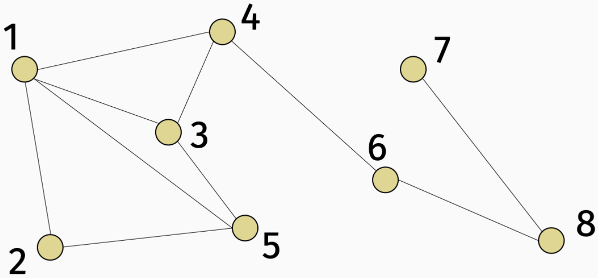
CS-GY 6763: Lecture 12
Spectral clustering, stochastic Block Model, subspace embeddings + ϵ-net arguments

NYU Tandon School of Engineering, Prof. Christopher Musco

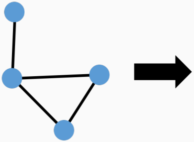
Main idea: Understand graph data by constructing natural matrix representations, and studying that matrix's spectrum (eigenvalues/eigenvectors).



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

MATRIX REPRESENTATIONS OF GRAPHS

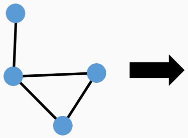
Two most common representations: $n \times n$ adjacency matrix $\mathbf{A}$ and graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ where $\mathbf{D}$ is the diagonal degree matrix.



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

Also common to look at normalized versions of both of these: $\bar{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ and $\bar{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$.

THE LAPLACIAN VIEW



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

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

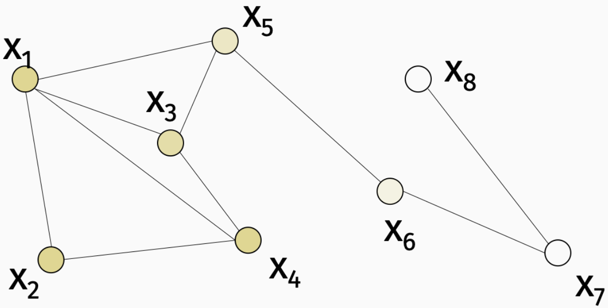
$\mathbf{B}$ has a row for every edge in G. The row for edge (i, j) has a $+1$ at position i, a -1 at position j, and zeros elsewhere.

Conclusions from $\mathbf{L} = \mathbf{B}^T\mathbf{B}$

- $\mathbf{L}$ is positive semidefinite: $\mathbf{x}^T\mathbf{L}\mathbf{x} \geq 0$ for all $\mathbf{x}$.
- $\mathbf{L} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T$ where $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ is $\mathbf{B}$'s SVD. Columns of $\mathbf{V}$ are eigenvectors of $\mathbf{L}$.
- For any vector $\mathbf{x} \in \mathbb{R}^n$,

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

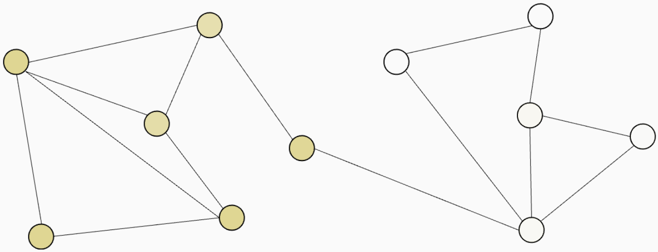
$x^T L x = \sum_{(i,j) \in E} (x(i) - x(j))^2$. So $x^T L x$ is small if x is a "smooth" function with respect to the graph.



Another conclusion from $\mathbf{L} = \mathbf{B}^T\mathbf{B}$:

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 $i \in T = V \setminus S$:

$$\mathbf{c}^T L\mathbf{c} = \sum_{(i,j)\in E} (\mathbf{c}(i) - \mathbf{c}(j))^2 = 4 \cdot \text{cut}(S, T). \tag{1}$$

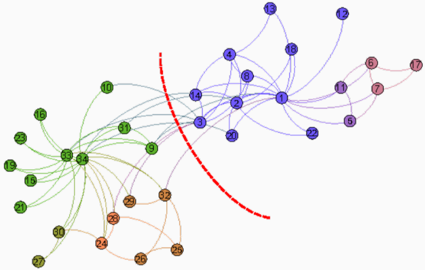


SPECTRAL GRAPH PARTITIONING

- Introduce NP-hard graph partitioning problem important in:
  - Understanding social networks.
  - Unsupervised machine learning (spectral clustering).
  - Graph visualization.
  - Mesh partitioning.
- See how this problem can be solved heuristically using Laplacian eigenvectors.
- Give an "average case" analysis of the method for a common random graph model.
- Use two tools: matrix concentration and eigenvector perturbation bounds.

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

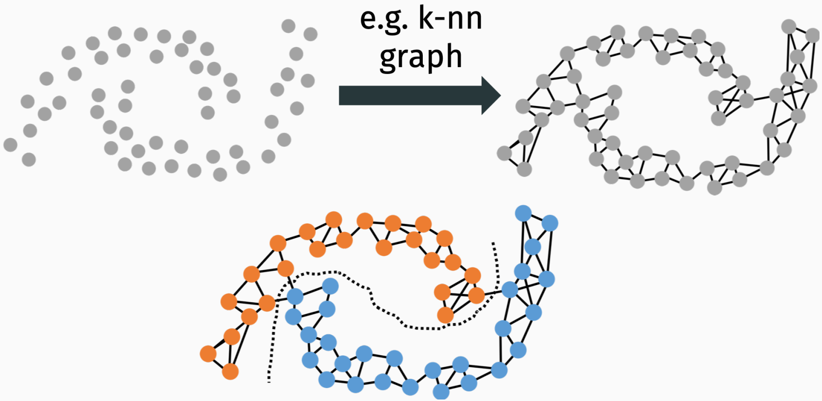
Example application: Understanding community structure in social networks.

Wayne W. Zachary (1977). An Information Flow Model for Conflict and Fission in Small Groups.

"At the beginning of the study there was an incipient conflict between the club president, John A., and Mr. Hi over the price of karate lessons. Mr. Hi, who wished to raise prices, claimed the authority to set his own lesson fees, since he was the instructor. John A., who wished to stabilize prices, claimed the authority to set the lesson fees since he was the club's chief administrator. As time passed the entire club became divided over this issue, and the conflict became translated into ideological terms by most club members."

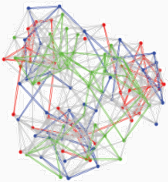
Zachary constructed a social network by hand and used a minimum cut algorithm to correctly predict who sided with who in the conflict. Beautiful paper – definitely worth checking out!

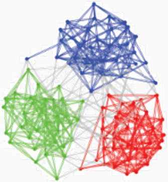
Idea: Construct synthetic graph for data that is hard to cluster.

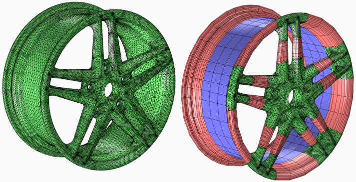


Spectral Clustering, Laplacian Eigenmaps, Locally linear embedding, Isomap, etc.

Balanced cut algorithms are also use in distributing data in graph databases, for partitioning finite element meshes in scientific computing (e.g., that arise when solving differential equations), and more.







Lots of good software packages (e.g. METIS).

There are many way's to formalize Zachary's problem:

β-Balanced Cut:

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

Sparsest Cut:

$$\min_S \frac{\text{cut}(S, V \setminus S)}{\min(|S|, |V \setminus S|)}$$

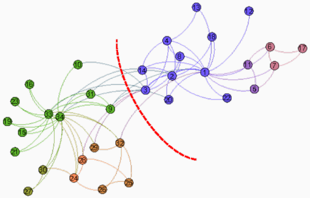
All natural formalizations lead to NP-hard problems. Lots of interest in designing polynomial time approximation algorithms, but tend to be slow. In practice, much simpler methods based on the graph spectrum are used.

Spectral methods run no more than $O(n^3)$ time (must faster if you use iterative methods for computing eigenvectors).

Basic spectral clustering method:

- 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.

This shouldn't make much sense yet! We will see that is a "relax and round" algorithm in disguise.

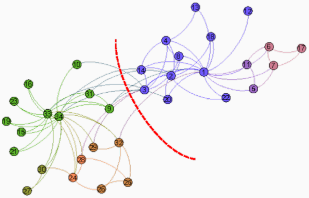


(a) Zachary Karate Club Graph

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 $i \in T$:

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

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



(a) Zachary Karate Club Graph

Equivalent formulation if we divide everything by $\sqrt{n}$ so that $\mathbf{c}$ has norm 1. Then $\mathbf{c} \in \{-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\}^n$ and:

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

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

Perfectly balanced balanced cut problem:

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

Relaxed perfectly balanced balanced cut problem:

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

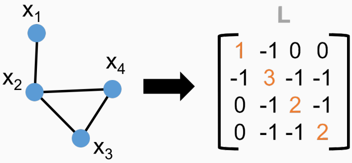
Claim: The relaxed problem is exactly minimized by the second smallest eigenvector $\mathbf{v}_{n-1}$ of $\mathbf{L}$.

Approach: Relax, find $\mathbf{v}_{n-1}$, then round back to a vector with $-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}$ entries.

Claim: The smallest eigenvector/singular vector of any graph Laplacian $\mathbf{L}$ always equals:

$$\mathbf{v}_n = \underset{v \in \mathbb{R}^n \text{ with } \|\mathbf{v}\|=1}{\arg\min} \mathbf{v}^T L \mathbf{v} = \frac{1}{\sqrt{n}} \cdot \mathbf{1}$$

with $\mathbf{v}_n^T L \mathbf{v}_n = 0$.



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

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

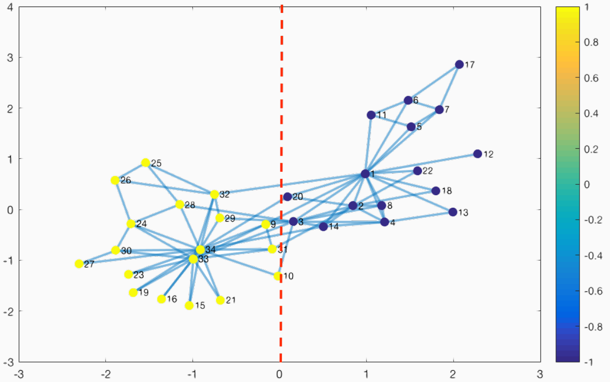
which is equivalent to

$$\mathbf{v}_{n-1} = \operatorname*{arg\,min}_{\|\mathbf{v}\|=1,\ \mathbf{1}^T\mathbf{v}=0} \mathbf{v}^T L \mathbf{v}.$$

Final relax and round algorithm: Compute

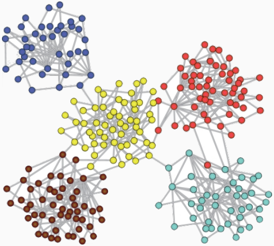
$$\mathbf{v}_{n-1} = \underset{v \in \mathbb{R}^n \text{ with } \|\mathbf{v}\|=1,\ \mathbf{v}^T\mathbf{1}=0}{\arg\min} \mathbf{v}^T L \mathbf{v}$$

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 $\mathbf{c} = \text{sign}(\mathbf{v}_{n-1})$



Lots of different variants used in practice:

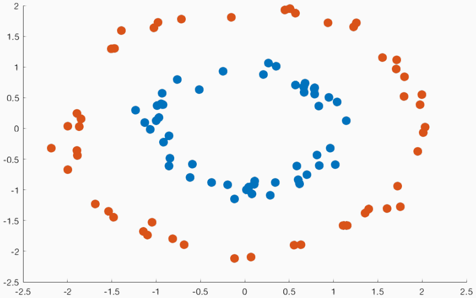
- Often do some sort of normalization of edge weights by degree. E.g. the Shi-Malik normalized cuts algorithm use the normalized Laplacian $\bar{\mathbf{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$.
- Different methods for how to choose the threshold to partition the second smallest eigenvector.
- Lots of variants to split the graph into more than two parts.



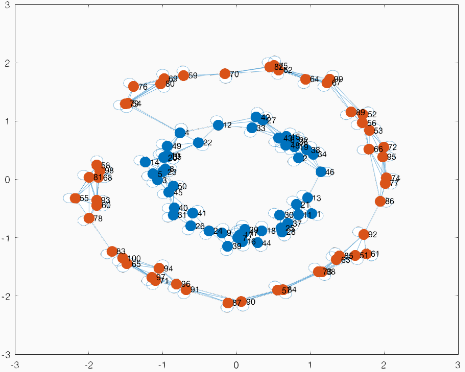
Multiway spectral partitioning:

- Compute smallest ℓ eigenvectors $\mathbf{v}_{n-1}, \ldots, \mathbf{v}_{n-\ell}$ of $\mathbf{L}$.
- Represent each node by its corresponding row in $\mathbf{V} \in \mathbb{R}^{n\times\ell}$ whose rows are $\mathbf{v}_{n-1}, \ldots \mathbf{v}_{n-\ell}$.
- Cluster these rows using k-means clustering (or really any clustering method).
- Often we choose $\ell = k$, but not necessarily.

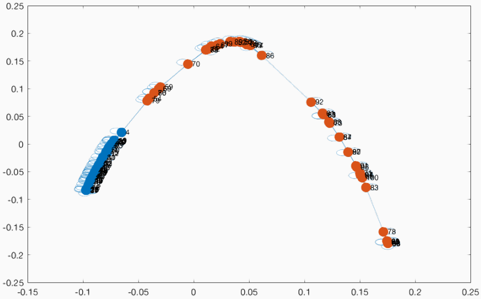
Original Data: (not linearly separable)



k-Nearest Neighbors Graph:



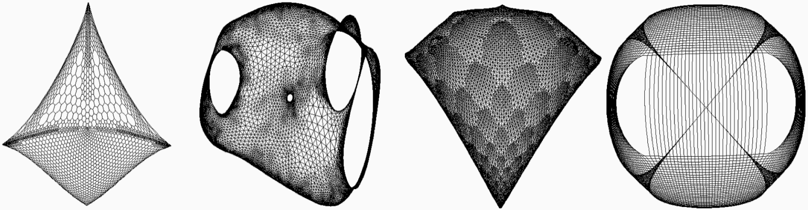
Embedding with eigenvectors v_{n-1}, v_{n-2}: (linearly separable)



Intuitively, since $\mathbf{v} \in \mathbf{v}_{n-1}, \ldots \mathbf{v}_{n-\ell}$ are smooth over the graph,

$$\sum_{i,j \in E} (\mathbf{v}[i] - \mathbf{v}[j])^2$$

is small for each coordinate. I.e. this embedding explicitly encourages nodes connected by an edge to be placed in nearby locations in the embedding.



Also useful e.g., in graph drawing.

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

- 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.

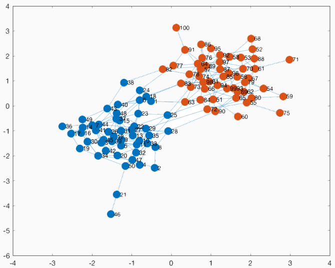
- Very common in algorithm design and analysis. Great way to start approaching a problem. Often our best way to understand why some algorithms "just work" in practice.
- Similar approach to Bayesian modeling in machine learning.

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

Stochastic Block Model (Planted Partition Model):

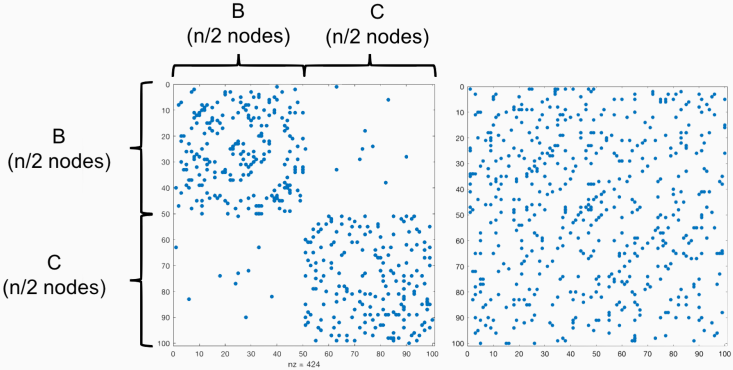
Let $G_n(p, q)$ be a distribution over graphs on n nodes, split equally into two groups B and C, each with $n/2$ nodes.

- 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. $q < p$.



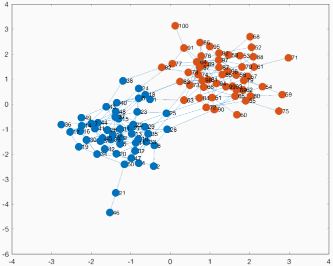
Let G be a stochastic block model graph drawn from $G_n(p, q)$.

- 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.

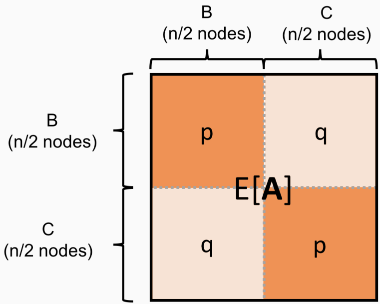


To do so, we need to understand the second smallest eigenvector of $\mathbf{L} = \mathbf{D} - \mathbf{A}$. We will start by considering the <u>expected value</u> of these matrices:

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

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. $(\mathbb{E}[\mathbf{A}])_{i,j} = p$ for i, j in same group, $(\mathbb{E}[\mathbf{A}])_{i,j} = q$ otherwise.



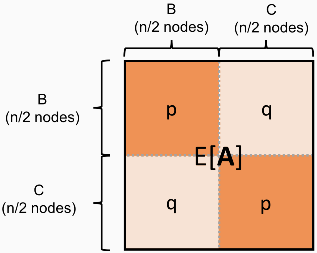
EXPECTED LAPLACIAN

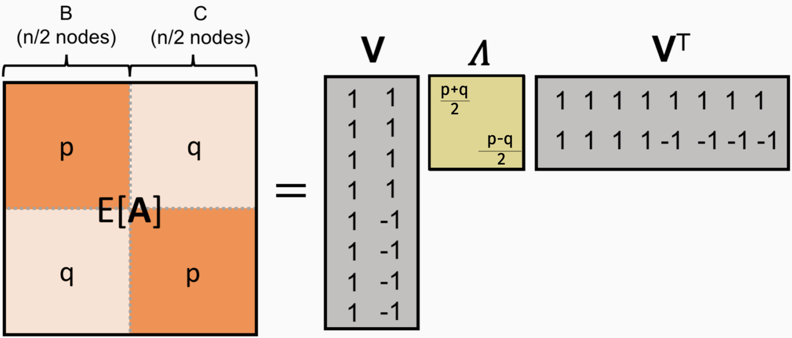
What is the expected Laplacian of $G_n(p, q)$?

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

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 $\mathbb{E}[\mathbf{A}]$?





- $\bar{\mathbf{v}}_1 \sim \mathbf{1}$ 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!

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.

Alon, Krivelevich, Vu, 2002:

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).

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

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

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 $\overline{\mathbf{A}}$.

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

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

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

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

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.

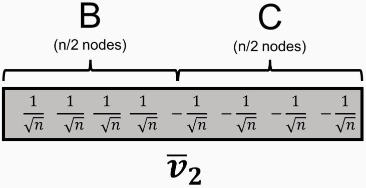
Claim 2 (Davis-Kahan): For $p \geq O\left(\frac{\log^4 n}{n}\right)$,

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

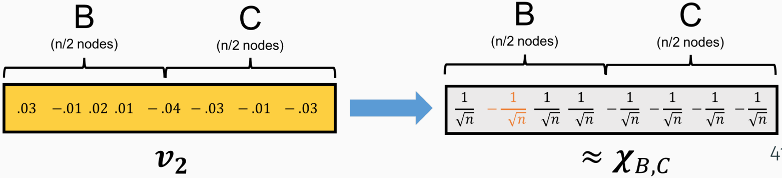
(A slightly trickier analysis can remove the qn term entirely.)

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

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



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



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.

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 $q = .8p$ 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.

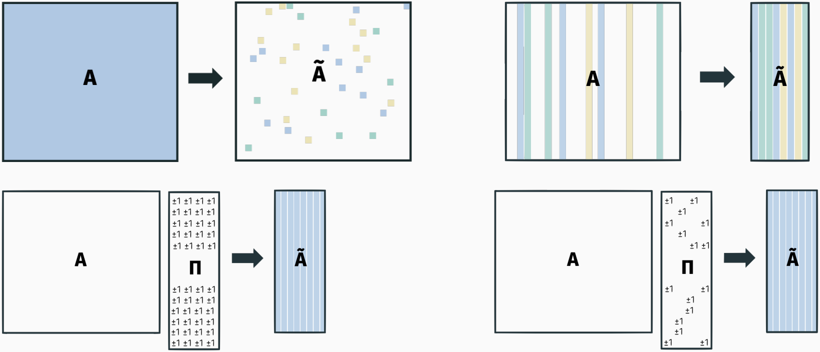
Forget about the previous problem, but still consider the matrix $\mathbf{M} = \mathbb{E}[\mathbf{A}]$.

- 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?

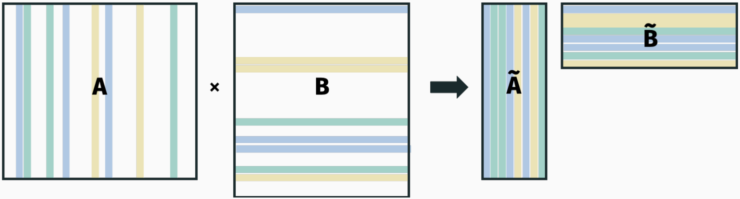
Main idea: If you want to compute singular vectors, multiply two matrices, solve a regression problem, etc.:

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 $\mathbf{A}$.

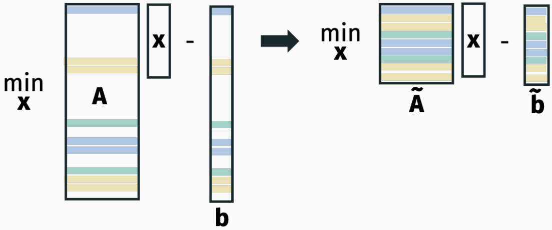


BREAK

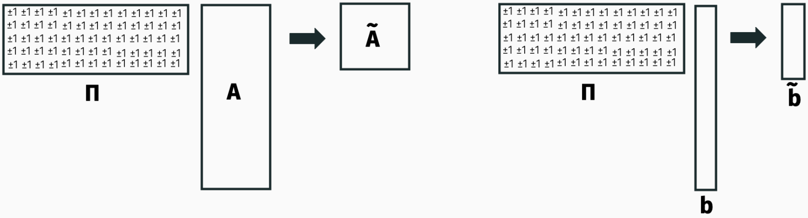
Approximate matrix multiplication:



Approximate regression:



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



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

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

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 part of lecture is as much about the technique as the final result.
- For the bonus problem on your last problem set you will use an ϵ-net argument to prove a matrix concentration inequality on your last 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.

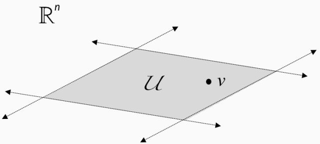
Note that all vectors of the form $(\mathbf{A}\mathbf{x} - \mathbf{b})$ lie in a low dimensional subspace: spanned by $d + 1$ vectors, where d is the width of $\mathbf{A}$. **So even though the set is infinite, it is "simple" in some way. Parameterized by just $d + 1$ numbers.**

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.

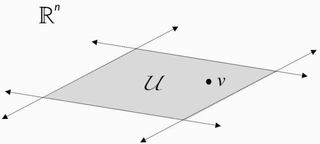
Proof: Apply Subspace Embedding Thm. to the $(d+1)$ dimensional subspace spanned by $\mathbf{A}$'s d columns and $\mathbf{b}$. Every vector $\mathbf{A}\mathbf{x}-\mathbf{b}$ lies in this subspace.

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 $1 - \delta$,

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

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 tons of other applications!

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

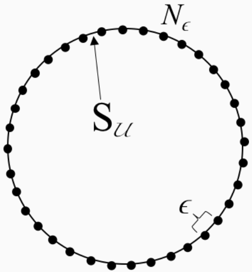
First Observation: The theorem holds as long as (2) holds for all $\mathbf{w}$ on the unit sphere in $\mathcal{U}$. Denote the sphere $S_{\mathcal{U}}$:

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

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

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



N_ϵ is called an "ϵ"-net.

If we can prove

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

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 $|N_\epsilon| = \left(\frac{4}{\epsilon}\right)^d$ 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.

1. Preserving norms of all points in net N_ϵ.

Set $\delta' = \frac{1}{|\mathcal{N}_\epsilon|} \cdot \delta = \left(\frac{\epsilon}{4}\right)^d \cdot \delta$. 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,

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

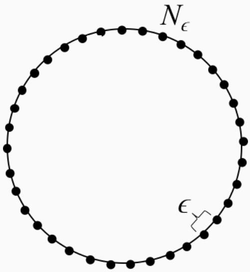
for <u>all</u> $\mathbf{w} \in N_\epsilon$,with probability $1-\delta$.

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



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:

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

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

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

$$\vdots$$

2. Extending to all points in the sphere.

Applying triangle inequality, we have that:

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

3. Preserving norm of v.

Similarly,

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

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 $1 - \delta$,

$$(1-\epsilon)\|\mathbf{v}\|_2^2 \leq \|\mathbf{\Pi}\mathbf{v}\|_2^2 \leq (1+\epsilon)\|\mathbf{v}\|_2^2 \tag{4}$$

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 $\mathbf{A}$.

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

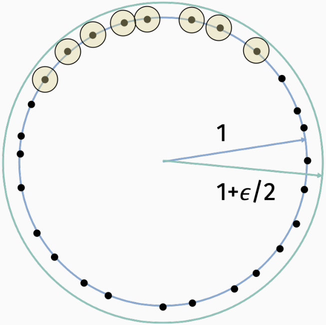
$$\min_{\mathbf{w} \in N_\epsilon} \|\mathbf{v} - \mathbf{w}\| \leq \epsilon.$$

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 of d dimensional ball of radius r is

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

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

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

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

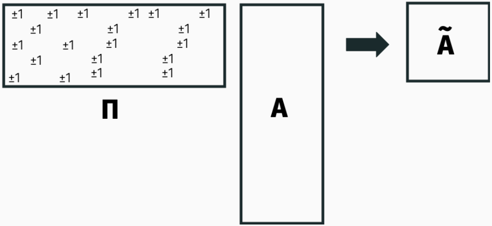
The trick is to show that a constant factor net is actually all that you need instead of an ϵ factor.

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.

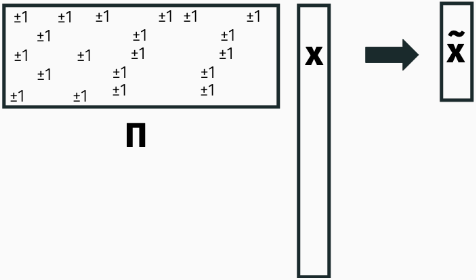


Typically using sparse and structured matrices.

Next class: 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:

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



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

Preview: Will involve Fast Fourier Transform in disguise.