Applications of Matrix Deviation Ineq. (MDI)

0. JL Lemma
1. Spectra of Random Matrices
2. Covariance Estimation
3. Random Projection of Sets
4. Random Section of Sets (M* Bound)
5. Escape Theorem (Gordon)
6. Compressed Sensing

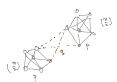
7. Community Detection } Today.

Community Detection in Networks (via Spectral Clustering)

Stochastic Block Model

Defn: Divide n vertices into two sets ("communities") of sizes $n/2$ each. Construct a random graph G by connecting each pair of vertices independently with prob p, if they belong to the same community, and q if they belong to different communities. This distribution on graphs is called the stochastic block model, and is denoted by $G(n, p, q)$

We assume $p > q$



Erdős Rényi Random Graphs $p = q$.

Here $p > q$ ⊥

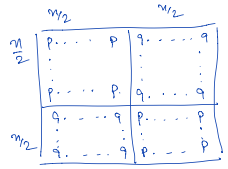
For a random graph $G \sim G(n, p, q)$
write A to denote the adjacency matrix of G

$A_{ij} \sim \text{Ber}(p)$ if i & j in the same community

$A_{ij} \sim \text{Ber}(q)$ if i & j are in different communities

Write $D :=$ Expected Adj. Matrix with the "right" indexing (unknown a priori)

$D =$



$A = D + R$

Signal Strength $\|D\| = \lambda_1 \approx n$

Noise Level $\|R\| \leq C\sqrt{n}$

$\text{rank}(D) = 2$. Its eigenvectors & values are

$u_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ $\lambda_1 = \left(\frac{p+q}{2}\right) \cdot n$ $\quad \|u_1\|_2 = \sqrt{n}$

$u_2 = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}$ u_2 $\lambda_2 = \left(\frac{p-q}{2}\right) \cdot n$

Eigenvector u_2 contains all the info about the communities simply by considering the sign of the components

---

Lemma [Norm of Symmetric Matrices with Sub-gaussian entries]

Let R be a $n \times n$ random matrix whose entries R_{ij} on & above the diagonal are ind., mean-zero, sub-gauss rvs.

Then, for any $t > 0$, we have $\|R\| \leq CK(\sqrt{n} + t)$

with prob at least $1 - 4\exp(-t^2)$.

where, $K := \max_{i,j} \|A_{ij}\|_{\psi_2}$

Proof Sketch:

- Decompose A into upper & lower triangular matrices A^+ A^-
- $\|A\| = \|A^+ + A^-\| \leq \|A^+\| + \|A^-\|$
- Union Bound.

Perturbation Theory

Thm: [Weyl's Ineq.] for any symmetric matrices S & T, with the same dimension we have

$$\max_i |\lambda_i(S) - \lambda_i(T)| \leq \|S - T\|$$

- The spectral norm of $(S-T)$ controls the stability of the spectrum.
- Follows from Courant-Fisher min-max characterization of eigenvalues.

Similar result for eigenvectors.

THM: [Davis Kahan] Let S & T be symmetric matrices with the same dimension. Fix i and assume that ith largest eigen value of S is well-separated from the rest of the spectrum.

$$\min_{j : j \neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0$$

Then, the ith largest (normalized) eigenvectors of S & T (i.e., eigenvectors corresponding to the ith eigenvalues) satisfy

$$\|v_i(S) - \theta v_i(T)\| \leq 2^{3/2} \frac{\|S-T\|}{\delta}$$

for some sign $\theta \in \{\pm 1\}$.

Spectral Clustering

We will apply Davis-Kahan with $S = D$ & $T = A = D + R$

- In particular, we want to show that v_2 (eigenvector of D) is not perturbed much by R.

- Need to show that λ_2 is well separated from the rest of the spectrum.

$\delta = \min(\lambda_2, \lambda_1 - \lambda_2)$

$= \min\left(\frac{p-q}{2}, q\right) n$

$\delta = \mu \cdot n$

- Applying Davis Kahan thm to bound the unit eigenvectors of D & A.

$$\|v_2(D) - \theta \cdot v_2(A)\| \leq \frac{C\|D-A\|}{\delta} \leq \frac{C\sqrt{n}}{\mu n}$$

w.p. at least $(1 - 4e^{-n})$

(Recall $\|R\| \le C\sqrt{n}$ whp)

Eigenvectors $u_i(D)$ have norm $\sqrt{n}$.
Therefore, normalizing we obtain

$$\| u_2(D) - \theta \cdot u_2(A) \|_2 \le \frac{C}{\mu}$$

Therefore,

$$\sum_j \left(u_2(D)_j - u_2(A)_j \right)^2 \le \frac{C}{\mu^2}$$

If $\operatorname{sgn}(u_2(D)_j) \neq \operatorname{sgn}(u_2(A)_j)$, then the difference is at least 1.

Hence, the number of components that differ in sign b/w $u_2(D)$ & $u_2(A)$ is at most $\frac{C}{\mu^2}$

Summarizing, we can use $u_2(A)$ to accurately estimate $u_2(D)$.

Spectral Clustering Algorithm

Input: Graph $G \sim G(n, p, q)$. i.e., the adj matrix A of G.

Output: A partition of G into two communities

1. Compute $u_2(A)$ - eigenvector corresponding to second largest eigenvalue of A
2. Partition the vertices of G based on the signs of the coeffs. of $u_2(A)$. (into two communities)

THM 8. (Spectral Clustering for the Stochastic Block Model)

Let $G \sim G(n, p, q)$ with $p > q$ and $\min\{p-q, q\} = \mu > 0$

Then, with probability at least $1 - 4\exp(-n)$, the spectral clustering alg. identifies the communities of G. correctly up to $\frac{C}{\mu^2}$ misclassified vertices.