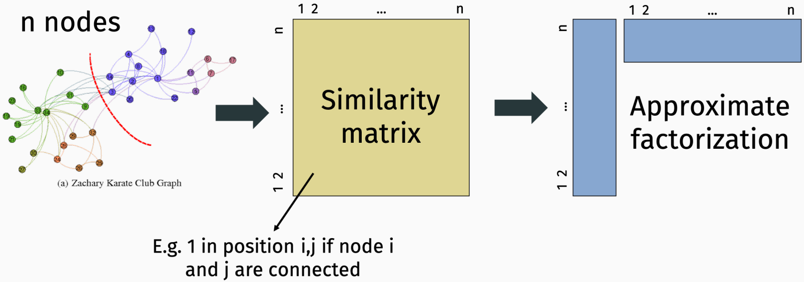
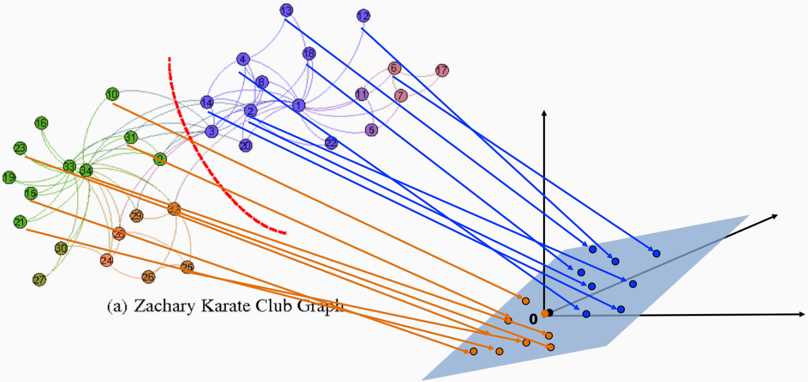
CS-GY 9223 I: Lecture 11
Spectral graph theory + randomized numerical linear algebra.

NYU Tandon School of Engineering, Prof. Christopher Musco

Often data is represented as a graph and similarities can be obtained from that graph:



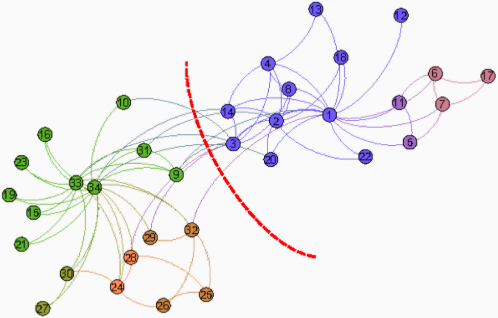
ENCODING GRAPH SIMILARITY



Spectral graph theory lets us formalize this heuristic idea.

Goal: 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.

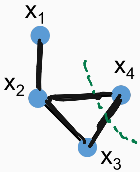


(a) Zachary Karate Club Graph

For a graph with adjacency matrix **A** and degree matrix **D**, **L** = **D** − **A** is the graph Laplacian.

$\|Bc\|_2^2 = \underline{c^T L c}$

$= c^T B^T B c$



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

$= c$

cut indicator vector $[1\ 1\ 1\ -1]$

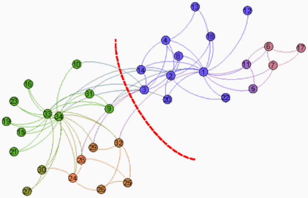
$\boxed{\mathbf{L} = \mathbf{B}^T\mathbf{B}}$ where B is the "edge-vertex incidence" matrix.

$$\mathbf{B} = \begin{array}{c|cccc} & \text{node1} & \text{node2} & 3 & 4.. \\ \hline \text{edges} & 1 & -1 & 0 & 0 \\ & 0 & 1 & -1 & 0 \\ & 0 & 1 & 0 & -1 \\ & 0 & 0 & 1 & -1 \end{array} \begin{bmatrix} \\ c \\ \\ \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 2 \end{bmatrix}$$

$Bc = \begin{cases} 0 & \text{for edges not across the cut} \\ 2 & \text{for edges across the cut} \end{cases}$

$\|Bc\|_2^2 = 4 \cdot$ (# edges across cut)

THE LAPLACIAN VIEW



(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 $i \in T$:

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

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

Courant–Fischer min-max principle

for symmetric matrix

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

$$\mathbf{v}_1 = \arg\max_{\|\mathbf{v}\|=1} \mathbf{v}^T \mathbf{L} \mathbf{v}$$

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

$$\mathbf{v}_3 = \arg\max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_1, \mathbf{v}_2} \mathbf{v}^T \mathbf{L} \mathbf{v}$$

$$\vdots$$

$$\mathbf{v}_n = \arg\max_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_1, \ldots, \mathbf{v}_{n-1}} \mathbf{v}^T \mathbf{L} \mathbf{v}$$

Courant–Fischer min-max principle

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

$$\begin{aligned}
\mathbf{v}_n &= \arg\min_{\|\mathbf{v}\|=1} \mathbf{v}^T \mathbf{L} \mathbf{v} \\
\mathbf{v}_{n-1} &= \arg\min_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_n} \mathbf{v}^T \mathbf{L} \mathbf{v} \\
\mathbf{v}_{n-2} &= \arg\min_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_n, \mathbf{v}_{n-1}} \mathbf{v}^T \mathbf{L} \mathbf{v} \\
&\vdots \\
\mathbf{v}_1 &= \arg\min_{\|\mathbf{v}\|=1, \mathbf{v} \perp \mathbf{v}_n, \ldots, \mathbf{v}_2} \mathbf{v}^T \mathbf{L} \mathbf{v}
\end{aligned}$$

The smallest eigenvector/singular vector $\mathbf{v}_n$ satisfies: for ~~any~~ L

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

$$\frac{1}{\sqrt{n}} \cdot \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

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

1) All of L's eigenvalues are ≥ 0.

L is positive semidefinite: $B^T B = L$.

$B^T B$

2) $L \cdot \vec{1} = B^T \underbrace{B \cdot 1}_{0} = B^T 0 = 0$.

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

$$L \cdot \left(\frac{1}{\sqrt{n}} 1\right) = 0 \left(\frac{1}{\sqrt{n}}\right) 1$$

$v_n = \frac{1}{\sqrt{n}} \cdot \mathbb{1}$

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

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

cut value for cut indicators

add constraint

$v_{n-1} \in \{ 1/\sqrt{n}, -1/\sqrt{n} \}^n$

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

- $\mathbf{v}_{n-1}^T L \mathbf{v}_{n-1} = cut(S, T)$ 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.

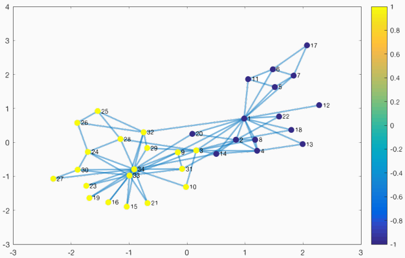
$\mathbf{v}_{n-1} \in \mathbb{R}^n$ is not generally binary, but still satisfies a 'relaxed' version of this property.

Find a good partition of the graph by using an eigendecomposition to 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 $\mathbf{v}_{n-1}(i) \geq 0$.

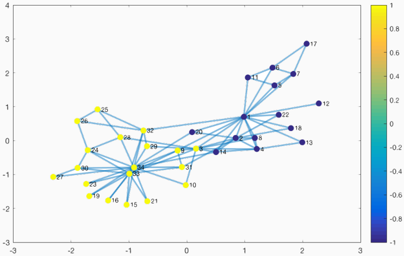
$$V_{n-1} = \begin{bmatrix} -.1 \\ .1 \\ -.5 \\ .4 \\ 1 \end{bmatrix}$$



Find a good partition of the graph by using an eigendecomposition to compute

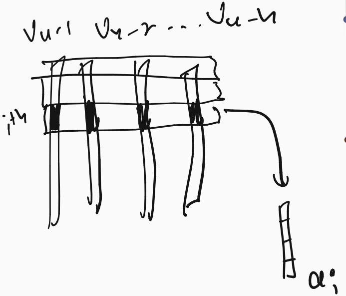
$$\mathbf{v}_{n-1} = \underset{v \in \mathbb{R}^n \text{ with } \|\mathbf{v}\|=1,\ \mathbf{v}^T\mathbf{1}=0}{\arg\min} \mathbf{v}^T L \mathbf{v}$$

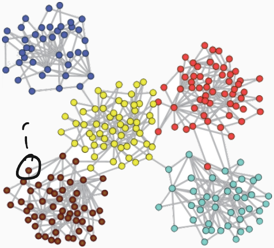
Set S to be all nodes with $\mathbf{v}_{n-1}(i) < 0$, and T to be all with $\mathbf{v}_{n-1}(i) \geq 0$.

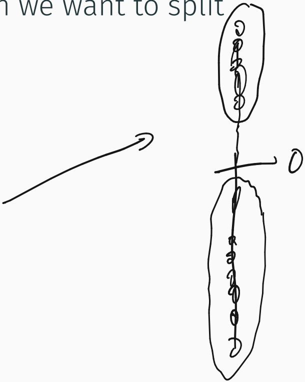


The Shi-Malik normalized cuts algorithm is one of the most commonly used variants of this approach, using the normalized Laplacian $\bar{\mathbf{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$.

Important consideration: What to do when we want to split the graph into more than two parts?





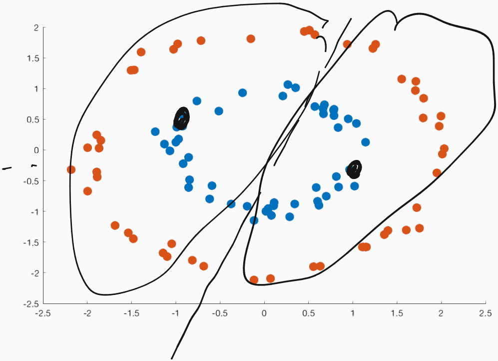


Spectral Clustering:

- Compute smallest k eigenvectors $\mathbf{v}_{n-1}, \ldots, \mathbf{v}_{n-k}$ of $\mathbf{L}$.
- Represent each node by its corresponding row in $\mathbf{V} \in \mathbb{R}^{n \times k}$ whose rows are $\mathbf{v}_{n-1}, \ldots \mathbf{v}_{n-k}$.
- Cluster these rows using k-means clustering (or really any clustering method).

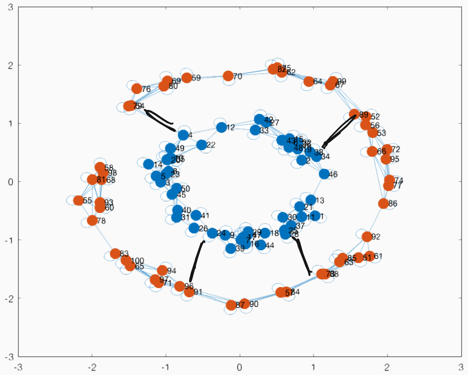
2-means

Original Data: (not linearly separable)

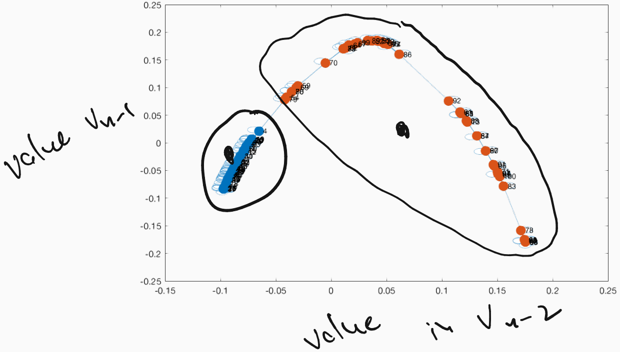


5 or 10

k-Nearest Neighbors Graph:



Embedding with eigenvectors $\mathbf{v}_{n-1}, \mathbf{v}_{n-2}$: (linearly separable)



So far: Spectral clustering partitions a graph along a small cut between large pieces.

- No formal guarantee on the 'quality' of the partitioning.
- Would be difficult to analyze for general input graphs.

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

- Very common in algorithm design for data analysis/machine learning (can be used to justify ℓ_2 linear regression, k-means clustering, PCA, etc.)

$a_i^T x = y_i$

$A = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$

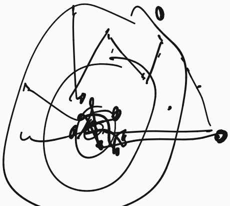
$x = \min \|Ax - b\|_{2,10}$

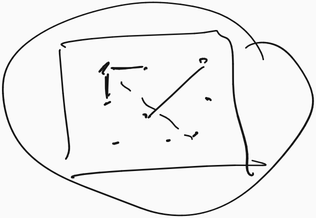
Ideas for a generative model for graphs that would allow us to understand partitioning?

n nodes

connect with prob
$p = .1$

add with prob
$p = .15$





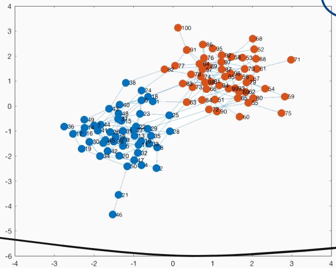
Stochastic Block Model (Planted Partition Model):

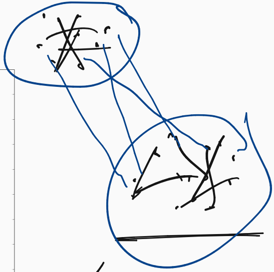
Let $G_n(p, q)$ be a distribution over graphs on n nodes, split equally into two groups B and C, each with $n/2$ nodes.

- Any two nodes in the same group are connected with probability p (including self-loops). $p = .1$
- Any two nodes in different groups are connected with prob. $q < p$.

Erdos-Renyi

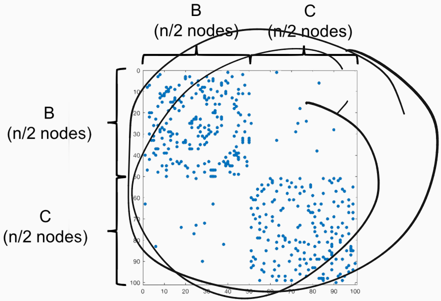
$G_{n/2}(p)$

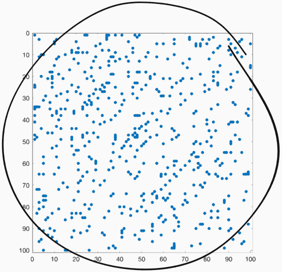




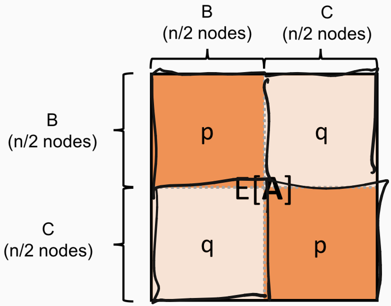
Let G be a stochastic block model graph drawn from $G_n(p, q)$.

- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be the adjacency matrix of G. What is $\mathbb{E}[\mathbf{A}]$?





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.



What are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{A}]$?

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, what are the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{A}]$?

$$\mathbb{E}[A] = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$$

$$\mathbb{E}[A] \cdot \mathbf{1} = \begin{bmatrix} \frac{(p+q)n}{2} \\ \vdots \\ \frac{(p+q)n}{2} \\ \vdots \\ (p+q)n/2 \end{bmatrix} = \underbrace{\frac{(p+q)n}{2}}_{\lambda} \mathbf{1}$$

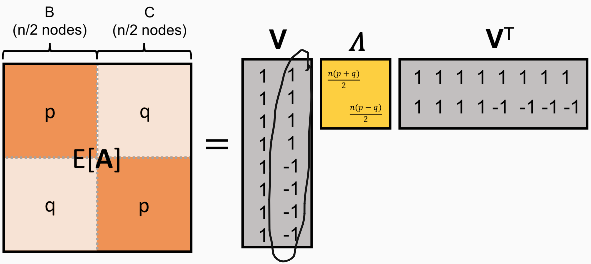
Eigenvector 1: $\frac{1}{\sqrt{n}} \mathbf{1}$

Eigenvalue 1: $\frac{(p+q)n}{2}$

Eigenvector 2: $\frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix} \to u$

$$\mathbb{E}[A] \cdot u = \begin{bmatrix} (p-q)n/2 \\ \vdots \\ (p-q)n/2 \\ (q-p)n/2 \\ \vdots \end{bmatrix} = \frac{(p-q)n}{2} u$$

EXPECTED ADJACENCY SPECTRUM



- $\mathbf{v}_1 = \mathbf{v}_1$ 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!

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

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

random

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

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

vector

For $X \in \mathbb{R}^{n \times d}$, $\|X\|_2 = \max_{z \in \mathbb{R}^d : \|z\|_2 = 1} \|Xz\|_2$. max $|z^T A z|$

Exercise: Show that $\|X\|_2$ is equal to the largest singular value of X. For symmetric X (like $A - \mathbb{E}[A]$) show that it is equal to the magnitude of the largest magnitude eigenvalue.

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

1974

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:

measure of closeness between $v_i, \bar{v}_i$

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

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

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

EIGENVECTOR PERTURBATION

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

$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\quad v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$\lambda_1 = 1+\varepsilon \quad \lambda_2 = 1$

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

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

$\lambda_1 = 1+\varepsilon \quad \lambda_2 = 1$

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

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

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

$$\sin\theta(v_2, \bar{v}_2) \leq \frac{O(\sqrt{pn})}{\min_{j \neq 2} |\lambda_2 - \lambda_j|} \leq \frac{O(\sqrt{pn})}{(p-q)n/2} = \boxed{O\left(\frac{\sqrt{p}}{(p-q)\sqrt{n}}\right)}$$

$i = 2$

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.

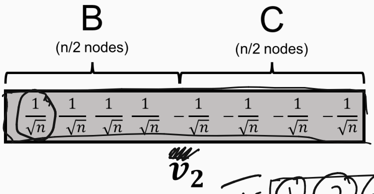
qn

$\frac{(p+q)n}{2}$ $\frac{(p-q)n}{2}$ 0 0 0 0 0 $\frac{(p-q)n}{2}$

APPLICATION TO STOCHASTIC BLOCK MODEL

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



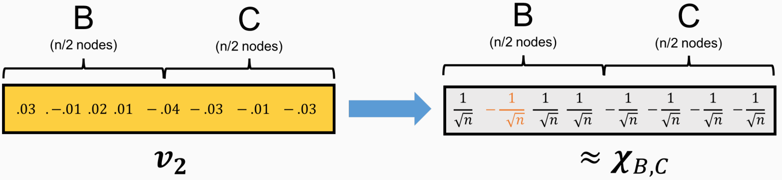
$v_2 = [.1 \;\; .2 \;\; -.001 \;\; \cdots \;\; -.1 \;\; -.5]$

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

$\|\bar{v}_2(i) - v_2(i)\|_2$

- So they differ in sign in at most $O\left(\frac{p}{(p-q)^2}\right)$ positions.

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



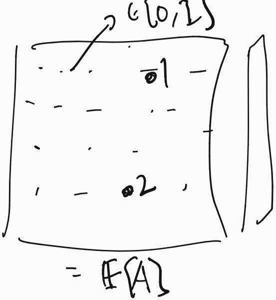
- Why does the error increase as q gets close to p?
- Even when $p - q = O(1/\sqrt{n})$, assign all but an $O(n)$ fraction of nodes correctly. E.g., assign 99% of nodes correctly.

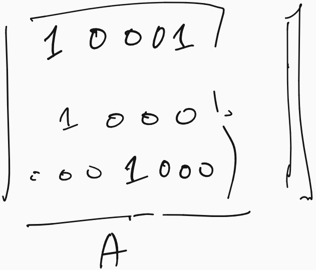
$p - q = O(1/\sqrt{n}) \Rightarrow \frac{p}{(p-q)^2} = .01\,n$

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?.





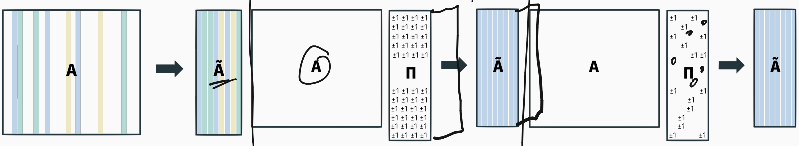
Main idea: If you want to compute singular vectors or eigenvectors, multiply two matrices, solve a regression problem, etc.:

sampling

1. Compress your matrices using a randomized method.
2. Solve the problem on the smaller or sparser matrix.
   - Ã called a "sketch" or "coreset" for **A**.

sparse JL matrix

column sampling



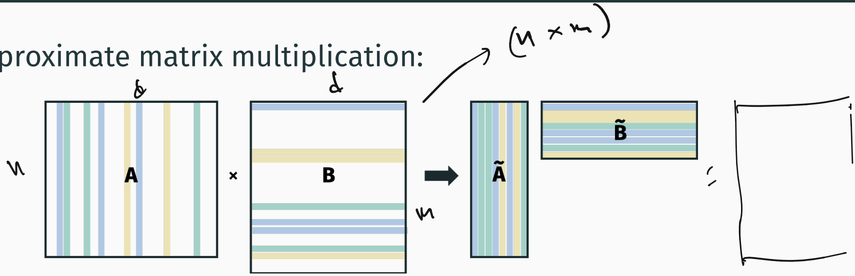
AA^T

$\tilde{A}\tilde{A}^T$

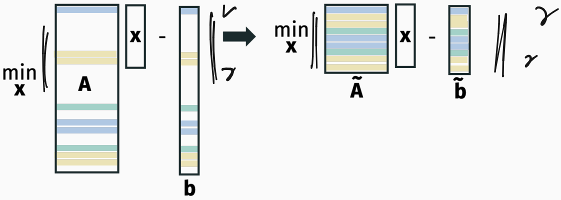
compute approximate left singular vectors

random matrix

Approximate matrix multiplication:

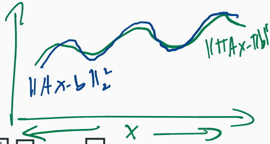


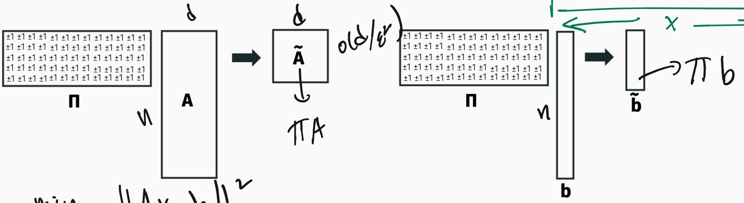
Approximate regression:



	Direct	Iterative	Randomized
Method:	Gaussian Elimination QR Algorithm	Gradient Descent + Power Method + Krylov Methods	sketch ↓ sampling, random projection
Speed:	$O(n^3)$ for inverse, eigendecomp Slow	$O(n^2) \cdot$ (# of iterations) Fast	$\approx O(n^2)$ + accuracy, matrix params etc. Faster (sometimes)
Accuracy:	Exact $\log\log(1/\varepsilon)$	$\log(1/\varepsilon)$ $1/\varepsilon$ $\quad 1/\sqrt{\varepsilon}$	$\frac{1}{\varepsilon^2}$ $\quad 1/\varepsilon$

Randomized approximate regression using a Johnson-Lindenstrauss Matrix:





$\min \|Ax-b\|_2^2$

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

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

Goal: Want ~~$\min_{\mathbf{x}} \|\mathbf{\Pi A x} - \mathbf{\Pi b}\|_2^2 \le (1+\epsilon)\min_{\mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|_2^2$~~

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

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

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

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* $\mathbf{y}$*,*

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

with probability $(1-\delta)$.

$y = Ax - b$

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

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

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

This statement requires establishing a Johnson-Lindenstrauss type bound for an infinity of possible vectors $(\mathbf{A}\mathbf{x} - \mathbf{b})$, which obviously can't be tackled with a union bound argument.

$x_1 \quad x_2 \quad \cdots \quad x_{100} \quad \cdots$

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

Then for each x_i

$\|Ax_i - b\|_2^2 = (1 \pm \varepsilon)\|\Pi A x_i - \Pi b\|_2^2$

with prob. $1-\delta$,

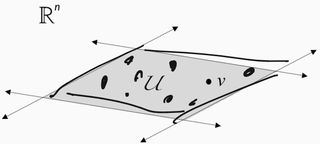
So with prob $1-100\delta$ this holds for $x_1, \ldots, x_{100}$

Theorem (Subspace Embedding from JL)

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

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

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



[1]It's possible to obtain a slightly tighter bound of $\boxed{O\left(\frac{d+\log(1/\delta)}{\epsilon^2}\right)}$. 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, then with high probability,

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

for all $\mathbf{x}$ and thus

for any y that can be written as $Ax-b$, want $(1-\epsilon)\|y\|_2^2 \leq \|\Pi y\|_2^2 \leq (1+\epsilon)\|y\|_2^2$

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

I.e. we can solve linear regression approximately using the $O\left(d/\epsilon^2\right) \times d$ matrix $\mathbf{\Pi}\mathbf{A}$ in place of $\mathbf{A}$.

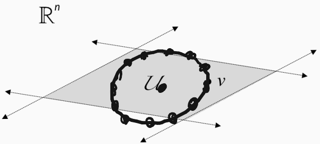
Proof: Apply Subspace Embedding Thm. to the $(d+1)$ dimensional subspace spanned by $\mathbf{A}$'s d columns and $\mathbf{b}$. Every vector $\mathbf{A}\mathbf{x}-\mathbf{b}$ lies in this subspace.

Theorem (Subspace Embedding from JL)

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

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

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

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

$c \cdot w$

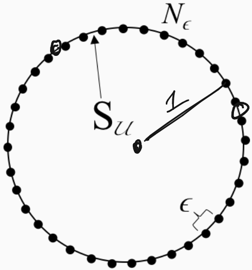
$v = \|v\|_2 \cdot \frac{v}{\|v\|_2}$

$\in S_{\mathcal{U}}$

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

$v \qquad v \qquad v$

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



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

$O(\frac{1}{\epsilon^2})$

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

N_ϵ is called an "ϵ"-net.

If we can prove

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

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

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

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

Want (1) $\to (1-\varepsilon)\|w\|_2^2 \le \|\Pi w\|_2^2 \le \|w\|_2^2 \cdot (1+\varepsilon)$

to hold for all $w \in N_\varepsilon$.

Set $\delta = \dfrac{1}{|N_\varepsilon|}$ $\quad \log(1/\delta) \quad \log(|N_\varepsilon|)$

$= \log\left(4/\varepsilon\right)^d$

$= d \log(4/\varepsilon)$

1. Set $\delta = \left(\frac{\epsilon}{8}\right)^d$. By a union bound, with high probability, for all $\mathbf{w} \in N_\epsilon$,

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

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

2. Consider any $\mathbf{v} \in S_{\mathcal{U}}$. You can check that, for some $\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2 \ldots \in N_\epsilon$, v can be written:

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

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

3. Applying triangle inequality, we have

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

4. Similarly,

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

So we have proven

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

for all v in $S_{\mathcal{U}}$.

Adjusting ϵ proves the Subspace Embedding theorem.