COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Andrew McGregor

Lecture 19
• Week 10 Quiz is due Monday at 8pm.
Last Class: Spectral Graph Theory

- View of a graph in terms of adjacency matrix and Laplacian.
- Spectral embedding for non-linear dimensionality reduction.
- Start on graph clustering for community detection and non-linear clustering.
- Idea of finding small cuts that separate large sets of nodes.

This Class: Spectral Clustering and the Stochastic Block Model

- Spectral clustering: finding good cuts via Laplacian eigenvectors.
- Stochastic block model: A simple clustered graph model where we can prove the effectiveness of spectral clustering.
For a graph with adjacency matrix $A$ and degree matrix $D$, $L = D - A$ is the graph Laplacian.

For any vector $\tilde{\vec{v}}$, its ‘smoothness’ over the graph is given by:

$$\sum_{(i,j) \in E} (\tilde{\vec{v}}(i) - \tilde{\vec{v}}(j))^2 = \tilde{\vec{v}}^T L \tilde{\vec{v}}.$$
For a cut indicator vector $\vec{v} \in \{-1, 1\}^n$ with $\vec{v}(i) = -1$ for $i \in S$ and $\vec{v}(i) = 1$ for $i \in T$:

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

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

**Next Step:** See how this dual minimization problem is naturally solved (sort of) by eigendecomposition.
The smallest eigenvector of the Laplacian is:

\[ \vec{v}_n = \frac{1}{\sqrt{n}} \cdot \vec{1} = \arg \min_{\vec{v} \in \mathbb{R}^n \text{ with } \|\vec{v}\| = 1} \vec{v}^T L \vec{v} \]

with eigenvalue \( \vec{v}_n^T L \vec{v}_n = 0 \). Why?

---

**n**: number of nodes in graph, \( A \in \mathbb{R}^{n \times n} \): adjacency matrix, \( D \in \mathbb{R}^{n \times n} \): diagonal degree matrix, \( L \in \mathbb{R}^{n \times n} \): Laplacian matrix \( L = A - D \).
By Courant-Fischer, the second smallest eigenvector is given by:

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

If $\vec{v}_{n-1}$ were in $\left\{ -\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}} \right\}^n$ it would have:

• $\vec{v}_{n-1}^T L \vec{v}_{n-1} = \frac{4}{\sqrt{n}} \cdot \text{cut}(S, T)$ as small as possible given that $\vec{v}_{n-1}^T \vec{v}_{n-1} = \frac{1}{\sqrt{n}} \vec{v}_{n-1}^T \vec{1} = \frac{|T|-|S|}{n} = 0$.

• I.e., $\vec{v}_{n-1}$ would indicate the smallest perfectly balanced cut.

• The eigenvector $\vec{v}_{n-1} \in \mathbb{R}^n$ is not generally binary, but still satisfies a ‘relaxed’ version of this property.

---

$n$: number of nodes in graph, $A \in \mathbb{R}^{n \times n}$: adjacency matrix, $D \in \mathbb{R}^{n \times n}$: diagonal degree matrix, $L \in \mathbb{R}^{n \times n}$: Laplacian matrix $L = A - D$. $S, T$: vertex sets on different sides of cut.
Find a good partition of the graph by computing

$$\vec{v}_2 = \arg \min_{\vec{v} \in \mathbb{R}^d \text{ with } \|\vec{v}\|=1} \vec{v}^T L \vec{v}$$

Set $S$ to be all nodes with $\vec{v}_2(i) < 0$, $T$ to be all with $\vec{v}_2(i) \geq 0$. 
The Shi-Malik normalized cuts algorithm is one of the most commonly used variants of this approach, using the normalized Laplacian \( \overline{L} = D^{-1/2}LD^{-1/2} \).

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

**Spectral Clustering:**
- Compute smallest \( k \) nonzero eigenvectors \( \tilde{\mathbf{v}}_n - 1, \ldots, \tilde{\mathbf{v}}_n - k \) of \( \overline{L} \).
The smallest eigenvectors of $L = D - A$ give the orthogonal ‘functions’ that are smoothest over the graph. I.e., minimize

$$\vec{v}^T L \vec{v} = \sum_{(i,j) \in E} [\vec{v}(i) - \vec{v}(j)]^2.$$ 

Embedding points with coordinates given by $[\vec{v}_{n-1}(j), \vec{v}_{n-2}(j), \ldots, \vec{v}_{n-k}(j)]$ ensures that coordinates connected by edges have minimum total squared Euclidean distance.

- Spectral Clustering
- Laplacian Eigenmaps
- Locally linear embedding
- Isomap
- Node2Vec, DeepWalk, etc. (variants on Laplacian)
Laplacian Embedding

Original Data: (not linearly separable)

$k$-Nearest Neighbors Graph:
**So Far:** Have argued that spectral clustering partitions a graph effectively, along a small cut that separates the graph into large pieces. But it is difficult to give any formal guarantee on the ‘quality’ of the partitioning in general graphs.

**Common Approach:** Give a natural generative model for random 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 least squares regression, $k$-means clustering, PCA, etc.)
Stochastic Block Model (Planted Partition Model): Let $G_n(p, q)$ be a distribution over graphs on $n$ nodes, split randomly 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. $q < p$.
- Connections are independent.
Let $G$ be a stochastic block model graph drawn from $G_n(p, q)$.

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

$G_n(p, q)$: stochastic block model distribution. $B, C$: groups with $n/2$ nodes each. Connections are independent with probability $p$ between nodes in the same group, and probability $q$ between nodes not in the same group.
Letting $G$ be a stochastic block model graph drawn from $G_n(p, q)$ and $A \in \mathbb{R}^{n \times n}$ be its adjacency matrix. $(\mathbb{E}[A])_{i,j} = p$ for $i, j$ in same group, $(\mathbb{E}[A])_{i,j} = q$ otherwise.

What is $\text{rank}(\mathbb{E}[A])$?
What are the eigenvectors and eigenvalues of $\mathbb{E}[A]$?

$G_n(p, q)$: stochastic block model distribution. $B, C$: groups with $n/2$ nodes each. Connections are independent with probability $p$ between nodes in the same group, and probability $q$ between nodes not in the same group.
Letting $G$ be a stochastic block model graph drawn from $G_n(p, q)$ and $A \in \mathbb{R}^{n \times n}$ be its adjacency matrix, what are the eigenvectors and eigenvalues of $\mathbb{E}[A]$?
If we compute $\vec{v}_2$ then we recover the communities $B$ and $C$!

- Can show that for $G \sim G_n(p, q)$, $A$ is close to $\mathbb{E}[A]$ with high probability (matrix concentration inequality).
- Thus, the true second eigenvector of $A$ is close to $[1, 1, 1, \ldots, -1, -1, -1]$ and gives a good estimate of the communities.
Goal is to recover communities – so adjacency matrix won’t be ordered in terms of community ID (or our job is already done!)

- Actual adjacency matrix is \( \mathbf{PAP}^T \) where \( \mathbf{P} \) is a random permutation matrix and \( \mathbf{A} \) is the ordered adjacency matrix.
- **Exercise**: The first two eigenvectors of \( \mathbf{PAP}^T \) are \( \mathbf{P}\vec{v}_1 \) and \( \mathbf{P}\vec{v}_2 \).
- \( \mathbf{P}\vec{v}_2 = [1, -1, 1, -1, \ldots, 1, 1, -1] \) gives community ids.
Letting $G$ be a stochastic block model graph drawn from $G_n(p, q)$, $A \in \mathbb{R}^{n \times n}$ be its adjacency matrix and $L$ be its Laplacian, what are the eigenvectors and eigenvalues of $\mathbb{E}[L]$?
Letting $G$ be a stochastic block model graph drawn from $G_n(p, q)$, $A \in \mathbb{R}^{n \times n}$ be its adjacency matrix and $L$ be its Laplacian, what are the eigenvectors and eigenvalues of $\mathbb{E}[L]$?
**Upshot:** The second smallest eigenvector of $\mathbb{E}[L]$ is $\chi_{B,C}$ — the indicator vector for the cut between the communities.

- If the random graph $G$ (equivalently $A$ and $L$) were exactly equal to its expectation, partitioning using this eigenvector would exactly recover the two communities $B$ and $C$.

How do we show that a matrix (e.g., $A$) is close to its expectation? Matrix concentration inequalities.

- Analogous to scalar concentration inequalities like Markovs, Chebyshevs, Bernsteins.

- Random matrix theory is a very recent and cutting edge subfield of mathematics that is being actively applied in computer science, statistics, and ML.
Everything after this slide is bonus material, if you are interested in how we formally prove that spectral clustering succeeds in the stochastic block model, using matrix concentration bounds.
Matrix Concentration Inequality: If \( p \geq O \left( \frac{\log^4 n}{n} \right) \), then with high probability

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

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

For any \( X \in \mathbb{R}^{n \times d} \), \( \| X \|_2 = \max_{z \in \mathbb{R}^d : \| z \|_2 = 1} \| Xz \|_2 \).

**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?
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 $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$ and $\overline{\mathbf{v}}_1, \overline{\mathbf{v}}_2, \ldots, \overline{\mathbf{v}}_d$. Letting $\theta(\mathbf{v}_i, \overline{\mathbf{v}}_i)$ denote the angle between $\mathbf{v}_i$ and $\overline{\mathbf{v}}_i$, for all $i$:

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

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

The errors get large if there are eigenvalues with similar magnitudes.
EIGENVECTOR PERTURBATION

\[
\begin{align*}
\mathbf{A} & = \begin{pmatrix} 1 + \varepsilon & 0 \\ 0 & 1 \end{pmatrix} \\
\tilde{\mathbf{A}} & = \begin{pmatrix} 1 & 0 \\ 0 & 1 + \varepsilon \end{pmatrix} \\
\mathbf{A} - \tilde{\mathbf{A}} & = \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}
\end{align*}
\]
Claim 1 (Matrix Concentration): For $p \geq O\left(\frac{\log^4 n}{n}\right)$,

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

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

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

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

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

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

A adjacency matrix of random stochastic block model graph. $p$: connection probability within clusters. $q < p$: connection probability between clusters. $n$: number of nodes. $v_2, \bar{v}_2$: second eigenvectors of $A$ and $\mathbb{E}[A]$ respectively.
**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 $\|v_2 - \bar{v}_2\|_2^2 \leq O \left( \frac{p}{(p-q)^2 n} \right)$ (exercise).
- $\bar{v}_2$ is $\frac{1}{\sqrt{n}} \chi_{B,C}$: the community indicator vector.

\[
\begin{bmatrix}
\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & -\frac{1}{\sqrt{n}} & -\frac{1}{\sqrt{n}} & -\frac{1}{\sqrt{n}} & -\frac{1}{\sqrt{n}} \\
\end{bmatrix}
\]

\[
\bar{v}_2
\]

- Every $i$ where $v_2(i), \bar{v}_2(i)$ differ in sign contributes $\geq \frac{1}{n}$ to $\|v_2 - \bar{v}_2\|_2^2$.
- So they differ in sign in at most $O \left( \frac{p}{(p-q)^2} \right)$ positions.

**A** adjacency matrix of random stochastic block model graph. $p$: connection probability within clusters. $q < p$: connection probability between clusters. $n$: number of nodes. $v_2, \bar{v}_2$: second eigenvectors of $A$ and $\mathbb{E}[A]$ respectively.
Upshot: If $G$ is a stochastic block model graph with adjacency matrix $A$, 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.