

CS 798 - Algorithmic Spectral Graph Theory, Fall 2015, Waterloo

Lecture 6: Improved Cheeger's inequality

We will analyze the performance guarantee of Cheeger's rounding using higher eigenvalues.

This provides better explanation of its success in practice.

Analysis of spectral partitioning

Assume the graph is d -regular. Let $\mathcal{L} = \frac{1}{d}$ be the normalized Laplacian matrix with eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n \leq 2$.

Cheeger's inequality states that $\frac{1}{2}\lambda_2 \leq \phi(G) \leq \sqrt{2\lambda_2}$, and the proof shows that the "sweep" algorithm

on the second eigenvector will find a set $S \subseteq V$ with $\phi(S) \leq \sqrt{2\lambda_2}$ and $|S| \leq |V|/2$.

Since $\phi(G) \geq \frac{1}{2}\lambda_2$, this implies that $\phi(S) \leq 2\sqrt{\phi(G)}$, and thus the spectral partitioning algorithm is a $\frac{1}{\sqrt{\phi(G)}}$ -approximation algorithm for graph expansion.

Notice that $\phi(G)$ could be as small as $\frac{1}{n^2}$ for a simple graph, and so the worst case approximation ratio could be as bad as $\Omega(n)$.

This does not explain the success of this algorithm in applications in image segmentation and data clustering.

There are difficult approaches to explain this phenomenon in practice.

One approach is to show that spectral partitioning works very well in the planted random instances, explaining its performance in some average-case sense.

Another direction is to prove upper bounds on the eigenvalues of special graph classes, e.g. it is proved that $\lambda_2 = O(\frac{d}{n})$ for planar graphs with maximum degree d , and so the spectral partitioning algorithm will find a set with small conductance.

Both directions are well-studied with very nice results; see project page for further references.

There are some notions of "stability" of a solution to capture practical instances in clustering.

Roughly speaking, an instance is stable if there is an outstanding sparse cut S (stable solution) such that any good sparse cut is similar to S (e.g. in terms of symmetric difference).

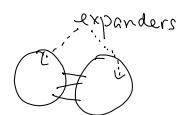
This arguably captures the "meaningful" instances for image segmentation and data clustering.


We work on a related question of similar spirit.

Suppose $\phi(G)$ is small but $\phi_2(G)$ is large. Can we prove that spectral partitioning works better?

The assumption is saying that there is a good way to cut the graph into two parts, but there is no good way to cut the graph into three parts.

So, the graph should look like a sparse cut separating two expanders.



Then, since the two parts are expanders, to minimize the Rayleigh quotient, the best embedding should map vertices in the same expander to similar values, and it should look like , and so the sweep algorithm should find the sparse cut.

The condition $\Phi_3(G)$ is large is combinatorial and is not as easy to use.

By the higher-order Cheeger's inequality, $\Phi_3(G)$ is large if and only if λ_3 is large.

So, we ask the question whether spectral partitioning works better when λ_k is large.

The following theorem is a strengthening of Cheeger's inequality using higher eigenvalues.

Theorem [KLLT13] For any $k \geq 2$, the spectral partitioning algorithm outputs a set S with $\phi(S) = O\left(\frac{k\lambda_2}{\lambda_k}\right)$.

This shows that the spectral partitioning algorithm is a $O\left(\frac{k}{\lambda_k}\right)$ -approximation algorithm for any $k \geq 2$.

For example, if λ_{10} is a constant, then it is a constant factor approximation algorithm.

In practical instances of image segmentation and data clustering, it is usually true that λ_k is large for a small k , and this provides a better explanation of its empirical success.

Note that the inequality is tight for cycles, up to a small constant factor.

Proof ideas

We don't directly follow the above intuition to prove the theorem, as it involves too many transitions between combinatorial properties and the spectral characterization.

The above intuition suggests that λ_k is large would imply that the second eigenvector looks like a k -step function (i.e. a vector with k distinct values).

The proof consists of two steps.

① If the second eigenvector is "close" to a k -step function, then the rounding algorithm performs better.

Intuitively, this should be true, as ideally we hope that the vector is binary-valued and this would correspond to a cut, and now the vector has only k distinct values (think of k as a small number),

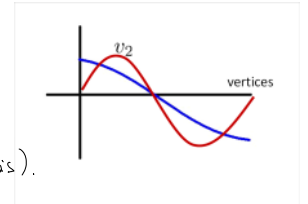
It is instructive to think about how to do a better analysis when the vector is an exact k -step function.

② The second step is a contrapositive to the above intuition.

We will prove that if the second eigenvector doesn't look like a k -step function, then λ_k must be small.

To get some intuition, look at the cycle example, if the second eigenvector is very "smooth" (i.e. the points are distributed evenly), then it is possible to find an orthogonal vector with similar Rayleigh quotient, proving that λ_3 is also small.

The reason that the orthogonal vector has small Rayleigh quotient is because every edge is of similar length as in the original vector (term-by-term analysis).



We will prove the following two lemmas corresponding to the two steps.

Lemma 1 (2k-step approximation) For any vector $x \in \mathbb{R}^n$ with $\|x\|=1$, there is a vector $y \in \mathbb{R}^n$ with only $2k$ distinct values such that $\|x-y\|^2 \leq O\left(\frac{R(x)}{\lambda_k}\right)$.

Lemma 2 (2k-step rounding) Let $y \in \mathbb{R}^n$ be a vector with only $2k$ distinct values. Let $x \in \mathbb{R}^n$ with $\|x\|=1$. The sweep algorithm applied on x will give a set S with $\phi(S) \leq O(kR(x) + k\sqrt{R(x)}\|x-y\|)$.

It is easy to see that combining the two lemmas with a vector x with $R(x) = O(\lambda_2)$ will prove the theorem, because $\phi(S) \leq O(k\lambda_2 + k\sqrt{\lambda_2}\|x-y\|) \leq O(k\lambda_2 + k\sqrt{\lambda_2}\sqrt{\frac{\lambda_2}{\lambda_k}}) = O\left(\frac{k\lambda_2}{\sqrt{\lambda_k}}\right)$.

One technical remark: We will assume that x is a non-negative vector with $R(x) \leq \lambda_2$ and $|\text{supp}(x)| \leq n/2$.

This can be achieved by the truncation argument used in LO3.

2k-step approximation

To prove Lemma 1, we won't use the idea of constructing orthogonal vectors with small Rayleigh quotient, as these are difficult to reason about.

Instead, we will use the strategy to construct k disjointly supported vectors $\psi_1, \dots, \psi_k \in \mathbb{R}^n$ such that $R(\psi_i)$ is not much larger than $R(x)$ if x is "smooth".

The claim that we use is the following, which is a generalization of the easy direction of higher-order Cheeger's inequality, whose proof is left as a homework problem.

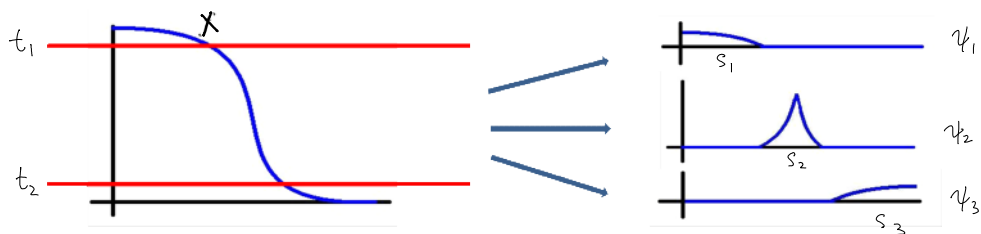
Claim If $\psi_1, \dots, \psi_k \in \mathbb{R}^k$ are vectors with disjoint support, then $\frac{1}{k}\lambda_k \leq \max_{1 \leq i \leq k} R(\psi_i)$.

Given a vector x that is "smooth", we will construct from it k disjointly supported vectors with small

Rayleigh quotient, and use the above claim to conclude that λ_k is small.

To get the idea, let's consider the case for $k=3$.

We will pick two threshold values $t_1 \geq t_2$ and partition the vertices into three groups.



Let $S_1 = \{i \mid x_i \geq t_1\}$, $S_2 = \{i \mid t_1 > x_i \geq t_2\}$, and $S_3 = \{i \mid t_2 > x_i \geq 0\}$.

Each ψ_ℓ is a vector with support in S_ℓ , defined as follows: $\psi_\ell(i) = \begin{cases} \min\{|x_i - t_1|, |x_i - t_2|\} & \text{if } i \in S_\ell \\ 0 & \text{otherwise} \end{cases}$.

We want to show that if x is smooth, then $R(\psi_\ell)$ is small for $1 \leq \ell \leq 3$.

Again, we will use a term-by-term analysis to compare $R(x)$ and $R(\psi_\ell)$.

By our construction of ψ_ℓ , it is clear that $|\psi_\ell(i) - \psi_\ell(j)| \leq |x_i - x_j|$, and so the numerator of $R(\psi_\ell)$ is no larger than that of $R(x)$.

For the denominator, the idea is that if x is smooth, then we can choose t_1, t_2 such that the denominator of ψ_ℓ is not too small compared to that of the denominator of $R(x)$.

To do this, we choose t_1, t_2 so that the denominators of ψ_ℓ are the same, i.e. $\|\psi_1\|^2 = \|\psi_2\|^2 = \|\psi_3\|^2 = C$.

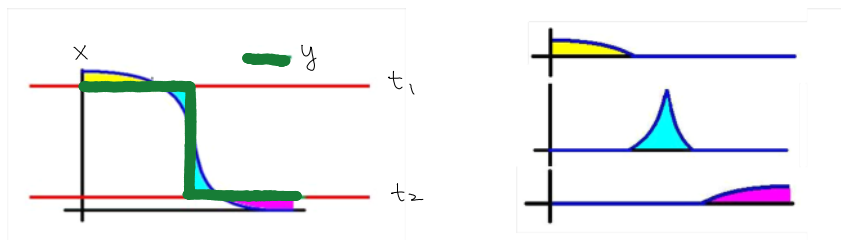
$$\begin{aligned} \text{Then, } R(\psi_\ell) &= \frac{\sum_{i,j} |\psi_\ell(i) - \psi_\ell(j)|^2}{d \|\psi_\ell\|^2} \leq \frac{\sum_{i,j} |x_i - x_j|^2}{dC} \quad (\text{by construction } |\psi_\ell(i) - \psi_\ell(j)| \leq |x_i - x_j|) \\ &= \frac{\lambda_2 - d \|x\|^2}{dC} = \frac{\lambda_2}{C} \quad (\text{by assumption } \|x\| = 1). \end{aligned}$$

By the claim, this implies that $\frac{\lambda_3}{2} \leq \max_{1 \leq \ell \leq 3} R(\psi_\ell) \leq \frac{\lambda_2}{C}$, and thus $C \leq \frac{2\lambda_2}{\lambda_3}$.

Now, notice that if we use t_1 and t_2 as a two step approximation y of x ,

$$\begin{aligned} \text{then } \|x - y\|^2 &= \sum_{\ell=1}^3 \|\psi_\ell\|^2 \\ &= 3C \leq \frac{6\lambda_2}{\lambda_3}, \end{aligned}$$

proving Lemma 1 in this case.



If we follow the same argument for general k , then we use $k-1$ thresholds t_1, \dots, t_{k-1} to partition the vertices into k groups, and define ψ_ℓ in the same way.

The same argument would imply that $C \leq \frac{2\lambda_2}{\lambda_k}$, but this only imply that $\|x - y\|^2 = kC \leq \frac{2k\lambda_2}{\lambda_k}$.

with an additional factor k compared to Lemma 1.

To remove this extra factor k , we use a simple trick that is also useful in higher-order Cheeger's inequality.

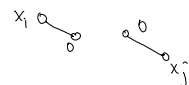
We use $2k-1$ thresholds $t_1 \geq t_2 \geq \dots \geq t_{2k-1}$ to divide the vertex set into $2k$ groups, and we choose

these thresholds such that $\|\psi_\ell\|^2 = C$ for all $1 \leq \ell \leq 2k$.

Let us sort the ψ_ℓ by the numerator such that $\sum_{i,j} \|\psi_1(i) - \psi_1(j)\|^2 \leq \sum_{i,j} \|\psi_2(i) - \psi_2(j)\|^2 \leq \dots \leq \sum_{i,j} \|\psi_{2k}(i) - \psi_{2k}(j)\|^2$.

The point is that $\sum_{\ell=1}^{2k} \sum_{i,j} \|\psi_\ell(i) - \psi_\ell(j)\|^2 \leq \sum_{i,j} \|x_i - x_j\|^2$, (think about each edge's contribution)

and thus $\sum_{i,j} \|\psi_\ell(i) - \psi_\ell(j)\|^2 \leq \frac{1}{2k} \sum_{i,j} \|x_i - x_j\|^2 \quad \forall 1 \leq \ell \leq 2k$.



Therefore, for $1 \leq \ell \leq 2k$, $R(\psi_\ell) = \frac{\sum_{i,j} \|\psi_\ell(i) - \psi_\ell(j)\|^2}{d \|\psi_\ell\|^2} \leq \frac{\frac{1}{2k} \sum_{i,j} \|x_i - x_j\|^2}{dC} = \frac{\frac{1}{2k} \lambda_2 d \|x\|^2}{dC} = \frac{\lambda_2}{2kC}$.

Hence, $\frac{\lambda_k}{2} \leq \max_{1 \leq \ell \leq 2k} R(\psi_\ell) \leq \frac{\lambda_2}{2kC}$, and thus $C \leq \frac{\lambda_2}{k\lambda_k}$.

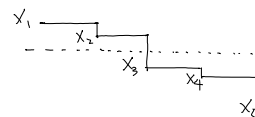
It follows that $\|x - y\|^2 = \sum_{\ell=1}^{2k} \|\psi_\ell\|^2 = 2kC \leq \frac{2\lambda_2}{\lambda_k}$, proving Lemma 1.

2k-step rounding

Ideal case

It is instructive to work out the ideal case when x is an exact $2k$ -step function, i.e.

there are only $2k$ distinct values $x_1 \geq x_2 \geq \dots \geq x_{2k} > 0$ in x .



Naturally, we like to cut in places where $x_i - x_{i+1}$ is large,

as the edges crossing must be long, and so there can't be too many edges.

Let $S_\ell := \{i \mid x_i \geq x_\ell\}$.

We output S_ℓ with probability $(x_\ell - x_{\ell+1})^2$, assuming $\sum_{\ell=1}^{2k} (x_\ell - x_{\ell+1})^2 = 1$ by scaling.

We will show that $\frac{\mathbb{E}[|S|]}{\mathbb{E}[d|S|]} \leq kR(x)$ and conclude that there exists ℓ with $\phi(S_\ell) \leq kR(x)$.

$$\mathbb{E}[|S|] = \sum_{i,j} \Pr(i,j \text{ is cut}) = \sum_{i,j} \sum_{\ell=1}^{i-1} (x_\ell - x_{\ell+1})^2 \leq \sum_{i,j} (x_i - x_j)^2$$



$$a^2 + b^2 + c^2 \leq (a+b+c)^2$$

$$\mathbb{E}[d|S|] = \sum_{i \in V} d \Pr(i \text{ in } S) = \sum_{i \in V} d \sum_{\ell=1}^{2k} (x_\ell - x_{\ell+1})^2$$

$$\geq d \sum_{i \in V} \frac{((x_\ell - x_{\ell+1}) + (x_{\ell+1} - x_{\ell+2}) + \dots + (x_{2k} - 0))^2}{2k}$$

$$= d \sum_{i \in V} \frac{x_i^2}{2k}$$

(Cauchy Schwarz says $\sum_{i=1}^k a_i^2 \geq \left(\sum_{i=1}^k a_i\right)^2 / k$)

Therefore, $\frac{\mathbb{E}[|S|]}{\mathbb{E}[d|S|]} \leq \frac{\sum_{i,j} (x_i - x_j)^2}{\sum_{i \in V} x_i^2} = 2kR(x)$, proving Lemma 2 in this ideal case.

$$= d \sum_{i \in V} \frac{1}{2k}.$$

Therefore, $\frac{E[|S|]}{E[d|S|]} \leq \frac{\sum_{i,j} (x_i - x_j)^2}{d \sum_{i \in V} \frac{x_i^2}{2k}} = 2k R(x)$, proving Lemma 2 in this ideal case.

General case

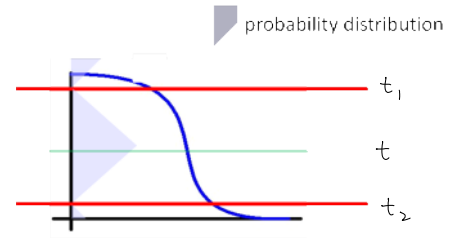
In the general case, we are only given a vector x that is "close" to a $2k$ -step vector y .

There is an elegant way to generalize the above argument.

We pick t with probability proportional to $\min_{1 \leq l \leq 2k} \{ |t - t_l| \}$

(in words, the distance to the closest step), and

output $S = \{ i \mid x_i \geq t \}$.

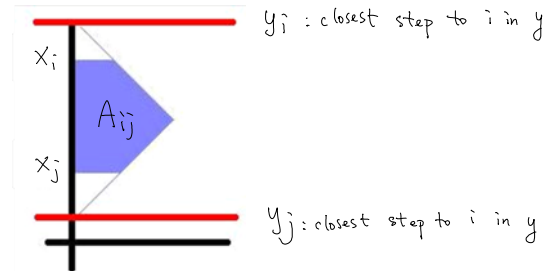


Without loss, we assume these probabilities sum to one by scaling x .

Let's analyze $E[|S|]$.

$$E[|S|] = \sum_{i,j} \Pr(i,j \text{ is cut})$$

$$\leq \sum_{i,j} \text{area}(A_{ij}) \quad (\text{this is worst possible: if there is another threshold in the middle then the area is } \frac{1}{2} \text{ of the big triangle})$$



big triangle

the two small triangles on the sides

$$= \sum_{i,j} \frac{1}{4} (|x_i - y_i| + |x_i - x_j| + |x_j - y_j|)^2 - \frac{1}{2} (x_i - y_i)^2 - \frac{1}{2} (x_j - y_j)^2$$

$$\leq \sum_{i,j} \frac{1}{4} [|x_i - x_j|^2 + 2|x_i - x_j| (|x_i - y_i| + |x_j - y_j|)]$$

$$= \frac{1}{4} R(x) \cdot d \|x\|^2 + \frac{1}{2} \sum_{i,j} |x_i - x_j| (|x_i - y_i| + |x_j - y_j|)$$

$$= \frac{1}{4} R(x) d \|x\|^2 + \frac{1}{2} \sqrt{\sum_{i,j} (x_i - x_j)^2} \sqrt{\sum_{i,j} (|x_i - y_i| + |x_j - y_j|)^2}$$

Cauchy-Schwarz

$$= \frac{1}{4} R(x) d \|x\|^2 + \frac{1}{2} \sqrt{R(x) \cdot d \|x\|^2} \sqrt{\sum_{i,j} (2(x_i - y_i)^2 + 2(x_j - y_j)^2)}$$

$$(a+b)^2 \leq 2a^2 + 2b^2$$

$$= \frac{1}{4} R(x) d \|x\|^2 + \frac{1}{2} \sqrt{R(x) \cdot d \|x\|^2} \sqrt{\sum_{i \in V} 2d (x_i - y_i)^2}$$

$$= \frac{d}{4} R(x) + \frac{d}{\sqrt{2}} \sqrt{R(x)} \|x - y\|$$

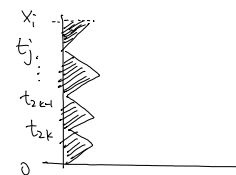
$$E[|S|] = \sum_{i \in V} \Pr(i \text{ in } S)$$

$$= \sum_{i \in V} [\text{area below } x_i]$$

$$\geq \sum_{i \in V} \frac{1}{4} (x_i - t_j)^2 + \sum_{l=j}^{2k} \frac{1}{4} (t_l - t_{l+1})^2$$

$$\geq \sum_{i \in V} \frac{1}{4} \left((x_i - t_j) + (t_j - t_{j+1}) + \dots + (t_{2k-1} - t_{2k}) + (t_{2k} - 0) \right)^2 / 2k$$

by Cauchy Schwarz
 $\sum_{i=1}^k a_i^2 \geq \left(\sum_{i=1}^k a_i \right)^2 / k$



$$\geq \sum_{i \in V} \frac{1}{4} \left((x_i - t_j) + (t_j - t_{j+1}) + \dots + (t_{2k-1} - t_{2k}) + (t_{2k} - 0) \right)^2 / 2k \quad \text{by Cauchy Schwarz}$$

$$= \sum_{i \in V} \frac{x_i^2}{8k} = \frac{1}{8k} \quad \text{by our assumption } \|x\|^2 = 1.$$

$$\sum_{i=1}^k a_i^2 \geq \left(\sum_{i=1}^k a_i \right)^2 / k$$

Therefore, $\frac{\mathbb{E}[|\delta(s)|]}{\mathbb{E}[d(s)]} \leq 2kR(x) + 4\sqrt{2}k\sqrt{R(x)}\|x-y\|$, proving Lemma 2.

Related results

Using some ideas from analysis of mixing time of random walks, it is possible to show that the spectral partitioning algorithm performs better when the robust vertex expansion is large.

Let $\Phi(S) = \min_{T \subseteq V-S} \left\{ |T| \mid E(S, T) \geq \frac{1}{2} E(S, \bar{S}) \right\} / |S|$, and $\Phi(G) = \min_{S \subseteq V, |S| \leq |V|/2} \Phi(S)$.

Theorem $\phi(G) \leq O(\lambda_2 / \Phi(G))$.

Also, one can prove a tight bound using $\phi_k(G)$.

Theorem $\phi(G) \leq O(k\lambda_2 / \phi_k)$ for any $k \geq 2$.

These show that the spectral partitioning algorithm is a $1/\Phi(G)$ and a k/ϕ_k -approximation algorithm.

Similar techniques can be applied to improve the analysis of higher-order Cheeger's inequality and the analog of Cheeger's inequality for bipartiteness ratio.

For example, one could prove that $\phi_k(G) \leq O(\text{poly}(k) \frac{\lambda_k}{\sqrt{\lambda_{k+1}}})$.

This says that if there is a large gap between λ_k and λ_{k+1} , the spectral k -way partitioning performs better. Indeed, in this case, the embedding is closer to the good case we discussed last time.

In practical application of clustering, one often does not know k , and the eigengap heuristic is to partition the data in k groups if $\lambda_{k+1} - \lambda_k$ is large, so the above result provides some justification of this heuristic.

For max cut, for expander graphs, one could prove that if the optimal solution cuts $1-\epsilon$ fraction of edges, then the spectral algorithm in L_4 can cut $1-O(\epsilon)$ fraction of edges (instead of $1-O(\sqrt{\epsilon})$ fraction).

References

[KLLO13] Improved Cheeger's inequality : analysis of spectral partitioning algorithms through higher order spectral gap. by Kwok, Lau, Lee, Oveis Gharan, Trevisan, 2013.

[KLL16] Improved Cheeger's inequality and analysis of local graph partitioning using vertex expansion and expansion profile, by Kwok, Lau, Lee, 2016.