CS 466/666 Algorithm Design and Analysis, Spring 2022

Lecture 13: Cheeger's inequalities

We will study Cheeger's inequality which relates the second eigenvalue of the Laplacian matrix to the (combinatorial) expansion of a graph. We will also discuss some recent generalizations.

Graph expansion

Recall that $\lambda_2 = 0$ if and only if G is disconnected.

Cheeger's inequality will show that λ_2 is "small" if and only if G is "close" to be disconnected.

First, let us make precise what it means for a graph to be close to be disconnected.

There are different definitions to measure how well a graph is connected.

The <u>expansion</u> of a graph is defined as $\Phi(G) := \min_{S \subseteq V,\ |S| \le |V|/2} \Phi(S)$, where $\Phi(S) := \frac{|\delta(S)|}{|S|}$, the ratio of the number of edges cut to the number of vertices in the set.

The <u>conductance</u> of a graph is defined as $\phi(G) := \min_{S \subseteq V,\ \mathrm{vol}(S) \le |E|} \phi(S)$, where $\phi(S) := \frac{|\delta(S)|}{\mathrm{vol}(S)}$ and $\mathrm{vol}(S) := \sum_{v \in S} \deg(v)$, the ratio of the number of edges cut to the total degree in the set.

These definitions are basically equivalent when the graphs are d-regular ($\Phi(S) = d\,\phi(S)$).

In non-regular graphs, we will relate the graph conductance to the second eigenvalue.

We say a graph is an <u>expander graph</u> if $\phi(G)$ is large (e.g. $\phi(G) \ge 0.1$), and we say $S \subseteq V$ a sparse cut if $\phi(S)$ is small. Note that $0 \le \phi(S) \le 1$ for every $S \subseteq V$.

Both concepts are very useful. As we have seen, sparse expander graphs are "magical" and have algorithmic applications, and they are also useful in derandomization.

Finding a sparse cut is useful in designing divide-and-conquer algorithms, and have applications in image segmentation, data clustering, community detection in social networks, VLSI design, etc.

The spectral partitioning algorithm

This is a popular heuristic in finding a sparse cut in practice.

① Compute the second eigenvector $x \in \mathbb{R}^n$ of $\mathcal{L}$.

② Sort the vertices so that $x_1 \ge x_2 \ge \dots \ge x_n$.

③ Let $S_i = \begin{cases} \{1, \dots, i\} & \text{if } i \le n/2 \\ \{i+1, \dots, n\} & \text{if } i \ge n/2 \end{cases}$

Return $\min_{1 \le i \le n} \{\phi(S_i)\}$.

There is a near-linear time algorithm, known as the "power method", to compute the second eigenvector. So, the whole algorithm can be implemented in near-linear time, and easy to code using MATLAB. This is one reason that this heuristic is popular.

Another reason is that it performs very well in various applications, especially in image segmentation and clustering, and it was a major breakthrough in image segmentation around 2000. The proof of Cheeger's inequality will provide some performance guarantee of this algorithm.

Normalized matrices

To state Cheeger's inequality nicely, we will use the "normalized" Laplacian matrix, which allows us to remove the dependence on the maximum degree of the graph in the theorem statement.

Given an adjacency matrix A, let $\mathcal{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ be the normalized adjacency matrix, and let $\mathcal{L} = I - \mathcal{A}$ be the normalized Laplacian matrix, where D is the diagonal matrix whose i-th entry is the degree of vertex i. Note that $\mathcal{L} = I - \mathcal{A} = D^{-\frac{1}{2}}(D-A)D^{-\frac{1}{2}} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$.

Let $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n$ be the eigenvalues of $\mathcal{A}$ and let $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ be the eigenvalues of $\mathcal{L}$.

Claim $1 = \alpha_1 \geq \alpha_n \geq -1$ and $0 = \lambda_1 \leq \lambda_n \leq 2$.

Proof We prove the result for normalized adjacency, and the result for normalized Laplacian follows directly.

Note that 0 is an eigenvalue for $\mathcal{L}$, as $\mathcal{L}(D^{\frac{1}{2}}\vec{1}) = (D^{-\frac{1}{2}} L D^{-\frac{1}{2}})(D^{\frac{1}{2}}\vec{1}) = D^{-\frac{1}{2}} L \vec{1} = 0$

To prove $\lambda_1 = 0$, we will show that $\mathcal{L}$ is a positive semidefinite matrix.

To see it, observe that $x^T \mathcal{L} x = x^T D^{-\frac{1}{2}} L D^{-\frac{1}{2}} x = \sum_{e \in E} x^T D^{-\frac{1}{2}} L_e D^{-\frac{1}{2}} x = \sum_{e=ij \in E} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2 \geq 0$,

where $L_e = b_e b_e^T$ that we defined last time (i.e. $L_e = \begin{matrix} \\ i \\ j \end{matrix} \begin{matrix} \begin{matrix} i & j \end{matrix} \\ \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{matrix}$ for $e = ij$).

This implies that $I - \mathcal{A} \succcurlyeq 0$, and thus $\alpha_1 \leq 1$.

Also, we can write $x^T(I+\mathcal{A})x = x^T \mathcal{L} x + 2x^T \mathcal{A} x = \sum_{e=ij \in E} \left(\left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2 + \frac{2x_i x_j}{\sqrt{d_i d_j}} \right) = \sum_{e=ij \in E} \left(\frac{x_i}{\sqrt{d_i}} + \frac{x_j}{\sqrt{d_j}} \right)^2 \geq 0$,

and this implies that $I + \mathcal{A} \succcurlyeq 0$, and thus $\alpha_n \geq -1$, and hence $\lambda_n = 1 - \alpha_n \leq 2$. □

Cheeger's inequality

Theorem $\frac{1}{2}\lambda_2 \leq \phi(G) \leq \sqrt{2\lambda_2}$, where λ_2 is the second smallest eigenvalue of $\mathcal{L}$ of G.

For simplicity, we only prove the theorem when G is a d-regular graph, in which case $\mathcal{L} = \frac{1}{d} L$.

The general case is similar but a little bit more involved.

The first inequality is called the easy direction, and the second inequality is called the hard direction.

So, naturally we prove the easy direction first.

One nice thing about the Laplacian matrix is that we know that the first eigenvector is the all-one vector

so by the characterization of λ_2 using Rayleigh quotient, we have

$$\lambda_2 = \min_{x \perp \vec{1}} \frac{x^T \mathcal{L} x}{x^T x} = \min_{x \perp \vec{1}} \frac{x^T L x}{d x^T x} = \min_{x \perp \vec{1}} \frac{\sum_{ij \in E} (x_i - x_j)^2}{d \sum_{i \in V} x_i^2}.$$

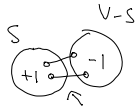
So, to upper bound λ_2, we just need to find a vector $x \perp \vec{1}$ and compute its Rayleigh quotient.

To get some intuition, let say $\phi(G) = \phi(S)$ and $|S| = n/2$.

We consider the "binary" solution: $x_i = +1$ if $i \in S$ and $x_i = -1$ if $i \notin S$.

Since $|S| = n/2$, $\sum_{i \in V} x(i) = 0$, and thus $x \perp \vec{1}$.

Then $\lambda_2 \le \frac{\sum_{ij \in E} (x_i - x_j)^2}{d \sum_{i \in V} x_i^2} = \frac{4 |\delta(S)|}{d|V|} = \frac{2|\delta(S)|}{d|S|} = 2\phi(S)$.



each edge in $\delta(S)$ contributes 4.

For general S, we consider the binary solution $x_i = \frac{+1}{|S|}$ if $i \in S$ and $x_i = \frac{-1}{|V-S|}$ if $i \notin S$.

By construction, $x \perp \vec{1}$.

So, $\lambda_2 \le \frac{\sum_{ij \in E} (x_i - x_j)^2}{d \sum_{i \in V} x_i^2} = \frac{|\delta(S)| \cdot \left(\frac{1}{|S|} + \frac{1}{|V-S|}\right)^2}{d\left(|S| \cdot \frac{1}{|S|^2} + |V-S| \cdot \frac{1}{|V-S|^2}\right)} = \frac{|\delta(S)| \cdot |V|}{d \cdot |S| \cdot |V-S|} \le 2\phi(S)$.

This proves the easy direction.

To summarize, if there is a sparse cut, then λ_2 is small.

We should think of λ_2 is a "relaxation" of the graph conductance problem, which is polynomial time solvable.

A consequence is that if λ_2 is large, then we know that G has no sparse cut.

This direction is useful in deterministic construction of expander graphs.

The hard direction: Intuition

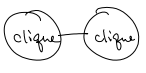
In the minimization problem $\min_{x \perp \vec{1}} \frac{\sum_{ij \in E} (x_i - x_j)^2}{d \sum_{i \in V} x_i^2}$, if we can only search for "binary" solutions,

then we are essentially optimizing over the Conductances.

But we are optimizing over a larger domain over $\mathbb{R}^n \perp \vec{1}$ (otherwise the problem is NP-hard),

and the optimal solutions could be some very non-binary solutions (e.g. very "smooth" vector),

for which it is not clear how to find a sparse cut from it.

To get some feeling, suppose we are given a graph like 

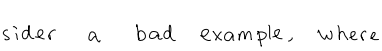
, which is a good case.

In this case, it is not good to "split" the vertices in a clique, because there are so many edges within it.

So, we would expect that the values in each clique are very similar, while the two cliques would have

different values so that $x \perp \vec{1}$.

Hence, we expect that the minimizer would look very similar to a binary vector, and we can find a good cut with $\phi(S) \approx \lambda_2$ by looking at the (binary) second eigenvector.

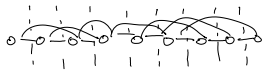
Now, consider a bad example, where the graph is like  .

Then, the minimizer x can do much better than a binary vector, by making each edge very short, while the values decrease smoothly from +1 to -1, in which case $\lambda_2 \approx \phi(G)^2$.

The key of Cheeger's inequality is to show that λ_2 cannot be smaller than $\phi(G)^2/2$.

In other words, if λ_2 is small, then we can extract a somewhat sparse cut from the eigenvector.

We can think of the optimize "embeds" the graph into a line, while most edges are short.



Then, it should be the case that some threshold gives a sparse cut (i.e. row and column argument).

The hard direction · Proof (optional)

The first step is to preprocess the second eigenvector so that at most half the entries are nonzero.

This would guarantee that the output set S satisfies $|S| \le |V|/2$.

This step is simple to describe.

Without loss of generality we assume that there are fewer positive entries in x than negative entries.

Consider the following vector y: $y_i = \begin{cases} x_i & \text{if } x_i \ge 0 \\ 0 & \text{if } x_i < 0 \end{cases}$, i.e. just zeroing out the negative part.

Denote the Rayleigh quotient by $R(x) := x^T \mathcal{L} x / x^T x = x^T L x / d x^T x$.

Claim $R(y) \le R(x)$.

proof For all i with $y_i > 0$, $(\mathcal{L}y)_i = y_i - \sum_{j \in N(i)} \frac{y_j}{d} \le x_i - \sum_{j \in N(i)} \frac{x_j}{d} = (\mathcal{L}x)_i = \lambda_2 x_i$.

Therefore, $y^T \mathcal{L} y = \sum_{i \in V} y_i \cdot (\mathcal{L}y)_i = \sum_{i: y(i) > 0} y_i (\mathcal{L}y)_i \le \sum_{i: y(i) > 0} \lambda_2 x_i^2 = \sum_{i \in V} \lambda_2 y_i^2$, proving the claim. □

Now, there is a very element argument to make the above intuition precise: just pick a random threshold!

Lemma Given any y, there exists a subset $S \subseteq \text{supp}(y)$ such that $\phi(S) \le \sqrt{2R(y)}$, where $\text{supp}(y) = \{ i \mid y(i) \ne 0 \}$.

Proof We can assume that $0 \le y_i \le 1$ for all i, by scaling y if necessary.

Let $t \in (0,1]$ be chosen uniformly at random.

Let $S_t = \{ i \mid y_i^2 \ge t \}$. Then $S_t \subseteq \text{supp}(y)$ by construction.

We analyze the expected value of $|\delta(S_t)|$ and the expected value of $|S_t|$.

$$\mathop{E}_t [|\delta(S_t)|] = \sum_{ij \in E} \left[\Pr(\text{the edge } ij \text{ is cut}) \right] \quad \text{by linearity of expectation}$$

$$= \sum_{ij\in E} \left[\Pr\left(y_i^2 < t \le y_j^2 \right) \right]$$

$$= \sum_{ij\in E} |\, y_i^2 - y_j^2 \,|$$

$$= \sum_{ij\in E} |y_i - y_j| \cdot |y_i + y_j|$$

$$\le \sqrt{\sum_{ij\in E} (y_i - y_j)^2} \sqrt{\sum_{ij\in E} (y_i + y_j)^2} \qquad \text{by Cauchy-Schwarz } \langle a,b\rangle \le \|a\| \cdot \|b\|$$

$$\le \sqrt{\sum_{ij\in E} (y_i - y_j)^2} \sqrt{2 \sum_{ij\in E} (y_i^2 + y_j^2)}$$

$$= \sqrt{\sum_{ij\in E} (y_i - y_j)^2} \sqrt{2d \sum_{i\in V} y_i^2}$$

$$= \sqrt{2R(y)} \left(d \sum_{i\in V} y_i^2 \right).$$

$$\mathop{E}_t [|S_t|] = \sum_{i\in V} \Pr[y_i^2 \ge t] = \sum_{i\in V} y_i^2.$$

Therefore, $\dfrac{\mathop{E}_t [|\delta(S_t)|]}{\mathop{E}_t [d\,|S_t|]} \le \sqrt{2R(y)}$.

This means that $\mathop{E}_t \left[|\delta(S_t)| - \sqrt{2R(y)} \cdot d \cdot |S_t| \right] \le 0$.

Hence, there exists t such that $\phi(S_t) = \dfrac{|\delta(S_t)|}{d \cdot |S_t|} \le \sqrt{2R(y)}$. □

Combining the claim and the lemma proves the hard direction of Cheeger's inequality.

Note that the proof shows that the spectral partitioning algorithm achieves the performance guarantee, because the output set S_t is a "threshold" set that the algorithm has searched for.

Discussions

① The proof can be generalized to weighted non-regular graphs, with suitable modifications.

② Both sides of Cheeger's inequality are tight, even the constants are tight.

To see an example where the hard direction is almost tight, consider a cycle of length n.

One can compute the spectrum of the cycle exactly, but we won't do it here.

Recall that $\lambda_2 = \min_{x \perp \vec{1}} \dfrac{x^T L x}{x^T x}$, so to give an upper bound on λ_2, we just need to demonstrate one vector.

Consider $x = \left(1, 1-\frac{1}{n}, 1-\frac{2}{n}, \ldots, \frac{1}{n}, 0, -\frac{1}{n}, \frac{-2}{n}, \ldots, -1+\frac{1}{n}, -1, -1-\frac{1}{n}, \ldots, -\frac{1}{n}, 0, \frac{1}{n}, \ldots, 1-\frac{1}{n}, 1\right)$.

Then $\lambda_2 \le \dfrac{\sum_{ij} (x_i - x_j)^2}{2 \sum_i x_i^2} = O\left(\dfrac{n(\frac{1}{n})^2}{n} \right) = O\left(\frac{1}{n^2}\right)$.

On the other hand, it is easy to verify that the conductance of a cycle is $\Omega(\frac{1}{n})$.

Therefore, in this example, $\phi(G) = \Omega(\sqrt{\lambda_2})$.

One may think that it is an artificial example in which the second eigenvalue clearly underestimates the conductance, but let's consider the following related example.

Two cycles of length n, and there is a perfect matching between the two cycles. where each edge

Two cycles of length n, and there is a perfect matching between the two cycles, where each edge in the matching has weight $100/n^2$.

Clearly, the optimal sparse cut is the perfect matching, with $\phi(G) = O(\frac{1}{n^2})$.

On the other hand, one can show that the second eigenvector would still be the same as in the cycle example, with two nodes in the perfect matching identified as one node.

Therefore, λ_2 is still $O(\frac{1}{n^2})$ and the value is correct, but the optimal cut is lost and every threshold cut is bad

③ Related to the above point, Cheeger's inequality gives an $O(\frac{1}{\sqrt{\lambda_2}})$-approximation algorithm for estimating $\phi(G)$.

When λ_2 is large (e.g. when λ_2 is a constant), then it is a good approximation.

But λ_2 could be as small as $O(\frac{1}{n^2})$, and so it could be an $\Omega(n)$-approximation.

This doesn't quite explain the good empirical performance in practice.

④ The second eigenvalue is closely related to the mixing time of random walks, and so Cheeger's inequality provides a combinatorial approach to bound the mixing time, which we will see next time.

Recent generalizations (optional)

The last eigenvalue

Note that a graph has a bipartite component iff $\alpha_n = 0$, where α_n is the smallest eigenvalue of $I + \mathcal{A}$.

There is a robust generalization of this spectral characterization.

Define $\beta(G) = \min_{y \in \{-1,0,+1\}^V} \frac{\sum_{ij \in E} |y(i) + y(j)|}{d \sum_{i \in V} |y(i)|} = \min_{\substack{S \subseteq V \\ (L,R) \text{ partition of } S}} \frac{2|\#\text{edges within } L| + 2|\#\text{edges within } R| + |\delta(S)|}{d|S|}$.

This is called the bipartiteness ratio of G, which is small if and only if G contains a subset $S \subseteq V$ which is close to a bipartite component, with most edges in S crossing L and R

Theorem (Trevisan) $\frac{1}{2}\alpha_n \le \beta(G) \le \sqrt{2\alpha_n}$, where α_n is the smallest eigenvalue of $I + \mathcal{A}$.

The statement and also the proof are similar to that of Cheeger's inequality.

This result can be used to design a nontrivial approximation algorithm for the "maximum cut" problem.

The k-th eigenvalue

Recall that $\lambda_k = 0$ iff G has k connected components, where λ_k is the k-th smallest eigenvalue of $\mathcal{L}(G)$.

It turns out that there are two meaningful ways to generalize this basic fact.

① Small sparse cut : If λ_k is small, then there is a sparse cut S with $|S| \approx |V|/k$.

② Many sparse cuts: If λ_k is small, then there are k vertex-disjoint sparse cuts.

The first result is proved by an argument using random walks, and the result roughly says that for large enough k, there is a cut S with $\phi(S) \approx \sqrt{\lambda_k}$ and $|S| \approx |V|/k$.

The second result is proved by a spectral embedding argument that maps each vertex to a point in $\mathbb{R}^k$ using its entries in first k eigenvectors, and then use some geometric method to partition the points, where this is a heuristic used in practice.

Theorem $\frac{1}{2}\lambda_k \le \phi_k(G) \le O(k^2)\cdot\sqrt{\lambda_k}$, where $\phi_k(G) := \min_{\substack{S_1, S_2, \dots, S_k \\ \text{disjoint}}} \max_{1 \le i \le k} \phi(S_i)$.

There is also a generalization of Cheeger's inequality using λ_k.

Theorem $\frac{1}{2}\lambda_2 \le \phi(G) \le O\left(\frac{k\lambda_2}{\sqrt{\lambda_k}}\right)$ for any $k \ge 2$.

So, when λ_k is large for a small k (e.g. $\lambda_{10} = 0.2$), then this theorem says that λ_2 is a constant factor approximation of $\phi(G)$, and indeed the same spectral partitioning algorithm achieves this performance guarantee.

In image segmentation and data clustering, practical instances usually have a small number of outstanding objects / clusters, and this implies that λ_k is large for a small k.

Therefore, this theorem rigorously explains the good empirical performance of the spectral partitioning algorithm in practice.

References Course notes of Dan Spielman, and Luca Trevisan, and CS 860 in Spring 2019.