CS 860 Spectral graph theory, Spring 2019, Waterloo.

Lecture 4: Higher order Cheeger's inequality

We see a recent generalization of Cheeger's inequality relating the k-th eigenvalue of the normalized Laplacian matrix to partitioning the graph into k disjoint sparse cuts.

Higher eigenvalues and graph multiway partitioning

Again, for simplicity, we assume the graph is d-regular.

Let $\mathcal{L} = \frac{1}{d} L$ be the normalized Laplacian matrix, with eigenvalues $0 = \lambda_1 \le \lambda_2 \le \dots \le \lambda_n \le 2$.

It is an exercise to show that $\lambda_k = 0$ if and only if the graph has at least k connected components.

In other words, $\lambda_k = 0$ if and only if there are disjoint subsets $S_1, S_2, \dots, S_k \subseteq V$ with $\phi(S_i) = 0$, where $\phi(S)$ denotes the conductance of S.

Similar to the pattern that we have seen in previous lectures, we will show that this spectral characterization allows us to prove a quantitative generalization, that λ_k is small if and only if there are disjoint subsets $S_1, S_2, \dots, S_k \subseteq V$ all with small conductance.

Let $\phi_k(G) = \min_{\substack{S_1, S_2, \dots S_k \subseteq V \\ S_i \cap S_j = \phi}} \max_{1 \le i \le k} \phi(S_i)$ be the k-way conductance of the graph G.

The following result is called the higher-order Cheeger's inequality.

Theorem [LOT12] $\frac{\lambda_k}{2} \le \phi_k(G) \le O(k^2)\sqrt{\lambda_k}$.

Another research group also proved a similar result using different techniques, which shows that the dependency on k can be improved considerably if we construct slightly fewer disjoint sparse cuts.

Theorem [LOT12, LRTV12] $\phi_k(G) \le O(\text{polylog}(k)) \sqrt{\lambda_{2k}}$.

Both results were inspired by the result in [ABS10] that if λ_k is small for a large k, then there exists a set S with $\phi(S) \approx O(\sqrt{\lambda_k})$ and $|S| \approx |V|/k$.

We may study the result in [ABS10] later when we talk about random walks in graphs.

The direction $\frac{1}{2}\lambda_k \le \phi_k(G)$ is called the easy direction, and the other direction is the hard direction.

Again, the easy direction can be seen as showing that λ_k is a relaxation of the k-way partitioning problem, while the hard direction can be seen as a rounding algorithm for the relaxation.

We will only prove the hard direction, and leave the easy direction as a homework problem.

Spectral embedding

We have proved the following result when we proved Cheeger's inequality.

Lemma (Cheeger's rounding) Given $x \in \mathbb{R}^n$, there is an efficient algorithm (the "sweep" algorithm) that finds $S \subseteq supp(x)$ such that $\phi(S) \le \sqrt{2R(x)}$, where $supp(x) := \{ i \mid x_i \neq 0 \}$ is the support of vector x and $R(x) := \frac{x^T \mathcal{L} x}{x^T x}$ is the Rayleigh quotient of x.

If λ_k is small, then we know that there are k orthogonal eigenvectors each with small Rayleigh quotient.
We can apply Cheeger's rounding on each eigenvector to find a sparse cut.
(Note that to apply the above lemma, we don't need x to be an eigenvector, just that $R(x)$ is small.)
Since the k vectors are orthogonal, intuitively, we expect that the sparse cuts produced by Cheeger's rounding on them are quite different (e.g. imagine the k vectors are all $\{0,1\}$-vectors).
So, this suggests that there are k different sparse cuts in the graph, and it should be possible to combine them to find k disjoint sparse cuts in the graph.
It is not clear how to proceed with this idea, by dealing with each vector independently.

Instead, we take a more global view that considers all the k vectors at the same time.
Let U be an $n \times k$ matrix where the i-th column is the i-th eigenvector, i.e. $U = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_k \\ | & | & & | \end{bmatrix} = \begin{bmatrix} — U_1 — \\ — U_2 — \\ \vdots \\ — U_n — \end{bmatrix}$
For each vertex $j \in V$, we map j to the j-th row of U, denoted by U_j.
We think of this maps vertex j to a point in the k-dimensional space.
This is what we mean by the spectral embedding of the graph.

The spectral embedding is actually used in pratical heuristics for graph k-way partitioning for some time.
One popular heuristic is to apply some well-known geometric clustering algorithms, in particular the k-means algorithm, to partition the point set in the spectral embedding into k groups, and use this partitioning to cut the graph into k sets.
It is reported that this heuristic works very well in applications, such as image segmentation and data clustering, but no theoretical guarantees were known about these heuristics.
The proofs of the higher-order Cheeger's inequality can be seen as proving rigorously the performance

guarantee of some variants of this type of algorithms, justifying the use of the spectral embedding.

<u>Isotropy condition</u>

For the spectral partitioning to provide useful information for us to find k disjoint sparse cuts, a necessary condition is that the points should be reasonably well spread out (e.g. if all the vertices are only mapped into only two distinct points in $\mathbb{R}^k$, then we have no clue how to find k disjoint sparse cuts for $k > 2$).

As you may expect, this kind of bad cases would not happen because the k vectors are orthogonal. The question is how to use the orthogonality to give us useful conditions for clustering.

Recall that U is the $n \times k$ matrix where the i-th column is the i-th eigenvector, and U_j denotes the j-th row of U, the k-dimensional point associated to vertex j in the spectral embedding.

Then $U^T U = I$, because the columns of U are orthonormal

Note that $U^T U = I$ can also be written as $\sum_{i=1}^{n} U_i U_i^T = I$.

We say the vectors $U_1, \ldots, U_n \in \mathbb{R}^k$ are in isotropy condition for the following reason.

<u>Claim</u> (isotropy condition) For any $x \in \mathbb{R}^k$ with $\|x\| = 1$, we have $\sum_{i=1}^{n} \langle x, U_i \rangle^2 = 1$.

<u>proof</u> $U^T U = I \Rightarrow x^T U^T U x = x^T x = 1 \Rightarrow 1 = x^T \left(\sum_{i=1}^{n} U_i U_i^T \right) x = \sum_{i=1}^{n} x^T U_i U_i^T x = \sum_{i=1}^{n} \langle x, U_i \rangle^2$. □

This claim says that for any direction $x \in \mathbb{R}^k$, the sum of the square of the projection of x_i on u is equal to one..

To get an idea, suppose $U_1 = U_2 = \cdots = U_\ell = y \in \mathbb{R}^k$ (that the first ℓ vertices all mapped to the same point y), then the claim implies that $1 = \sum_{i=1}^{n} \langle \frac{y}{\|y\|}, U_i \rangle^2 \geq \sum_{i=1}^{\ell} \langle \frac{U_i}{\|U_i\|}, U_i \rangle^2 = \sum_{i=1}^{\ell} \|U_i\|^2$.

On the other hand, $\sum_{i=1}^{n} \|U_i\|^2 = \sum_{i=1}^{k} \|v_i\|^2 = k$, where v_i is the i-th eigenvector (of unit length).

Therefore, if we think of the <u>mass</u> of a point i is $\|U_i\|^2$, then the above discussion says that the total mass mapped to the same point y is one, while the total mass is k.

That is, only $\frac{1}{k}$ fraction of the total mass is mapped to the same point.

This, for example, shows that the vertices cannot be mapped to less than k distinct points, and that's why the bad cases that we mentioned won't happen in the spectral embedding.

Generalizing this idea, suppose $U_1, U_2, \ldots, U_\ell$ are all pointing to similar direction, such that $\cos\theta_{ij} \geq 1-\varepsilon$.

By setting $x = U_1$, the isotropy condition tells us that

$$1 = \sum_{i=1}^{n} \left\langle \frac{U_1}{\|U_1\|}, U_i \right\rangle^2 \geq \sum_{i=1}^{\ell} \frac{1}{\|U_1\|^2} \langle U_1, U_i \rangle^2 = \sum_{i=1}^{\ell} \frac{1}{\|U_1\|^2} \|U_1\|^2 \|U_i\|^2 \cos^2 \theta_{1i} \geq \sum_{i=1}^{\ell} \|U_i\|^2 (1-\varepsilon)^2 \geq (1-2\varepsilon) \sum_{i=1}^{\ell} \|U_i\|^2.$$

This implies that $\sum_{i=1}^{\ell} \|U_i\|^2 \leq \frac{1}{1-2\varepsilon}$. When ε is small, the RHS is close to one.

This is saying that points in very similar direction can carry at most $\approx \frac{1}{k}$ fraction of total mass.

So, by using the direction of the points to guide us to cluster the points, we see that the points will be reasonably well spread out.

Clustering using directions

We formalize the above discussions into a lemma.

For a vertex i, the <u>mass</u> of i is defined as $\|U_i\|^2$.

Recall that the total mass is $\sum_{i=1}^{n} \|U_i\|^2 = k$.

For two vertices i, j, we define the <u>distance</u> $d(i,j)$ as $\left\| \frac{U_i}{\|U_i\|} - \frac{U_j}{\|U_j\|} \right\|$ (instead of Euclidean distance).

Note that $d(i,j) \approx \theta_{ij}$, where θ_{ij} is the angle between U_i and U_j.

<u>Lemma</u> Let $S \subseteq V$ be such that $d(i,j) \leq \Delta$ for all $i,j \in S$.

Then $\sum_{i \in S} \|U_i\|^2 \leq \frac{1}{1-\Delta^2}$.

<u>proof</u> For two unit vectors u, v, $\|u-v\|^2 = \|u\|^2 + \|v\|^2 - 2\langle u,v\rangle = 2 - 2\langle u,v\rangle$, and so $\langle u,v\rangle = 1 - \frac{d^2(u,v)}{2}$.

Pick an arbitrary vertex $j \in S$ and consider the unit vector $U_j / \|U_j\|$.

The isotropy claim says that
$$1 = \sum_{i=1}^{n} \left\langle \frac{U_j}{\|U_j\|}, U_i \right\rangle^2 \geq \sum_{i \in S} \|U_i\|^2 \left\langle \frac{U_j}{\|U_j\|}, \frac{U_i}{\|U_i\|} \right\rangle^2$$
$$= \sum_{i \in S} \|U_i\|^2 \left(1 - \frac{d^2(i,j)}{2}\right)^2 \quad \text{(by the above)}$$
$$\geq \sum_{i \in S} \|U_i\|^2 \left(1 - \frac{\Delta^2}{2}\right)^2 \quad \text{(by the assumption)}$$
$$\geq \sum_{i \in S} \|U_i\|^2 (1 - \Delta^2).$$
□

To apply the lemma, we will choose Δ so that $\frac{1}{1-\Delta^2} \leq 1 + \frac{1}{2k}$ (say when $\Delta = \frac{1}{2\sqrt{k}}$).

We will only take a subset S with diameter Δ (i.e. $d(i,j) \leq \Delta \;\; \forall i,j \in S$).

Then, the lemma implies that each such subset has mass at most $1 + \frac{1}{2k}$.

So, after taking $k-1$ subsets, the remaining mass is still at least $\frac{1}{2}$, a $\Omega(\frac{1}{k})$ fraction of the total mass.

This will ensure that we can form k groups of $\Omega(\frac{1}{k})$ fraction of mass by clustering using directions.

This distance function (using directions) is used in a practical heuristic before [NJW 01].

Disjoint Clusters

The previous section shows that bad cases won't happen if we do the clustering based on directions.

In this section, we think about the good case when the spectral embedding really gives what we want, k clusters that are far apart from each other.

Suppose there are k disjoint subsets $S_1, S_2, \ldots, S_k$ such that each has mass 1 and $d(S_i, S_j) \geq \delta \;\forall i,j$

where $d(S_i, S_j) := \min\{ d(a,b) \mid a \in S_i,\ b \in S_j \}$.

Can we conclude that these correspond to k disjoint sparse cuts in the graph?

Rayleigh Quotients

Let $v_1, v_2, \ldots, v_n \in \mathbb{R}^k$ be the spectral embedding of the graph.

We can define the Rayleigh quotient of the spectral embedding as $\dfrac{\sum_{i \sim j} \| v_i - v_j \|^2}{d \sum_{i \in V} \|v_i\|^2}$.

Note that $\|v_i - v_j\|^2 = \sum_{l=1}^{k} \left(v_i(l) - v_j(l) \right)^2$ (sum of the k coordinates)

$= \sum_{l=1}^{k} \left(v_l(i) - v_l(j) \right)^2$ (the l-th coordinate is the l-th eigenvector)

Let $R(v_l) = \dfrac{\sum_{i \sim j} \left(v_l(i) - v_l(j) \right)^2}{d \sum_{i \in V} v_l(i)^2} \overset{\text{def}}{=} \dfrac{A_l}{B_l}$, where A_l and B_l are just shorthands of the numerator and denominator.

Then observe that the Rayleigh quotient of the spectral embedding is $\dfrac{\sum_{l=1}^{k} A_l}{\sum_{l=1}^{k} B_l} \leq \max_{1 \leq l \leq k} R(v_l) \leq \lambda_k$.

On the other hand, if we have a spectral embedding in $\mathbb{R}^k$ with Rayleigh quotient $\leq \alpha$,

then we can also find a one dimensional embedding with Rayleigh quotient $\leq \alpha$.

The reasoning is similar: we write the Rayleigh quotient of the spectral embedding as $\dfrac{\sum_{l=1}^{k} A_l}{\sum_{l=1}^{k} B_l}$,

where A_l, B_l are the sums of the l-th coordinate.

Then, $\min_l \dfrac{A_l}{B_l} \leq \dfrac{\sum_{l=1}^{k} A_l}{\sum_{l=1}^{k} B_l}$, and so the best coordinate has Rayleigh quotient no larger than that of the k-dimensional embedding.

So, from a k-dimensional embedding, we can find a sparse cut by first restricting to the best coordinate, and then use Cheeger rounding in that coordinate to obtain a sparse cut.

The plan

Now we are ready to analyze the good case.

From the above discussion, we know that the Rayleigh quotient of the spectral embedding is $\leq \lambda_k$.

In the good case, there are k disjoint subsets $S_1, S_2, \ldots, S_k \subseteq V$, each has mass 1, and $d(S_i, S_j) \geq \delta \;\; \forall i, j$.

We would like to construct k spectral embeddings $\psi_1, \psi_2, \ldots, \psi_k$, such that ψ_ℓ is only supported on S_ℓ (i.e. in ψ_ℓ, every vertex $i \notin S_\ell$ is mapped to the zero vector).

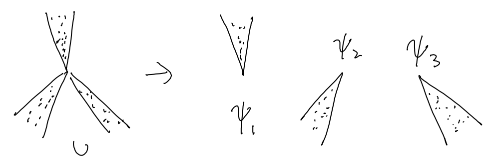
If we can do this in such a way that Rayleigh quotient of ψ_ℓ is small for all $1 \leq \ell \leq k$, then we can apply the argument in the above discussion to reduce to the best coordinate in ψ_ℓ

Then, we can apply the Cheeger rounding lemma to obtain a sparse cut, with support on S_ℓ.

The important question is how to construct ψ_ℓ and upper bound its Rayleigh quotient.

The most natural way to define ψ_ℓ is simply to zero out everything outside S_ℓ:

$$\psi_\ell(i) = \begin{cases} U_i & \text{if } i \in S_\ell \\ 0 & \text{otherwise.} \end{cases}$$



The Rayleigh quotient of ψ_ℓ is
$$\frac{\sum_{i \sim j} \|\psi_\ell(i) - \psi_\ell(j)\|^2}{d \sum_{i \in V} \|\psi_\ell(i)\|^2}$$
$$= \frac{\sum_{i \sim j,\, i \in S_\ell,\, j \in S_\ell} \|U_i - U_j\|^2 + \sum_{i \sim j,\, i \in S_\ell,\, j \notin S_\ell} \|U_i\|^2}{d \sum_{i \in S_\ell} \|U_i\|^2}$$

Compare it to the Rayleigh quotient of the spectral embedding of U.

Since $\sum_{i \in S} \|U_i\|^2 = 1$ (by our assumption of the good case) and $\sum_{i \in V} \|U_i\|^2 = k$, the denominator of ψ_ℓ is $\frac{1}{k}$ times the denominator of U.

Clearly, each term in the first summation of ψ_ℓ is at most that in U.

The interesting case is for those edges with $i \in S$ and $j \notin S$.

For those edges, it contributes $\|U_i\|^2$ in ψ_ℓ while it contributes $\|U_i - U_j\|^2$ in U.

By our assumption, $\theta_{ij} \approx d(i,j) \geq \delta$, the smallest $\|U_i - U_j\|^2$ one could get is $\|U_i\|^2 \sin^2\theta_{ij} \approx \|U_i\|^2 \theta_{ij}^2 \geq \|U_i\|^2 \delta^2$.

Therefore, each term in the numerator of ψ_ℓ is at most $\frac{1}{\delta^2}$ times that of the corresponding term in U, and thus the numerator of ψ_ℓ is at most $\frac{1}{\delta^2}$ times the numerator of U.

Combining, the Rayleigh quotient of ψ_ℓ is at most $\frac{k}{\delta^2}$ times the Rayleigh quotient of U, and thus $R(\psi_\ell) \leq \frac{k\lambda_k}{\delta^2}$.

So, by Cheeger's rounding, we get a sparse cut $S_\ell' \subseteq S_\ell$ with $\phi(S_\ell') \leq \frac{1}{\delta}\sqrt{2k\lambda_k}$.

Assuming δ is a constant, then we get k disjoint sparse cuts of conductance $O(\sqrt{k\lambda_k})$.

Smooth localization

In the general case, we use a similar strategy.

We would like to find k disjoint subsets $S_1, S_2, \ldots, S_k \subseteq V$ such that

① $\sum_{i \in S_\ell} \| v_i \|^2 \geq \Omega(1)$, or equivalently $\sum_{i \in S_\ell} \|v_i\|^2 \geq \Omega(\frac{1}{k}) \sum_{i=1}^{n} \|v_i\|^2$ for all $1 \leq \ell \leq k$.

② $d(S_i, S_j) \geq 2\delta \quad \forall\, 1 \leq i \neq j \leq k$, where $d(S_i, S_j) = \min_{a \in S_i, b \in S_j} d(a,b)$.

If we can do this, then we would like to define k disjoint supported functions $\psi_1, \psi_2, \ldots, \psi_k : V \to \mathbb{R}^k$ with Rayleigh quotient $R(\psi_\ell) \leq \frac{k\lambda_k}{\delta^2}$ as above.

There is one technical issue here.

There are some points $j \notin S_1 \cup \ldots \cup S_k$ but are very close to some point $i \in S_\ell$

In this case, when we define ψ_ℓ by zeroing out everything outside S_ℓ.

The length of this edge in ψ_ℓ is $\|\psi_\ell(i) - \psi_\ell(j)\|^2 = \|v_i\|^2$, which could be much larger than $\|v_i - v_j\|^2$.

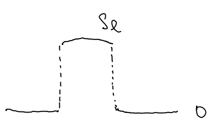
The ratio $\frac{\|\psi_\ell(i) - \psi_\ell(j)\|}{\|v_i - v_j\|}$ could be unbounded and the term-by-term analysis does not work.

To handle this issue, we use the condition that $d(S_i, S_j) \geq 2\delta$ to give us some room to "smoothly" decrease the length of the points close to S_ℓ to zero.

Formally, for each $j \notin S_\ell$, let $d(j, S_\ell) = \min_{i \in S_\ell} d(i,j)$.

We define $c_j = \max\left\{ 1 - \frac{d(j, S_\ell)}{\delta}, 0 \right\}$ and $\psi_\ell(j) = c_j v_j$.

So, if $d(j, S_\ell) \geq \delta$, then $\psi_\ell(j) = 0$; if $d(j, S_\ell) \leq \delta$, c_j decreases linearly with a slope $\frac{1}{\delta}$



sharp localization by zeroing out

smooth localization with slope $\frac{1}{\delta}$

By doing smooth localization, the following claim shows that the same term-by-term analysis still works.

Claim (smooth localization) $\|\psi_\ell(i) - \psi_\ell(j)\| \leq (1 + \frac{2}{\delta}) \|v_i - v_j\|$ for all $ij \in E$.

Proof Note that $|c_i - c_j| \leq \frac{1}{\delta} d(i,j)$, as $d(j,S) - d(i,S) \leq d(i,j)$.

Now, $\|\psi_\ell(i) - \psi_\ell(j)\| = \| c_i v_i - c_j v_j \| = \| c_i v_i - c_j v_i + c_j v_i - c_j v_j \|$

$$\leq |c_i - c_j| \|v_i\| + |c_j| \|v_i - v_j\| \leq |c_i - c_j| \|v_i\| + \|v_i - v_j\|.$$

Note that the first term $|c_i - c_j| \|v_i\| \leq \frac{1}{\delta} d(i,j) \|v_i\| = \frac{1}{\delta} \left\| \frac{v_i}{\|v_i\|} - \frac{v_j}{\|v_j\|} \right\| \|v_i\| = \frac{1}{\delta} \left\| v_i - \frac{\|v_i\|}{\|v_j\|} v_j \right\|$

$$\leq \frac{1}{\delta}\left(\|v_i - v_j\| + \left\| v_j - \frac{\|v_i\|}{\|v_j\|} v_j \right\| \right) \leq \frac{2}{\delta} \|v_i - v_j\|. \quad \square$$

With this (technical) claim, we are about to complete the proof.

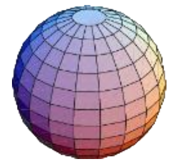
Partitioning space

The difficult case seems to be when the points distributed evenly in the space, in which it is not clear how to find the disjoint sets $S_1, \dots, S_k$ with the required properties.

The idea is to partition the directions in S^{k-1} (the k-dimensional sphere) into cubes each with side length $L = \frac{1}{2k}$.

(This cube argument is presented in [LL3].)

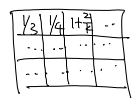
All the points with directions in the same cube are put into a block.



The diameter in each cube is $L\sqrt{k} = \frac{1}{2\sqrt{k}}$.

By the isotopy lemma, the points in each block has mass $\leq 1 + \frac{1}{2k}$.

To construct disjoint $S_1, \dots S_k \subseteq V$ where each S_i has mass $\geq \frac{1}{2}$, we just greedily group the blocks so that their mass is at least $1/2$.



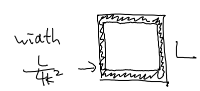
Since no block is too heavy, it is always possible.

Finally, we need to guarantee that the sets are δ-far apart.

To do this, for each cube, we delete all the directions that are of distance at most $\frac{L}{4k^2}$ from some side, and then we delete all the points with those directions in the corresponding block.

By doing so, the fraction of the volume of each cube remained is

$$\left(1 - \frac{1}{4k^2}\right)^k \geq \left(1 - \frac{1}{4k}\right).$$



Therefore, if we choose a uniformly random translation of the partition, the expected total mass removed is only $1/4$, and this is small enough so that it doesn't affect the grouping argument.

Now, the sets S_i are at least δ-far apart with $\delta = \frac{L}{2k^2} = \frac{1}{4k^3}$

So, following our plan (with smooth localization), we can construct disjoint supported functions, each with Rayleigh quotient $O\left(\frac{k\lambda_k}{\delta^2}\right) = O(k^7 \lambda_k)$, and hence we get k disjoint subsets with conductance $O(k^{3.5}\sqrt{\lambda_k})$ by Cheeger rounding.

Discussions

- The tighter bound $O(k^2 \sqrt{\lambda_k})$ is obtained by some random partitioning technique developed in metric embedding.
- The bound $\phi_k(G) \leq O(\text{polylog}(k) \sqrt{\lambda_{2k}})$ is proved by using some dimension reduction technique

to reduce the spectral embedding from k-dimensional space to $O(\log k)$-dimensional space.

This bound is tight as shown by the noisy hypercube example (we might talk about later).

- The algorithm in [LRTV12] is very simple. Compute the spectral embedding. Generate k random directions $r_1, \dots, r_k$. Put point j into cluster ℓ if $\langle v_j, r_\ell \rangle \geq \langle v_j, r_i \rangle \;\forall i$.

  This is it, but the analysis is less intuitive.

- It is still open whether $\phi_k(G) \leq O(\text{polylog}(k) \sqrt{\lambda_k})$.

References

[ABS10] Subexponential algorithms for unique games and related problems, by Arora, Barak, Steurer.

[LOT12] Multiway spectral partitioning and higher order Cheeger inequalities, by Lee, Oveis Gharan, Trevisan.

[LRTV12] Many sparse cuts via higher eigenvalues, by Louis, Raghavendra, Tetali, Vempala.

[NJW02] On spectral clustering: analysis and an algorithm, by Ng, Jordan, Weiss.

[L13] No frills proof of higher order Cheeger's inequality, blog post by Lee.