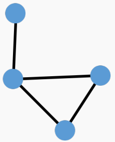
CS-GY 6763: Lecture 12
Stochastic Block Model, Randomized numerical linear algebra, ϵ-net arguments.

NYU Tandon School of Engineering, Prof. Christopher Musco

LAST CLASS

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



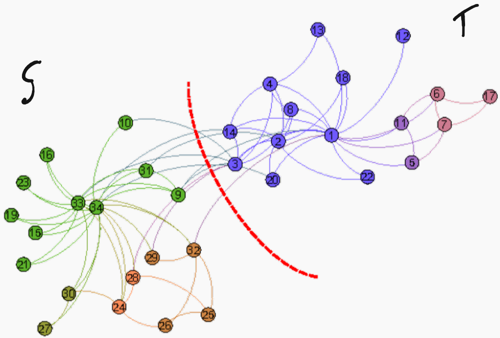
$$\overset{\mathbf{D}}{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}} - \overset{\mathbf{A}}{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}} = \overset{\mathbf{L}}{\begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 3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

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

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

Balanced Cut: Partition nodes along a cut that:

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



(a) Zachary Karate Club Graph

$c = [+1 \ -1 \ +1 \ +1 \ -1]$

n

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

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

Note: $\mathbf{c}$ often denote by $\chi_{S,T}$ to remind us what the cut is. And recall that we always have $S = V \setminus S$.

$|S| = |T| \qquad c^T \mathbb{1} = 0$

"Relax and round" algorithm: Perfectly balanced cut problem

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

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

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.
- Difficult to analyze for general 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.
- This is also the whole idea behind Bayesian Machine Learning (can be used to justify ℓ_2 linear regression, k-means clustering, PCA, etc.)

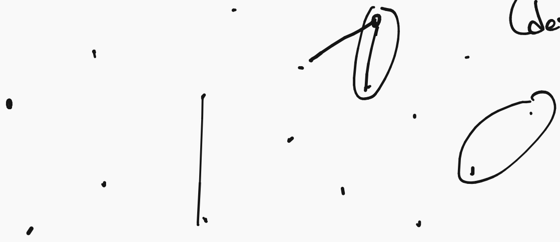
STOCHASTIC BLOCK MODEL

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

- ~~Small number of nodes w/ large number of connections~~

- Small well-connected communities.

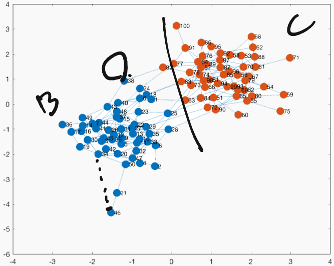
$(deg_i + 1)(deg_j + 1)$



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.

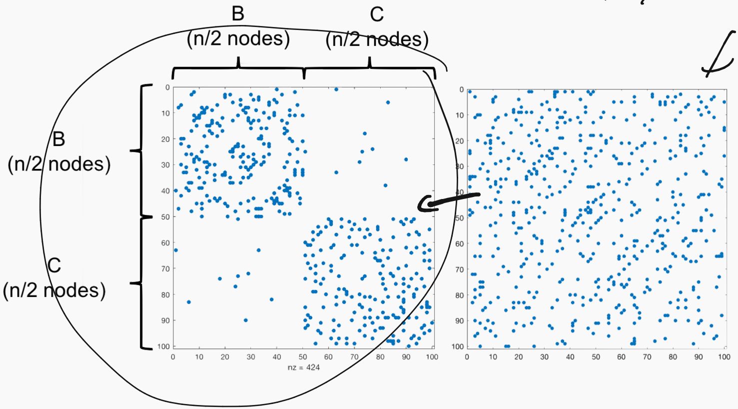
- Any two nodes in the **same group** are connected with probability p (including self-loops).
- Any two nodes in **different groups** are connected with prob. $q < p$.



LINEAR ALGEBRAIC VIEW

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

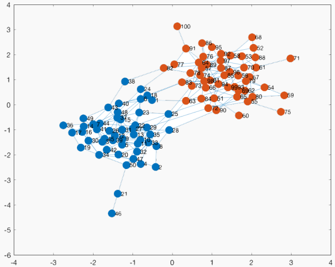
- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be the adjacency matrix of G. ~~What is $\mathbb{E}[\mathbf{A}]$?~~



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

STOCHASTIC BLOCK MODEL

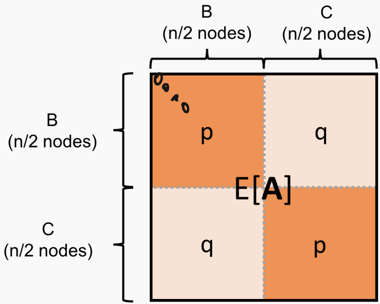
Goal is to find the "ground truth" balanced partition B, C using our standard spectal method.



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

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

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.



We are going to determine the eigenvectors and eigenvalues of $\mathbb{E}[\mathbf{A}]$.

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

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

v = eigenvector of $\mathbb{E}[A]$

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

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

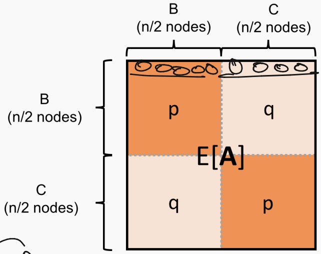
$\lambda_1, \ldots, \lambda_n$

$c-\lambda_1, \ldots, c-\lambda_n$

$\mathbb{E}$[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]

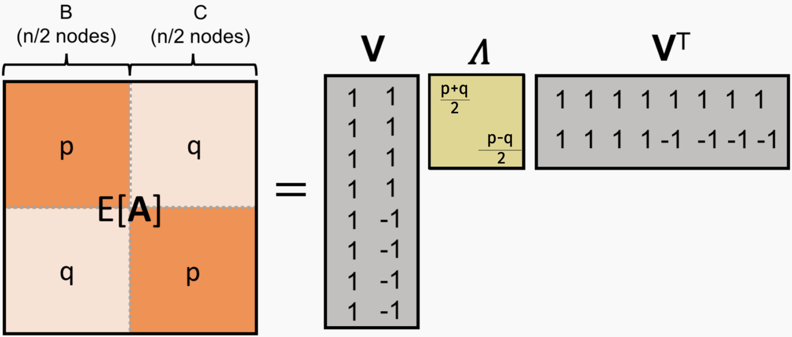
EXPECTED ADJACENCY SPECTRUM

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



$$\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{n}{2}p + \frac{n}{2}\cdot q \\ \vdots \\ \vdots \end{bmatrix} = \underbrace{\frac{p+q}{2}\cdot n}\cdot \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \\ \hline -1 \\ \vdots \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{n}{2}p - \frac{n}{2}q \\ \vdots \\ \hline \frac{n}{2}q - \frac{n}{2}p \end{bmatrix} = \boxed{\frac{p-q}{2}\cdot n} \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \hline -1 \\ \vdots \\ -1 \end{bmatrix}$$



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

If we compute $\mathbf{v}_2$ then we 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.

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

MATRIX CONCENTRATION

Alon, Krivelevich, Vu, 2002:

$$\begin{bmatrix} 1+\delta & 0 \\ 0 & 1-\delta \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1+\delta-\varepsilon & \\ & 1-\delta+\varepsilon \end{bmatrix}$$

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.

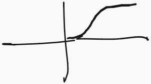
EIGENVECTOR PERTURBATION

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 $\bar{\mathbf{A}}$.



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

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

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

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

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

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

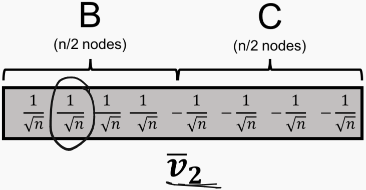
Assume $\frac{(p-q)n}{2}$ will be the minimum of these two gaps.

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

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?

→ true community indicator vector

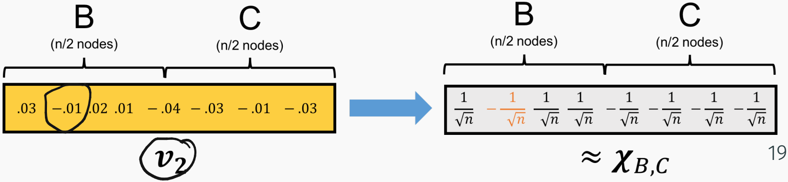
- Can show that this implies $\|\mathbf{v}_2 - \bar{\mathbf{v}}_2\|_2^2 \leq O\left(\frac{p}{(p-q)^2 n}\right)$ (exercise).
- $\bar{\mathbf{v}}_2$ is $\frac{1}{\sqrt{n}}\chi_{B,C}$: the community indicator vector.



$\frac{1}{\sqrt{n}} - (-.01)$

$> \frac{1}{\sqrt{n}}$

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



$p = \frac{c}{n}$ $q = \frac{c/2}{n}$

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.

$$\frac{c/n}{((c/2)/n)^2} = \frac{4}{c} n$$

$$\frac{c/n}{(c^2/4)/n^2} = \frac{4c}{c^2} n$$

$\frac{4}{100} \cdot n$

$\approx 4\%$

Upshot: If G is a stochastic block model graph with adjacency matrix $\mathbf{A}$, if we compute its second large 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:** $p = c/n$ for some factor c. Even when $p - q = O(1/n)$, assign all but an $O(n)$ fraction of nodes correctly. E.g., assign 99% of nodes to the right cluster.

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

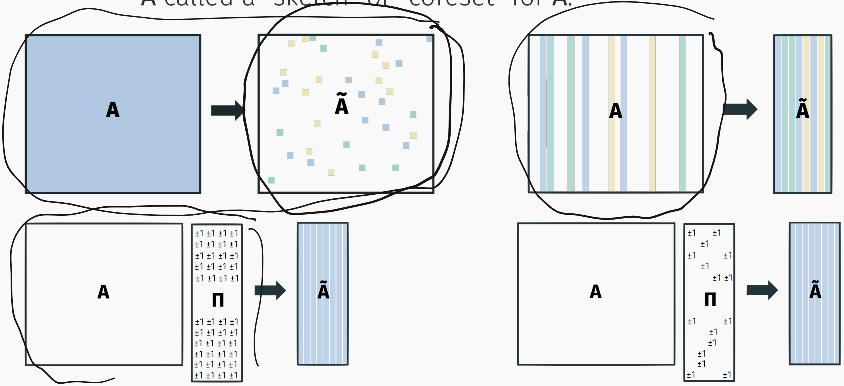
- Dense $n \times n$ matrix.
- Computing top eigenvectors takes $\approx O(n^2/\sqrt{\epsilon})$ time.

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

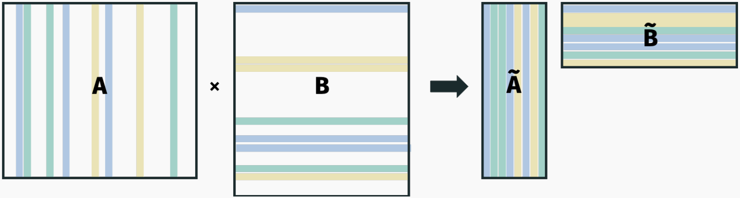
$O(n/\sqrt{\epsilon})$

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

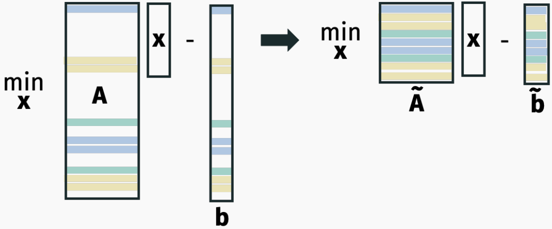
1. Compress your matrices using a randomized method (e.g. subsampling).
2. Solve the problem on the smaller or sparser matrix.
   - $\tilde{\mathbf{A}}$ called a "sketch" or "coreset" for $\mathbf{A}$.



Approximate matrix multiplication:

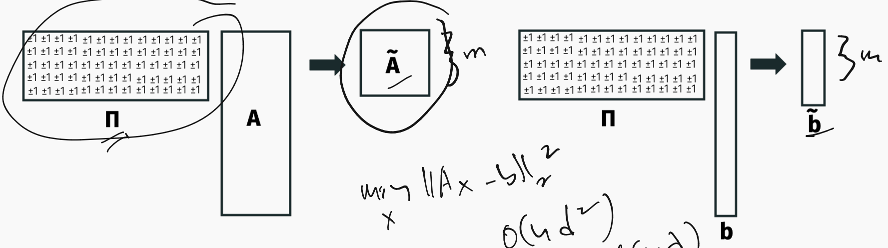


Approximate regression:



Today's example: Randomized approximate regression using a Johnson-Lindenstrauss Matrix.

$O(nd^2)$ $O(nd)$



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

$O(nd^2)$ $O(nd)$

$\|\tilde{A}x - \tilde{b}\|_2^2$

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

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

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

For fixed x $\quad \|Ax-b\|_2^2 \approx (1\pm\varepsilon)\|\tilde{A}x-\tilde{b}\|_2^2$

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

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

PLAN

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

$$= \Pi(Ax^* - b)$$

- 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.
- 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 $\mathbf{x} \in \mathbb{R}^d$, fixed

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

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

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

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

Lemma (Distributional JL)

If $\mathbf{\Pi}$ *is chosen to a properly scaled random Gaussian matrix, sign matrix, sparse random matrix, etc., with* $O\left(\frac{\log(1/\delta}{\epsilon^2}\right)$ *rows then for any fixed* $\mathbf{y}$,

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

with probability $(1-\delta)$.

$y = Ax - b$

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

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

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

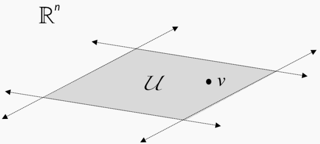
Note that all vectors of the form $(\mathbf{A}\mathbf{x} - \mathbf{b})$ lie in a low dimensional subspace: spanned by $d + 1$ vectors, where d is the width of $\mathbf{A}$. So even though the set is infinite, it is "simple" in some way. Parameterized by just $d + 1$ numbers.

Theorem (Subspace Embedding from JL)

Let $\mathcal{U} \subset \mathbb{R}^n$ be a d-dimensional linear subspace in $\mathbb{R}^n$. If $\mathbf{\Pi} \in \mathbb{R}^{m \times \cancel{d}\, n}$ 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.

SUBSPACE EMBEDDING TO APPROXIMATE REGRESSION

Corollary: If we choose $\mathbf{\Pi}$ and properly scale, then with $O\left(d/\epsilon^2\right)$ rows, $(d+1)$

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

for all $\mathbf{x}$ and thus $\quad m = O\left(\frac{d+1}{\epsilon^2}\right)$

$$\underbrace{\|\mathbf{A}\tilde{\mathbf{x}}-\mathbf{b}\|_2^2 \leq (1+O(\epsilon))\min_{\mathbf{x}}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2.}_{\text{I.e., our main theorem is proven.}} \quad = O(d/\epsilon)$$

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.

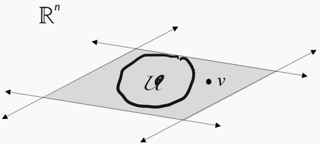
SUBSPACE EMBEDDINGS

Theorem (Subspace Embedding from JL)

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

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

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



Subspace embeddings have tons of other applications!

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

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

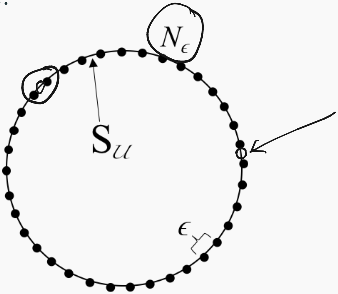
$$S_{\mathcal{U}} = \{\mathbf{w} \mid \mathbf{w} \in \mathcal{U} \text{ and } \|\mathbf{w}\|_2 = 1\}.$$

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

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

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

$\log(N_\epsilon) = d\log(1/\epsilon)$



N_ϵ is called an "ϵ"-net. if for all $y \in S_{\mathcal{U}}$, $\exists\, w \in N_\epsilon$ such that $\|y - w\|_2 \le \epsilon$.

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 $|N_\epsilon| \leq \left(\frac{4}{\epsilon}\right)^d$ 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,

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

$1-\delta$

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

$1-\delta'$ $\quad \delta' = \frac{1}{|N_\varepsilon|} \cdot \delta = \left(\frac{\varepsilon}{4}\right)^d \cdot \delta$

$\frac{\log(1/\delta')}{\varepsilon^2} \to \log(1/\delta) + \log\left(\frac{4}{a}^d\right) = \log(1/\delta) + d\log(4/\omega)$

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

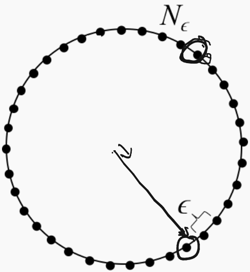
2. Writing any point in sphere as linear comb. of points in N_ϵ.

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

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

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

$|c_1| \leq \varepsilon \quad |c_2| \leq \varepsilon^2 \quad \ldots$



$$w_0 = \underset{w \in N_\varepsilon}{\arg\min} \|w - v\|_2$$

$$r_0 = w_0 - v$$

$$\|r_0\|_2 \leq \varepsilon$$

$$w_1 = \underset{w \in N_\varepsilon}{\arg\min} \left\| w - \frac{r_0}{\|r_0\|_2} \right\|_2$$

$$r_0 \simeq w_1 \|r_0\|_2$$

$$v = w_0 + \cancel{r_0}\ \|r_0\|_2 w_1 + r_1$$

$$\|r_1\| \leq \varepsilon \|r_0\|_2 \leq \varepsilon^2$$

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

3. Preserving norm of v.

$\|v\|_2 = 1$

Applying triangle inequality, we have that:

$$
\begin{aligned}
\|\boldsymbol{\Pi}\mathbf{v}\|_2 &= \|\boldsymbol{\Pi}\mathbf{w}_0 + c_1\boldsymbol{\Pi}\mathbf{w}_1 + c_2\boldsymbol{\Pi}\mathbf{w}_2 + \ldots\| \\
&\leq \|\boldsymbol{\Pi}\mathbf{w}_0\| + c_1\|\boldsymbol{\Pi}\mathbf{w}_1\| + c_2\|\boldsymbol{\Pi}\mathbf{w}_2\| + \ldots \\
&\leq \|\boldsymbol{\Pi}\mathbf{w}_0\| + \epsilon\|\boldsymbol{\Pi}\mathbf{w}_1\| + \epsilon^2\|\boldsymbol{\Pi}\mathbf{w}_2\| + \ldots \quad \epsilon^3 \ldots \\
&\leq (1+\epsilon) + \epsilon(1+\epsilon) + \epsilon^2(1+\epsilon) + \ldots \\
&\leq \cancel{1+2\epsilon} \\
& \quad 1+5\epsilon
\end{aligned}
$$

$(1+\epsilon) + (1+\epsilon)(\epsilon+\epsilon^2+\ldots)$

2ϵ

$(1+\epsilon)+(1+\epsilon)\,2\epsilon$

2

$\leq 1+5\epsilon$

3. Preserving norm of v.

Similarly,

$$
\begin{aligned}
\|\mathbf{\Pi v}\|_2 &= \|\mathbf{\Pi w}_0 + c_1\mathbf{\Pi w}_1 + c_2\mathbf{\Pi w}_2 + \ldots\| \\
&\geq \|\mathbf{\Pi w}_0\| - \epsilon\|\mathbf{\Pi w}_1\| - \epsilon^2\|\mathbf{\Pi w}_2\| - \ldots \\
&\geq (1-\epsilon) - \epsilon(1+\epsilon) - \epsilon^2(1+\epsilon) - \ldots \\
&\geq 1 - 5\epsilon.
\end{aligned}
$$

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 $1 - \delta$,

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

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

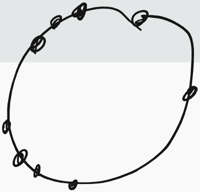
Subspace embeddings have many other applications!

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

Lemma (ϵ-net for the sphere)

For any $\epsilon \leq 1$, there exists a set $N_\epsilon \subset S_{\mathcal{U}}$ with $|N_\epsilon| = \left(\frac{3}{\epsilon}\right)^d$ such that $\forall \mathbf{v} \in S_{\mathcal{U}}$,

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

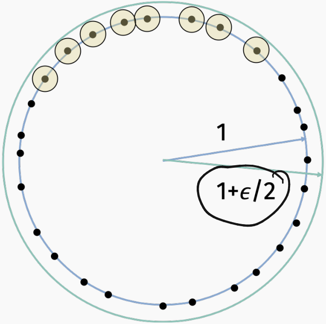


Imaginary algorithm for constructing N_ϵ:

- 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

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

where c is a constant that depends on d, but not r. From

previous slide we have:

$$\mathrm{vol}(d,\epsilon/2) \cdot |N_\epsilon| \le \mathrm{vol}(d, 1+\epsilon/2) \cdot 1$$

$$|N_\epsilon| \le \frac{\mathrm{vol}(d,1+\epsilon/2)}{\mathrm{vol}(d,\epsilon/2)} = \frac{\cancel{c}\,(1+\epsilon/2)^d}{\cancel{c}\,(\epsilon/2)^d}$$

$$\le \left(\frac{1+\epsilon/2}{\epsilon/2}\right)^d \le \left(\frac{3}{\epsilon}\right)^d$$

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

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.

RUNTIME CONSIDERATION

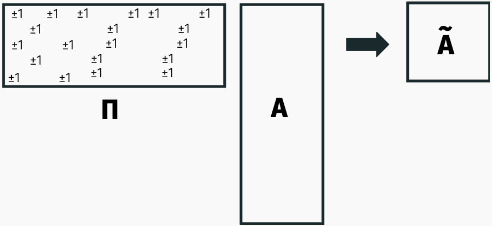
For $\epsilon, \delta = O(1)$, we need $\mathbf{\Pi}$ to have $m = O(d)$ rows.

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

$O(nd^2)$

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.

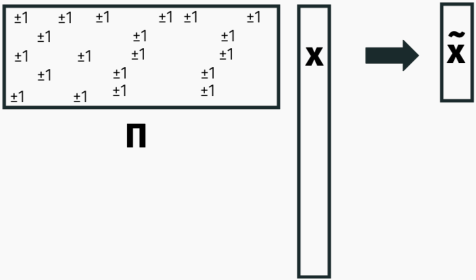


Typically using sparse and structured matrices.

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.