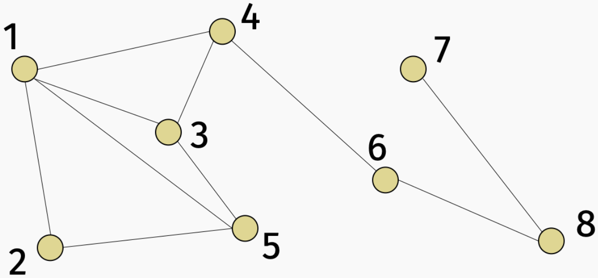
CS-GY 6763: Lecture 12
Stochastic Block Model, subspace embeddings + ϵ-net arguments

NYU Tandon School of Engineering, Prof. Christopher Musco

Main idea: Understand graph data by constructing natural matrix representations, and studying that matrix's spectrum (eigenvalues/eigenvectors).

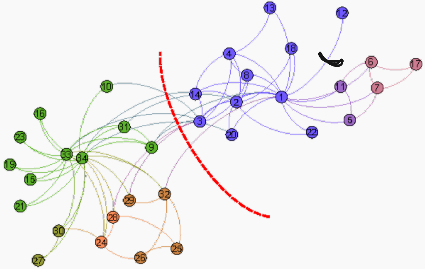


$G = (V, E)$ is an undirected, unweighted graph with n nodes.

BALANCED CUT

Goal: Given a graph $G = (V, E)$, 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

Applications: Understanding community structure in social networks, partitioning finite element meshes, non-linear clustering in machine learning, data visualization, etc. etc.

β-Balanced Cut:

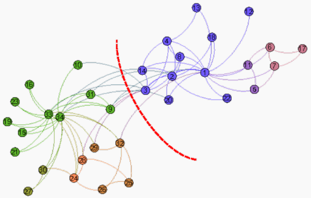
$$\min_S \text{cut}(S, V \setminus S) \quad \text{such that} \quad \min\left(|S|, |V \setminus S|\right) \geq \beta \cdot n \text{ for } \beta \leq .5$$

Last class we focused on the extreme case where $\beta = 1/2$.

Basic spectral clustering method:

Laplacian

- Compute second smallest eigenvector of graph, $\mathbf{v}_{n-1}$.
- $\mathbf{v}_{n-1}$ has an entry for every node i in the graph.
- If the i^{th} entry is positive, put node i in T.
- Otherwise if the i^{th} entry is negative, put i in S.



(a) Zachary Karate Club Graph

For a cut indicator vector $\mathbf{c} \in \{-1, 1\}^n$ with $\mathbf{c}(i) = -1$ for $i \in S$ and $\mathbf{c}(i) = 1$ for $i \in T$:

$c \in \{-1/\sqrt{n}, 1/\sqrt{n}\}$

- $\mathbf{c}^T L \mathbf{c} = 4 \cdot cut(S, T)$.
- $\mathbf{c}^T \mathbf{1} = |T| - |S|$. $= 0$

Want to minimize both $\mathbf{c}^T L \mathbf{c}$ (cut size) and $|\mathbf{c}^T \mathbf{1}|$ (imbalance).

$\frac{1}{n} \cdot 4 \cdot$ cut value

Perfectly balanced balanced cut problem:

$$\min_{\mathbf{c} \in \{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\}^n} \mathbf{c}^T \mathbf{L} \mathbf{c} \text{ such that } \mathbf{c}^T \mathbf{1} = 0.$$

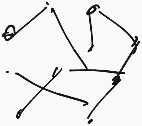
$|S| = |T|$

Relaxed perfectly balanced balanced cut problem:

$$\left(\min_{\|\mathbf{c}\|_2 = 1} \mathbf{c}^T \mathbf{L} \mathbf{c} \text{ such that } \mathbf{c}^T \mathbf{1} = 0. \right)$$

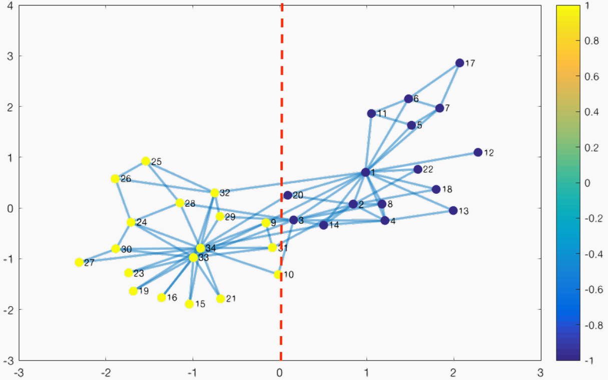
Main result: The relaxed problem is exactly minimized by the second smallest eigenvector $\mathbf{v}_{n-1}$ of $\mathbf{L}$.

Final relax and round algorithm: Compute



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

Set S to be all nodes with $\mathbf{v}_{n-1}(i) < 0$, and T to be all with $\mathbf{v}_{n-1}(i) \geq 0$. I.e. set $\mathbf{c} = \text{sign}(\mathbf{v}_{n-1})$



So far: Showed that spectral clustering partitions a graph along a small cut between large pieces.

$O(n^3)$

- No formal guarantee on the 'quality' of the partitioning.
- Can fail for worst case input graphs.

Common approach: Design a natural generative model that produces random but realistic inputs and analyze how the algorithm performs on inputs drawn from this model.

- Very common in algorithm design and analysis. Great way to start approaching a problem. Often our best way to understand why some algorithms "just work" in practice.
- Similar approach to Bayesian modeling in machine learning.

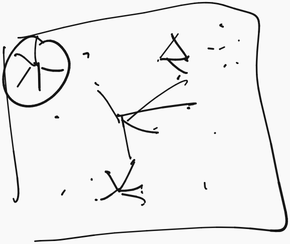
n nodes

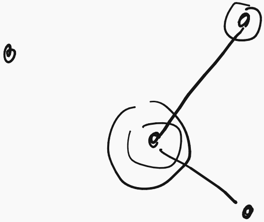
Ideas for a generative model for **social network graphs** that would allow us to understand partitioning?

Graphs

Geometric

$x^{-\alpha}$

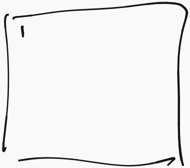


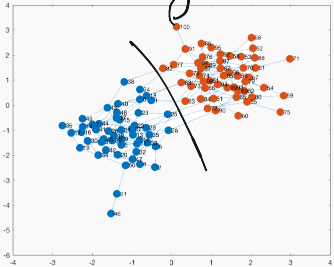


Stochastic Block Model (Planted Partition Model): $p, q \in [0,1]$

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. $q < p$.



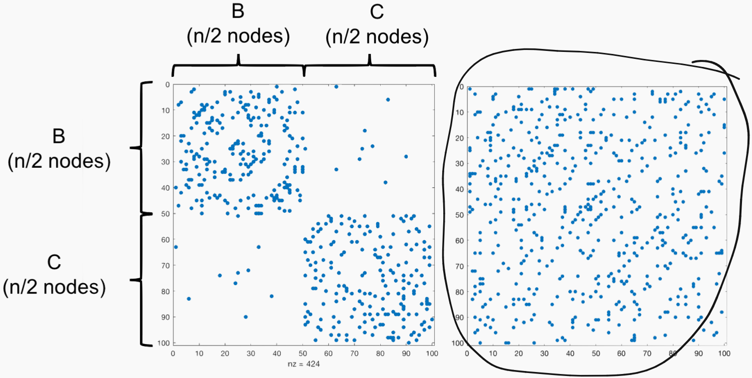


$1 \quad 2 \quad \ldots \quad n/2$

$n/2+1 \quad \ldots \quad n$

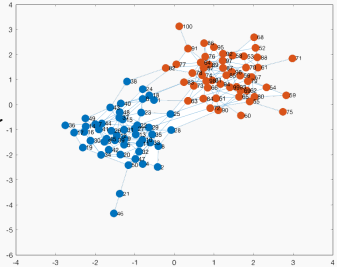
Let G be a stochastic block model graph drawn from $G_n(p, q)$.

- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ denote the adjacency matrix of G.



Note that we are arbitrarily ordering the nodes in A by group. In reality A would look "scrambled" as on the right.

Goal is to find the “ground truth” balanced partition B, C using our standard spectal method.



L → second smallest eigenvector

$\tilde{B}$ $\tilde{C}$

To do so, we need to understand the second smallest eigenvector of $\mathbf{L} = \mathbf{D} - \mathbf{A}$. We will start by considering the expected value of these matrices:

$E[L]_{ij} = \mathbb{E}[L_{ij}]$

$$\mathbb{E}[\mathbf{L}] = \mathbb{E}[\mathbf{D}] - \mathbb{E}[\mathbf{A}].$$

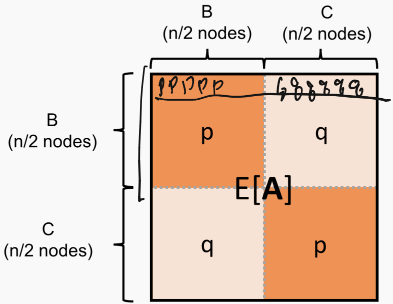
a matrix

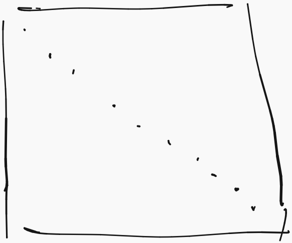
EXPECTED ADJACENCY SPECTRUM

$$\mathbb{E}[L] = \mathbb{E}[D] - \mathbb{E}[A]$$

Letting G be a stochastic block model graph drawn from $G_n(p, q)$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$ be its adjacency matrix. $(\mathbb{E}[\mathbf{A}])_{i,j} = p$ for i, j in same group, $(\mathbb{E}[\mathbf{A}])_{i,j} = q$ otherwise.

$$\mathbb{E}[D] = \frac{(p+q)n}{2} \cdot I$$





$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}[\text{node } i \text{ connects to node } j]\right]$$

What is the expected Laplacian of $G_n(p, q)$?

$$\mathbb{E}(L) = \mathbb{E}[D] - \mathbb{E}[A] = \underbrace{cI} - \mathbb{E}(A)$$

Suppose v is eigenvector of $\mathbb{E}[L]$.

$\mathbb{E}[L] \circ v = \lambda v$ for some $\lambda \in \mathbb{R}$.

$$(\mathbb{E}[D] - \mathbb{E}[A])v = \lambda v \quad = \quad -\mathbb{E}[A]v = \lambda v - \mathbb{E}[D]v$$

$$= \lambda v - cv$$

$$\mathbb{E}[A]v = \underline{\underline{(c - \lambda)}}v.$$

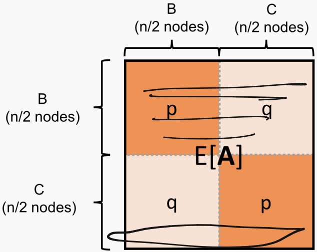
$\lambda_1 \to c - \lambda_1$
$\lambda_2 \to c - \lambda_2$
$\vdots$
$\lambda_n \quad c - \lambda_n$

$\mathbb{E}$[A] and $\mathbb{E}$[L] have the same eigenvectors and eigenvalues are equal up to a shift/inversion. So (second largest eigenvector) of $\mathbb{E}$[A] is the same as the second smallest of $\mathbb{E}$[L]

EXPECTED ADJACENCY SPECTRUM

Letting G be a stochastic block model graph drawn from $G_n(p,q)$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$ be its adjacency matrix, what are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{A}]$?

$Mv = \lambda v.$

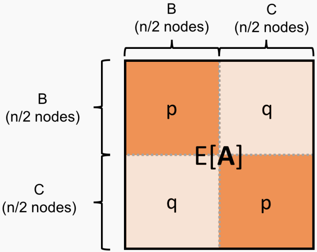


$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} (p-q)\,n/2 \\ (p-q)\,n/2 \\ \vdots \\ (q-p)\,n/2 \end{bmatrix} = \frac{(p-q)n}{2} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}$$

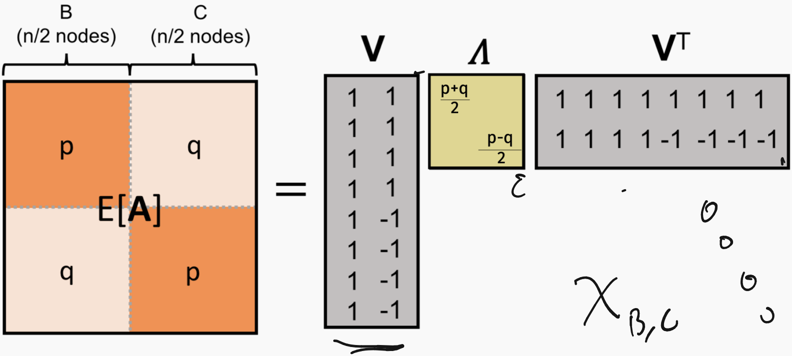
n-2 eigenvalues $= 0$

$$\begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} (p+q)\,n/2 \\ \vdots \\ \vdots \\ (p+q)\,n/2 \end{bmatrix} = \frac{(p+q)n}{2} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Letting G be a stochastic block model graph drawn from $G_n(p,q)$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$ be its adjacency matrix, what are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{A}]$?



EXPECTED ADJACENCY SPECTRUM



- $\bar{\mathbf{v}}_1 \sim \mathbf{1}$ with eigenvalue $\lambda_1 = \frac{(p+q)n}{2}$.
- $\bar{\mathbf{v}}_2 \sim \chi_{B,C}$ with eigenvalue $\lambda_2 = \frac{(p-q)n}{2}$.

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

EXPECTED LAPLACIAN SPECTRUM

Upshot: The second smallest eigenvector of $\mathbb{E}[\mathbf{L}]$, equivalently the second largest of $\mathbb{E}[\mathbf{A}]$, is exactly $\boldsymbol{\chi}_{B,C}$ – 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.

$\|A - \mathbb{E}[A]\|_2$

$\begin{bmatrix} p & q \\ q & p \end{bmatrix}$

Alon, Krivelevich, Vu, 2002:

Matrix Concentration Inequality: If $p \geq O\left(\frac{\log^4 n}{n}\right)$, then with high probability $1 - 1/\text{poly}(n)$

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

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

Recall that $\|\mathbf{X}\|_2 = \max_{z \in \mathbb{R}^n : \|z\|_2 = 1} \|\mathbf{X}z\|_2 = \sigma_1(\mathbf{X})$.

$\|\mathbf{A}\|_2$ is on the order of $O(pn)$ so another way of thinking about the right hand side is $\frac{\|\mathbf{A}\|_2}{\sqrt{np}}$. I.e. get's better with p and n.

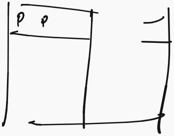
$\|A\|_2 \approx O(pn)$

$\|A\|_2$ is on the order of $O(pn)$ so another way of thinking about the right hand side is $\frac{\|A\|_2}{\sqrt{np}}$. I.e. get's better with p and n.

$$\frac{\|A\mathbb{1}\|_2}{\|\mathbb{1}\|} \le \|A\|_2 = \max_z \frac{\|Az\|_2}{\|z\|_2}$$

$\|\mathbb{1}\| \to n$

$A\mathbb{1} \to \begin{bmatrix} \approx (p+q)\,n \\ \\ \\ \end{bmatrix}$



$$\frac{\sqrt{(p+q)^2 \cdot n^2 \cdot n}}{\sqrt{n}}$$

$$\frac{(p+q)\,n^{1.5}}{\sqrt{n}} = (p+q)\,n$$

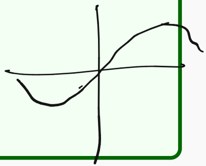
EIGENVECTOR PERTURBATION

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 $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ 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 $\bar{\mathbf{A}}$.



We will apply with $\bar{\mathbf{A}} = \mathbb{E}[\mathbf{A}]$.

EIGENVECTOR PERTURBATION

$$\|A-\bar{A}\|_2 = \varepsilon$$

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

$$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \quad \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \quad \begin{bmatrix}1\\ 0\end{bmatrix} \quad \begin{bmatrix}0\\ 1\end{bmatrix}$$

$$\underset{\downarrow\, v_1}{\left(\begin{bmatrix}1\\0\end{bmatrix}, (1+\varepsilon)\right)} \quad \underset{\downarrow\, v_2}{\left(\begin{bmatrix}0\\1\end{bmatrix}, 1\right)} \qquad \underset{\downarrow\, \bar{v}_2}{\left(\begin{bmatrix}1\\0\end{bmatrix}, 1\right)}, \underset{\downarrow\, \bar{v}_1}{\left(\begin{bmatrix}0\\1\end{bmatrix}, 1+\varepsilon\right)}$$

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

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

$\mathbb{E}[A] = \bar{A}$

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 $i \geq 3$.

$((p+q)-(p-q))\frac{n}{2}$

$i=2$

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

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

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

$O(1/\sqrt{n})$

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

(A slightly trickier analysis can remove the qn term entirely.)

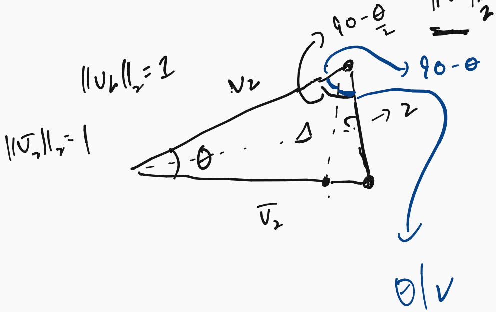
$= \Delta$

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 $\|\mathbf{v}_2 - \bar{\mathbf{v}}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$.

$\sin(\theta(v_2, \bar{v}_2)) = \Delta$

$\|z\|_2 \quad \|v_2 - \bar{v}_2\|_2$



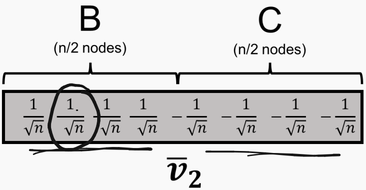
$\|z\|_2 \leq \Delta + \Delta z$

$\|z\|_2^2 = O(\Delta^2)$

APPLICATION TO STOCHASTIC BLOCK MODEL

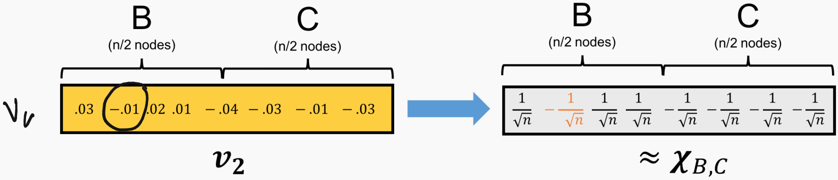
So far: $\|\mathbf{v}_2 - \bar{\mathbf{v}}_2\|_2^2 \le O\left(\frac{p}{(p-q)^2 n}\right)$ $\frac{1}{\sqrt{n}}\,\mathrm{sign}(v_2)$

- $\bar{\mathbf{v}}_2$ is $\frac{1}{\sqrt{n}}\chi_{B,C}$: the community indicator vector.



$\ge 1/\sqrt{n}$

- We want to show that $\mathrm{sign}(\mathbf{v}_2)$ and $\bar{\mathbf{v}}_2$ are close. They only differ at locations where $\mathbf{v}_2$ and $\bar{\mathbf{v}}_2$ differ in sign.



$| \; 1/\sqrt{n} \;\; -1/\sqrt{n} \;\; 1/\sqrt{n} \; \cdots \; |$

Main argument:

- Every i where $v_2(i)$, $\bar{v}_2(i)$ differ in sign contributes $\geq \frac{1}{n}$ to $\|\mathbf{v}_2 - \bar{\mathbf{v}}_2\|_2^2$.
- 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.

$p = \frac{250}{n}$

Upshot: If G is a stochastic block model graph with adjacency matrix $\mathbf{A}$, if we compute its second largest eigenvector $\mathbf{v}_2$ 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:** Suppose $\underline{q = .8p}$ so $\frac{p}{(p-q)^2} = 25/p$. Even if p is really small, i.e. $p = 250/n$, then we assign roughly 90% of nodes to the right partition.

$$\frac{p}{(p-.8)^2} = \frac{p}{(p-.8p)^2} = \frac{p}{(p/5)^2} = \frac{25}{p} = \frac{25}{250/n} = \underline{.10n}$$

RANDOMIZED NUMERICAL LINEAR ALGEBRA

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

$$\begin{bmatrix} P & Q \\ Q & P \end{bmatrix}$$

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

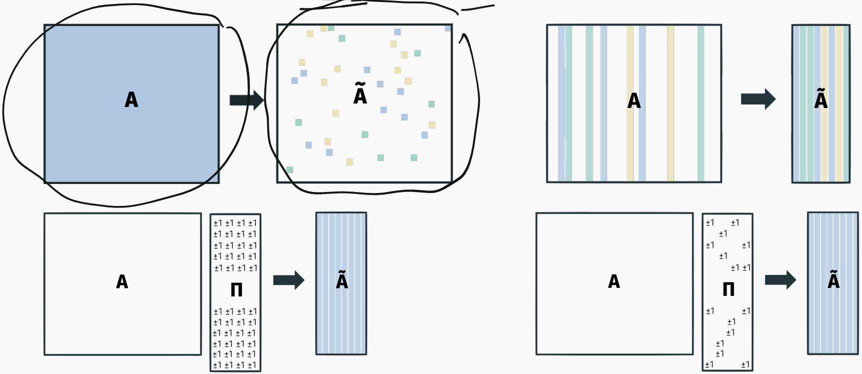
If someone asked you to speed this up and return <u>approximate</u> top eigenvectors, what could you do?

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

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

3:30pm

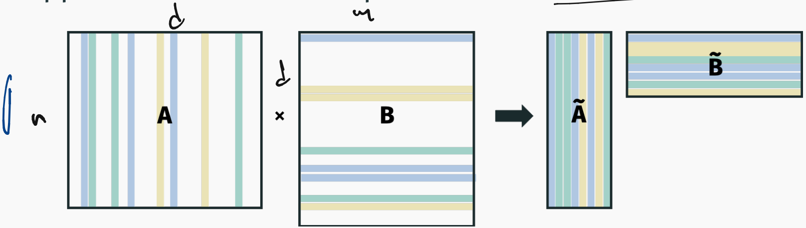
1. Compress your matrices using a randomized method (e.g. subsampling).
2. Solve the problem on the smaller or sparser matrix.
   - $\tilde{\mathbf{A}}$ called a "sketch" or "coreset" for $\mathbf{A}$.



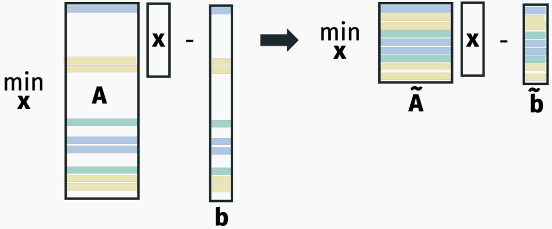
BREAK

Approximate matrix multiplication:

$m \times k \quad k \times m$

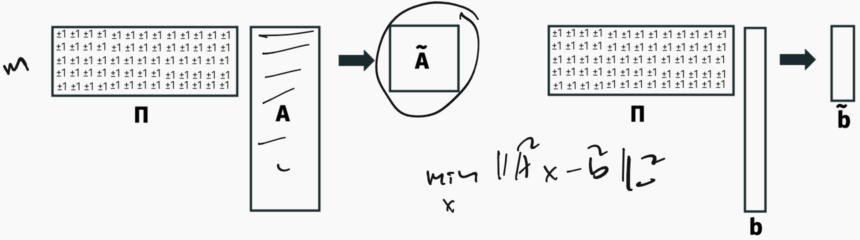


Approximate regression:



$O(nd^2)$ $O(md^2)$ $m \approx O(d)$

Today's example: Randomized approximate regression using a Johnson-Lindenstrauss matrix for compression.



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

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

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

Theorem (Randomized Linear Regression)

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

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

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

PLAN

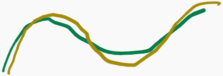
- Prove this theorem using an ϵ-net argument, which is a popular technique for applying our standard concentration inequality + union bound argument to an infinite number of events.
- These sort of arguments appear all the time in theoretical algorithms and ML research, so this part of lecture is as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

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

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

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

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



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

optimality of $\tilde{x}$

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

Lemma (Distributional JL)

If $\mathbf{\Pi}$ *is chosen to a random Gaussian matrix, sign matrix, sparse random matrix, etc. (scaled by* $1/\sqrt{m}$*) with* $m = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$ *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 $(1-\delta)$,

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

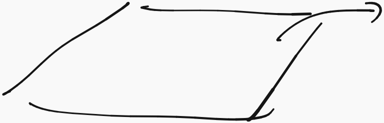
$$= \|\Pi(Ax-b)\|_2^2$$

FOR ANY TO FOR ALL

$\begin{bmatrix} A \end{bmatrix} \begin{bmatrix} x \end{bmatrix} - \begin{bmatrix} b \end{bmatrix}$

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 infinity of possible vectors $(\mathbf{A}\mathbf{x} - \mathbf{b})$ which can't be tackled directly with a union bound argument.



d + 1
dimensional

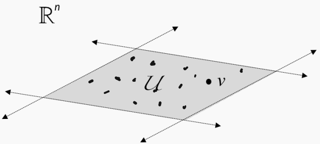
Note that all vectors of the form $(\mathbf{A}\mathbf{x} - \mathbf{b})$ lie in a low dimensional subspace: spanned by $d + 1$ vectors, where d is the width of $\mathbf{A}$. **So even though the set is infinite, it is "simple" in some way. Parameterized by just $d + 1$ numbers.**

Theorem (Subspace Embedding from JL) [Sarlos, 2006]

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] $\approx O\left(\frac{d}{\epsilon^2}\right)$



[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,

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

for all $\mathbf{x}$ and thus

$Ax - b$ $\quad$ $d+1$

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

I.e., our main theorem is proven.

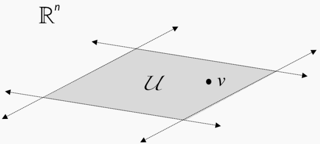
Proof: Apply Subspace Embedding Thm. to the $(d+1)$ dimensional subspace spanned by $\mathbf{A}$'s d columns and $\mathbf{b}$. Every vector $\mathbf{A}\mathbf{x}-\mathbf{b}$ lies in this subspace.

Theorem (Subspace Embedding from JL) 1

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



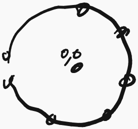
Subspace embeddings have tons of other applications!

SUBSPACE EMBEDDING PROOF

$w = c_1 e_1 + c_2 e_2 + \dots + c_d e_d \qquad \Pi w = c_1 \Pi e_1 \dots c_d \Pi e_d$

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

First 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 $\mathbf{w} \in S_{\mathcal{U}}$.

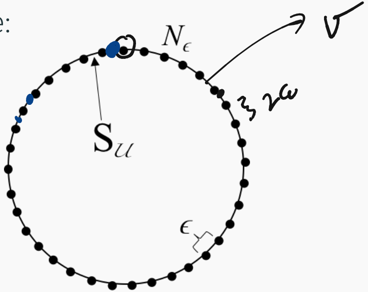
$v = c \cdot w$

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

$\|\Pi w\| \leq \|c_1 \Pi e_1\| + \dots + \|c_d \Pi e_d\|_2$

SUBSPACE EMBEDDING PROOF

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



Goal: Find a set N_ϵ such that, for every $\mathbf{v} \in S_\mathcal{U}$, there is some point $\mathbf{w} \in N_\epsilon$ such that $\|\mathbf{w} - \mathbf{v}\|_2 \leq \epsilon$. N_ϵ is called an "ϵ"-net.
If we can prove

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

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 $|N_\epsilon| = \left(\frac{3}{\epsilon}\right)^d$ such that $\forall \mathbf{v} \in S_{\mathcal{U}}$,

$$\min_{\mathbf{w} \in N_\epsilon} \|\mathbf{v} - \mathbf{w}\|_2 \leq \epsilon.$$

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

$$\log\left(\left(\frac{3}{\varepsilon}\right)^d\right) = d \log(3/\varepsilon)$$

$1-\delta'$

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

1. Preserving norms of all points in net N_ϵ.

Set $\delta' = \frac{1}{|\mathcal{N}_\epsilon|} \cdot \delta = \left(\frac{\epsilon}{3}\right)^d \cdot \delta$. 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, then by a union bound,

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

for all $\mathbf{w} \in N_\epsilon$,with probability $1-\delta$.

$\delta' = \frac{1}{|N_\epsilon|} \cdot \delta = \left(\frac{\epsilon}{3}\right)^d \cdot \delta$

δ

$v - w_0$

2. Extending to all points in the sphere.

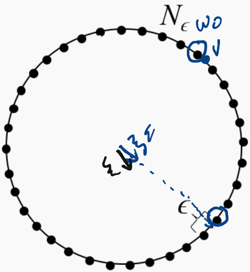
For some $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2 \ldots \in N_\epsilon$, any $\mathbf{v} \in S_\mathcal{U}$ can be written:

$$\mathbf{v} = 1\mathbf{w}_0 + c_1\mathbf{w}_1 + c_2\mathbf{w}_2 + \ldots$$

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

residual

$$r_i = (w_0 + \ldots + c_{i-1} w_{i-1}) - v$$



$|c_1| \leq \epsilon$

$|c_2| \leq \epsilon^2$

$\vdots$

$|c_i| \leq \epsilon^i$

$$\|r_i\|_2 \leq \epsilon^i$$

c_1

$$\exists\ w_1 : \left\| w_1 - \frac{r_1}{\|r_1\|} \right\| \leq \epsilon \quad = \|w_1 \|r_1\| - r_1\|_2 \leq \epsilon \|r_1\|$$

ϵ^2

$\frac{r_1}{\|r_1\|}$

2. Extending to all points in the sphere.

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

Greedy construction:

$$\mathbf{w}_0 = \min_{\mathbf{w} \in \mathcal{N}_\epsilon} \|\mathbf{v} - \mathbf{w}_0\|_2 \qquad \mathbf{r}_0 = \mathbf{v} - \mathbf{w}_0$$

$$\mathbf{w}_1 = \min_{\mathbf{w} \in \mathcal{N}_\epsilon} \left\| \frac{\mathbf{r}_0}{\|\mathbf{r}_0\|} - \mathbf{w}_0 \right\|_2 \quad c_1 = \|\mathbf{r}_0\|_2 \quad \mathbf{r}_1 = \mathbf{v} - \mathbf{w}_0 - c_1\mathbf{w}_1$$

$$\mathbf{w}_2 = \min_{\mathbf{w} \in \mathcal{N}_\epsilon} \left\| \frac{\mathbf{r}_1}{\|\mathbf{r}_1\|} - \mathbf{w}_0 \right\|_2 \quad c_2 = \|\mathbf{r}_1\|_2 \quad \mathbf{r}_2 = \mathbf{v} - \mathbf{w}_0 - c_1\mathbf{w}_1 - c_2\mathbf{w}_2$$

$$\vdots$$

SUBSPACE EMBEDDING PROOF

$(1-\epsilon)\, 1 \le \|\Pi v\|_2 \le (1+\epsilon)\, 1$

2. Extending to all points ~~in the sphere.~~

Applying triangle inequality, we have that: $v = w_0 + c_1 w_1 + c_2 w_2 + \dots$

$$
\begin{aligned}
\|\Pi \mathbf{v}\|_2 &= \|\Pi \mathbf{w}_0 + c_1 \Pi \mathbf{w}_1 + c_2 \Pi \mathbf{w}_2 + \dots\| \\
&\le \|\Pi \mathbf{w}_0\| + c_1 \|\Pi \mathbf{w}_1\| + c_2 \|\Pi \mathbf{w}_2\| + \dots \\
&\le \|\Pi \mathbf{w}_0\| + \epsilon \|\Pi \mathbf{w}_1\| + \epsilon^2 \|\Pi \mathbf{w}_2\| + \epsilon^3 \|\Pi w_3\| \dots \\
&\le (1+\epsilon) + \epsilon \underbrace{(1+\epsilon)}_{\le 2} + \epsilon^2 (1+\epsilon) + \dots \\
&\le 1 + 4\epsilon.
\end{aligned}
$$

$1 + \epsilon + 2\epsilon + 2\epsilon^2$

$\le (1+\epsilon) + (1+\epsilon)(\underbrace{\epsilon + \epsilon^2 + \dots}_{\le 2\epsilon}) \le 1 + 4\epsilon$

3. Preserving norm of v.

Similarly,

$$\begin{aligned}\|\boldsymbol{\Pi}\mathbf{v}\|_2 &= \|\boldsymbol{\Pi}\mathbf{w}_0 + c_1\boldsymbol{\Pi}\mathbf{w}_1 + c_2\boldsymbol{\Pi}\mathbf{w}_2 + \ldots\| \\ &\geq \|\boldsymbol{\Pi}\mathbf{w}_0\| - \epsilon\|\boldsymbol{\Pi}\mathbf{w}_1\| - \epsilon^2\|\boldsymbol{\Pi}\mathbf{w}_2\| - \ldots \\ &\geq (1-\epsilon) - \epsilon(1+\epsilon) - \epsilon^2(1+\epsilon) - \ldots \\ &\geq 1 - 4\epsilon.\end{aligned}$$

$O(\epsilon)$

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 $1 - \delta$,

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

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

$\|A - A\tilde{V}_k\tilde{V}_k^T\|_F \leq (1+\epsilon)\|A - AV_kV_k^T\|_F$

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 $|N_\epsilon| = \left(\frac{3}{\epsilon}\right)^d$ such that $\forall \mathbf{v} \in S_\mathcal{U}$,

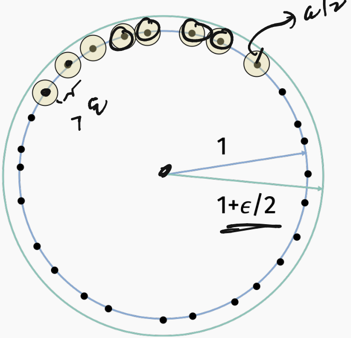
$$\min_{\mathbf{w} \in N_\epsilon} \|\mathbf{v} - \mathbf{w}\| \leq \epsilon.$$

Imaginary algorithm for constructing N_ϵ:

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

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

How many steps does this procedure take?



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

VOLUME ARGUMENT

Volume of d dimensional ball of radius r is

$$\left(\frac{3}{\epsilon}\right)^d$$

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

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

$$\begin{aligned}
\text{vol}(d, \epsilon/2) \cdot |N_\epsilon| &\le \text{vol}(d, 1 + \epsilon/2) \\
|N_\epsilon| &\le \frac{\text{vol}(d, 1 + \epsilon/2)}{\text{vol}(d, \epsilon/2)} \\
&\le \left(\frac{1 + \epsilon/2}{\epsilon/2}\right)^d \le \left(\frac{3}{\epsilon}\right)^d
\end{aligned}$$

$$\frac{\cancel{c}\,(1+\epsilon/2)^d}{\cancel{c}\,(\epsilon/2)^d}$$

$$= \left(\frac{2+\epsilon}{\epsilon}\right)^d \le \left(\frac{3}{\epsilon}\right)^d$$

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.

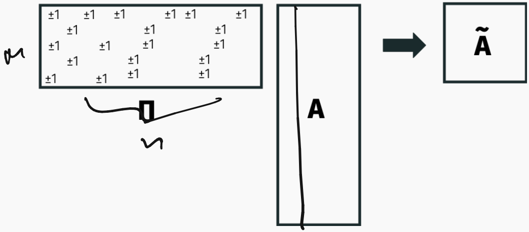
- Cost to solve $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$: $O(nd^2)$
  - $O(nd^2)$ time for direct method. Need to compute $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$.
  - $O(nd) \cdot (\text{\# of iterations})$ time for iterative method (GD, AGD, conjugate gradient method).
- Cost to solve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

$$\underline{\Pi A} \qquad (\underset{\varepsilon^2}{d} \times n)(n \times d) = \cancel{O(nd^2)}$$

$$\underline{O(nd)}$$

But time to compute **ΠA** is an $(m \times n) \times (n \times d)$ matrix multiply: $O(mnd) = O(nd^2)$ time!

Goal: Develop faster Johnson-Lindenstrauss projections.



$O(mn)$

$O(n)$

Typically using sparse and structured matrices.

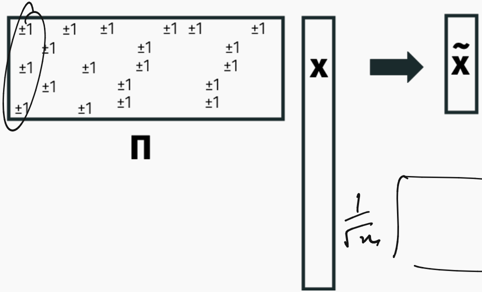
Next class: We will describe a construction where **ΠA** can be computed in $O(nd \log n)$ time.

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:

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

$\log(1/\delta)$

$\frac{1}{\delta\epsilon^2}$



$\frac{1}{\sqrt{m}}$

There is a truly brilliant method that runs in $O(n \log n)$ time.

Preview: Will involve Fast Fourier Transform in disguise.