Lecture 13

Last time

- Proof of Davis-Kahan $\sin\Theta$.
- Wedin's theorem
- Community detection

Today

- Community detection continued
- Nets, coverings, and packings.

Community Detection Continued

In expectation this matrix has a block structured. Assuming the first community is $[n/2]$, we have

$$\mathbb{E}A = \left[\begin{array}{ccc|ccc} p & \cdots & p & q & \cdots & q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \cdots & p & q & \cdots & q \\ \hline q & \cdots & q & p & \cdots & p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q & \cdots & q & p & \cdots & p \end{array}\right]$$

Check that this matrix is rank 2 with $\lambda_1 = \frac{p+q}{2} n$, $\lambda_2 = \frac{p-q}{2} n$,

$$u_1 = \frac{1}{\sqrt{n}}\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{and} \quad u_2 = \frac{1}{\sqrt{n}}\begin{bmatrix} \left.\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}\right\} n/2 \\ -1 \\ \vdots \\ -1 \end{bmatrix}$$

Key Insight: Thus, if A is close to $\mathbb{E}A$ we could use its second eigenvector to identify the communities.

Spectral Clustering Algorithm

Input: Graph G

Step 1: Compute adjecency matrix A

Step 2: Compute $u_2(A)$ ← eigenvalue associated with 2nd largest λ_i

Step 3: Return $\operatorname{sign}(u_2(A))$.

We shall prove the following theorem

Theorem: Let $G \sim G(n,p,q)$ and $q \wedge (p-q) =: \mu > 0$. Then, with probability at least $1-4e^{-n}$, the spectral clustering algorithm identifies the communities of G with at most C/μ^2 misclassified vertices. ⊣

C ← Universal constant

Remark:

- In HW3 you will get rid of the dependency in q from μ.
- This result allows for an expected average degree of $\sqrt{n}$. This is highly suboptimal, state of the art results

allow for $O(\log n)$. See Abbe, Bandeira, & Hall (2015).

Proof: Applying the Davis-Kahan $\sin\theta$ theorem

Unit norm eigenvector associated to $\lambda_2(\mathbb{E}A)$

$$\sin\angle(u_2(\mathbb{E}A), u_2(A)) \le \frac{2\|A-\mathbb{E}A\|_{op}}{\min\{\lambda_2(\mathbb{E}A),\ \lambda_1(\mathbb{E}A)-\lambda_2(\mathbb{E}A)\}}.$$

Check (HW 3)

$$\le 2\frac{\|A-\mathbb{E}A\|_{op}}{\mu n}.$$

Next, we use a fact that will occupy the next few lectures.

Fact (☺): We have that

$$\mathbb{P}(\|A-\mathbb{E}A\|_{op} \ge C\sqrt{n}) \le 4e^{-n}$$

for some universal constant $C>0$.

Thus, assuming we are in the event † we get

$$\sin\angle(u_2(\mathbb{E}A), u_2(A)) \le \frac{C\sqrt{n}}{\mu n} \le \frac{C}{\mu\sqrt{n}}.$$

Therefore,

$$\min_{s\in\{\pm1\}} \| u_2(\mathbb{E}A) - s\, u_2(A)\|_2 \lesssim \frac{C}{\mu\sqrt{n}}$$

Now, consider $v_2 := \sqrt{n}\, u_2$ we get

±1 entries

$$\#\{ i \mid \operatorname{sign}(v_2^{(i)}(A)) \neq v_2^{(i)}(\mathbb{E}A) \}$$

$$\leq \sum^n \mathbb{1}\{\operatorname{sign}(v_2^{(i)}(A)) \neq v_2^{(i)}(\mathbb{E}A)\}$$

Check!

$$\leq \sum^n (v_2^{(i)}(A) - v_2^{(i)}(\mathbb{E}(A)))^2$$

$$\leq C^2/\mu^2$$

$\square$

Nets, coverigs, and packings

Our next goal is to prove Fact(8).
Suppose $A \in \mathbb{R}^{n\times m}$ random matrix and our goal is to get high probability bounds on

$$\|A\|_{op} = \max_{v\in S^{m-1}} \|Av\|_2 .$$

Before, we derived probability bounds $\max_{i\in[n]} X_i$ using the union bound.
However, this only applied for

finitely many variables. In $\|A\|_{op}$ we have infinitely many of them. We will develop a powerful method to bound maxima $\max_{v \in K} X_v$ of random variables X_v that change continuously wrt to the index v.

Key idea: If we substitute the infinite set S^{m-1} by a finite set $\mathcal{N} \subseteq S^{m-1}$ such that

$$\max_{v \in S^{m-1}} \|Av\|_2 \approx \max_{v \in \mathcal{N}} \|Av\|_2 .$$

Then, we can apply the same union bound strategy as before. $\dashv$

In what follows we learn how to construct such finite sets $\mathcal{N}$.

<u>Def (ε-nets)</u>: Let (T, d) be a metric space. Consider a set $K \subseteq T$ and a number $\varepsilon > 0$. A subset $\mathcal{N} \subseteq K$ is called an ε-net of K if

$$\forall x \in K \ \exists x_0 \in \mathcal{N} \ \text{s.t.} \ d(x, x_0) \leq \varepsilon.$$

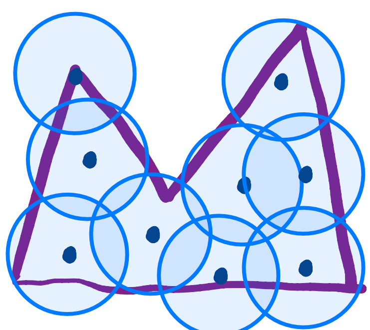
The covering number of K denoted $\mathcal{N}(K, d, \varepsilon)$ is the cardinality of the smallest ε-net of K.

$+$

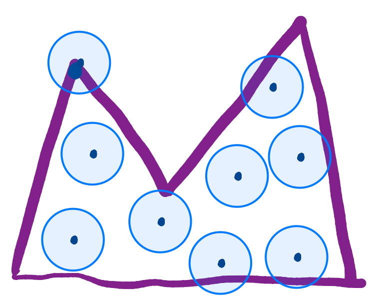
Metric space: $d : T \times T \to \mathbb{R}_+$ s.t $\forall x, y, z$

1. $d(x,x) = 0$,
2. if $x \neq y \Rightarrow d(x,y) > 0$,
3. $d(x,y) = d(y,x)$,
4. $d(x,y) \leq d(x,z) + d(z,y)$.

Example: $(\mathbb{R}^d, d)$ with $d(x,y) = \|x-y\|_2$.
In that case we cover the set with round balls.



ε-net



$\varepsilon/2$-packing

Def (Packing number): A subset $\mathcal{N}$ of a metric space (T,d) is ε-separated if

$$d(x,y) > \varepsilon \quad \text{if} \quad x, y \in \mathcal{N} \text{ and } x \neq y.$$

The cardinality of the largest ε-separa ted subset of $K \subseteq T$ is called the **packing number** of K denoted $\mathcal{P}(K, d, \varepsilon)$.

In turn, covering and packing numbers are almost the same.

Lemma ⊙: For any subset $K \subseteq T$ and $\varepsilon > 0$, we have

$$\mathcal{P}(K, d, 2\varepsilon) \leq \mathcal{N}(K, d, \varepsilon) \leq \mathcal{P}(K, d, \varepsilon).$$

Proof: To prove the lower bound, Take a 2ε-packing $\mathcal{P}$ and an ε-covering $\mathcal{N}$. Let $x \in \mathcal{P}$, then by def. there is $y \in \mathcal{N}$ s.t. $d(x, y) \leq \varepsilon$. By def

$\forall z \in \mathcal{P} \setminus \{x\}$

$$d(z, y) \geq d(z, x) - d(x, y) > \varepsilon.$$

Thus, for each $x \in \mathcal{P}$ there exist a unique $y \in \mathcal{N}$ and so $|\mathcal{P}| \leq |\mathcal{N}|$. Since $\mathcal{P}$ and $\mathcal{N}$ are arbitrary, the lower bound follows.

To prove the upper bound. Let $\mathcal{N}$

be a maximal ε-separated set of K, i.e., $|\mathcal{N}| = \mathcal{P}(K, d, \varepsilon)$. We want to show that $\mathcal{N}$ is an ε-net. Let $x \in K$, suppose in search of contradiction that $\forall y \in \mathcal{N}$

$$d(x, y) > \varepsilon.$$

This, implies that $\mathcal{N} \cup \{x\}$ is a larger ε-separated set. $\lightning$

Thus,

$$\mathcal{N}(K, d, \varepsilon) \leq |\mathcal{N}| = \mathcal{P}(K, d, \varepsilon). \qquad \square$$