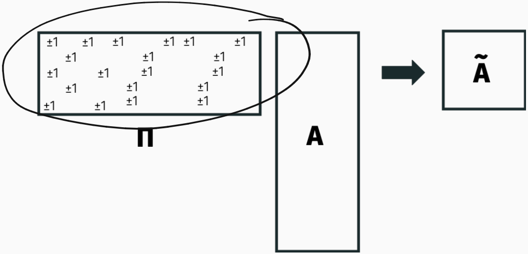
- Cost to solve $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$:
  - $O(nd^2)$ time for direct method. Need to compute $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$.
  - $O(nd) \cdot$ (# of iterations) time for iterative method (GD, AGD, conjugate gradient method).
- Cost to solve $\|\mathbf{\Pi}\mathbf{A}\mathbf{x}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

ΠA



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.

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 $2^i \times 2^i$.

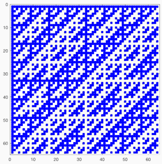
$$H_0 = 1 \quad H_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad H_2 = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

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

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 $\mathbf{\Pi} \in \mathbb{R}^{m \times n}$.



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, <u>each bounded by $O(\log(n/\delta))$</u>, 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:

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

Johnson-Lindenstrauss with SHRTs

Theorem (JL from SRHT)

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

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

Upshot for regression: Compute $\mathbf{\Pi}\mathbf{A}$ 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:

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

$\mathbf{\Pi}$ 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:

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

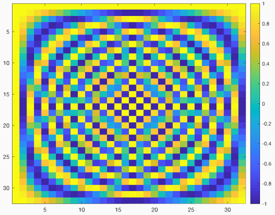
Simple, inspired algorithm that has been used for accelerating:

- 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);
```

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

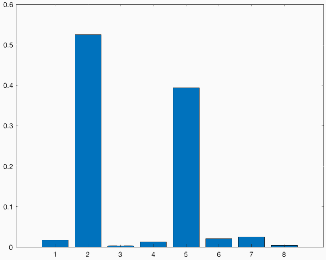
$$\mathbf{F}_{j,k} = e^{-2\pi i \frac{j \cdot k}{n}}, \qquad\qquad \mathbf{F}^*\mathbf{F} = \mathbf{I}.$$



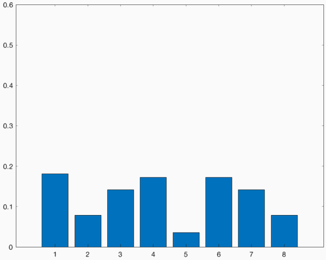
Real part of $\mathbf{F}_{j,k}$.

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