University of Edinburgh

INFR11156: Algorithmic Foundations of Data Science (2019)

Lecture 12: The Cheeger Inequality

For any undirected graph $G = (V, E)$ and a set $S \subset V$, let

$$h_G(S) = \frac{|\partial S|}{\min\{\mathrm{vol}(S), \mathrm{vol}(V \setminus S)\}},$$

where $\mathrm{vol}(S) = \sum_{u\in} d_u$ and $\partial S = E(S, V \setminus S)$ denotes the set of edges with one endpoint in S and the other endpoint in $V \setminus S$. The **Cheeger constant** or the **conductance** of graph G is defined as

$$h_G = \min_S h_G(S).$$

By definition, the set S achieving h_G corresponds to the sparsest cut in G, and finding such set S has numerous applications in computer science. For instance, when analysing a physical network, one can view the servers and the links connecting different servers as the vertices and edges in G. Hence, a higher value of h_G shows that the underlying network is more reliable, since one have to remove many links to make the network disconnected. Moreover, as the number of edges corresponds to the construction cost, it is desired to construct a network G with higher value of h_G while in the mean time keeping the number of edges in G as small as possible. For image segmentation, a common approach is to construct a graph G based on the RGB values and the pairwise distances among different pixels, and the sparsest cuts in G are used to identify different objects in a picture. While we can formulate a sparse cut in different ways with respect to different settings, one of the simplest formulations is as follows:

Problem 1 (**The Sparsest Cut Problem**). *Given an undirected graph $G = (V, E)$ of n vertices as input, find a set $S \subset V$ such that $h_G(S) = h_G$.*

The sparest cut problem is NP-hard, and the current best approximation algorithm achieves approximation ratio $O(\sqrt{\log n})$, which is based spectral geometry and semi-definite programming. Designing approximation algorithms for the sparsest cut problem is one of the most central problems in approximation algorithms.

In this lecture, we will see how h_G relates to λ_2, which is polynomial-time computable, and design an approximation algorithm for the sparsest cut problem. We will also briefly discuss the high-order generalisation of the Cheeger inequality.

1 The Cheeger Inequality

We have seen several equivalent formulations for λ_2 from the last lecture. From these formulations, we can write λ_2 as the minimum of a function $g(x)$ over possible $x \in \mathcal{D} \subseteq \mathbb{R}^n$, and $g(x)$ for any $x \in \mathcal{D}$ gives an upper bound of λ_2. Now, we use the same method to show that λ_2 can be upper bounded with respect to h_G.

Lemma 2. $\lambda_2 \leq 2 \cdot h_G$.

Proof. Let $C = (A, B)$ be the optimal cut that achieves h_G, and let $|C|$ be the number of edges in this cut. We define a vector $x \in \mathbb{R}^n$ such that $x_u = 1/\operatorname{vol}(A)$ if $u \in A$, and $x_u = -1/\operatorname{vol}(B)$ of $u \in B$. Since

$$\langle x, D\mathbf{1} \rangle = \sum_{u \in A} \frac{d_u}{\operatorname{vol}(A)} - \sum_{u \in B} \frac{d_u}{\operatorname{vol}(B)} = 0,$$

it holds that

$$\begin{aligned}
\lambda_2 &\leq \frac{\sum_{u \sim v} (x_u - x_v)^2}{\sum_u d_u \cdot x_u^2} = \frac{|C| \cdot (1/\operatorname{vol}(A) + 1/\operatorname{vol}(B))^2}{1/\operatorname{vol}(A) + 1/\operatorname{vol}(B)} \\
&= |C| \cdot \left(\frac{1}{\operatorname{vol}(A)} + \frac{1}{\operatorname{vol}(B)} \right) \leq \frac{2|C|}{\min\{\operatorname{vol}(A), \operatorname{vol}(B)\}} = 2 \cdot h_G,
\end{aligned}$$

which proves the statement. □

Next, we will show that h_G can be upper bounded with respect to λ_2 as well.

Theorem 3 (Cheeger Inequality)**.** *It holds that* $h_G \leq \sqrt{2 \cdot \lambda_2}$.

The core behind the proof of the Cheeger inequality is the following fact, which corresponds to an approximation algorithm for finding a sparse cut. For any vector $y \in \mathbb{R}^n$, we assume that $y_1 \leq \ldots \leq y_n$. For any $t \in \mathbb{R}$, define

$$S_t = \{u : \ y_u \leq t\}.$$

We call these $\{S_t\}_{t=1}^n$ **sweep sets**.

Lemma 4. *For any vector y satisfying $y^\intercal D\mathbf{1} = 0$, there is a number t such that*

$$h_G(S_t) \leq \sqrt{2 \cdot \frac{y^\intercal L y}{y^\intercal D y}}.$$

Notice that the vector $y = D^{-1/2} f_2$ satisfies

$$\frac{y^\intercal L y}{y^\intercal D y} = \frac{f_2^\intercal D^{-1/2} L D^{-1/2} f_2}{f_2^\intercal D^{-1/2} D D^{-1/2} f_2} = \frac{f_2^\intercal \mathcal{L} f_2}{f_2^\intercal f_2} = \lambda_2.$$

Hence, based on Lemma 4 we have the following Algorithm 1, whose output is a set S satisfying $h_G(S) \leq \sqrt{2 \cdot \lambda_2}$.

Proof of Lemma 4. Let

$$\rho = \frac{y^\intercal L y}{y^\intercal D y} = \frac{\sum_{u \sim v} (y_u - y_v)^2}{\sum_u d_u \cdot y_u^2}.$$

Without loss of generality, we assume that $y_1 \leq \ldots \leq y_n$, and let j be the smallest number such that $\sum_{i \leq j} d_i \geq \operatorname{vol}(G)/2$.

We introduce another vector $z \in \mathbb{R}^n$ such that $z_u = y_u - y_j$, and hence $z_j = 0$. Moreover, it is easy to show that

$$\frac{z^\intercal L z}{z^\intercal D z} = \frac{y^\intercal L y}{y^\intercal D y + \operatorname{vol}(G) \cdot y_j^2} \leq \rho.$$

We further scale vector z such that $z_1^2 + z_n^2 = 1$, and define set

$$V_t = \{u : z_u \leq t\}.$$

Algorithm 1 Algorithm for finding a sparse cut

$f = D^{-1/2} f_2$
Sort all the vertices such that $f(u_1) \leq \ldots \leq f(u_n)$
$t = 0$
$S = \emptyset$
$S^\star = \{u_1\}$
while $t \leq n$ **do**
  $t = t + 1$
  $S = S \cup \{u_t\}$
  if $h_G(S) \leq h_G(S^\star)$ **then** $S^\star = S$
  end if
end while
return $S^\star$

Since
$$h_G(V_t) = \frac{|\partial V_t|}{\min\{\mathrm{vol}(V_t), \mathrm{vol}(V \setminus V_t)\}},$$
our goal is to define a distribution on t such that
$$\frac{\mathbb{E}\left[|\partial V_t|\right]}{\mathbb{E}\left[\min\{\mathrm{vol}(V_t), \mathrm{vol}(V \setminus V_t)\}\right]} \leq \sqrt{2\rho}, \tag{1}$$
Notice that (1) is equivalent to show that $\mathbb{E}\left[\sqrt{2\rho} \cdot \min\left\{\mathrm{vol}(V_t), \mathrm{vol}(V \setminus V_t)\right\} - |\partial V_t|\right] \geq 0$. Therefore, there is a set V' such that $\sqrt{2\rho} \cdot \min\left\{\mathrm{vol}(V'), \mathrm{vol}(V \setminus V')\right\} \geq |\partial V'|$, i.e.,
$$\frac{|\partial V'|}{\min\left\{\mathrm{vol}(V'), \mathrm{vol}(V \setminus V')\right\}} \leq \sqrt{2\rho}.$$

To define such a distribution, we choose t according to the probability density function $2|t|$. Hence, the probability that a value between $[a, b]$ is chosen is
$$\mathbb{P}\left[t \in [a, b]\right] = \int_a^b 2|t| \mathrm{d}t = \mathrm{sgn}(b) \cdot b^2 - \mathrm{sgn}(a) \cdot a^2.$$
Since $z_1^2 + z_n^2 = 1$, we have that
$$\mathbb{P}\left[t \in [z_1, z_n]\right] = \int_{z_1}^{z_n} 2|t| \mathrm{d}t = \mathrm{sgn}(z_n) \cdot z_n^2 - \mathrm{sgn}(z_1) \cdot z_1^2 = 1.$$
So it suffices to analyse $\mathbb{E}\left[\min\left\{\mathrm{vol}\left(V_t\right), \mathrm{vol}(V \setminus V_t)\right\}\right]$ and $\mathbb{E}\left[|\partial V_t|\right]$.

Analysis of $\mathbb{E}\left[\min\left\{\mathrm{vol}\left(V_t\right), \mathrm{vol}(V \setminus V_t)\right\}\right]$. Notice that
$$\mathbb{E}\left[\mathrm{vol}\left(V_t\right)\right] = \sum_u \mathbb{P}\left[z_u \leq t\right] \cdot d_u.$$
By the choice of j, we know that $t < 0$ implies that $\mathrm{vol}(V_t) < \mathrm{vol}(G)/2$, while $t > 0$ implies that $\mathrm{vol}(V \setminus V_t) \leq \mathrm{vol}(G)/2$. Hence, it holds that
$$\begin{aligned}
&\mathbb{E}\left[\min\left\{\mathrm{vol}\left(V_t\right), \mathrm{vol}(V \setminus V_t)\right\}\right] \\
&= \sum_u \mathbb{P}\left[z_u \leq t \text{ and } t < 0\right] \cdot d_u + \sum_u \mathbb{P}\left[z_u > t \text{ and } t \geq 0\right] \cdot d_u \\
&= \sum_{u:\ z_u \leq t} d_u \cdot z_u^2 + \sum_{u:\ z_u > t} d_u \cdot z_u^2 \\
&= z^\intercal D z.
\end{aligned}$$

Analysis of $\mathbb{E}\left[|\partial V_t|\right]$. Notice that an edge $u \sim v$ with $z_u \le z_v$ is in ∂V_t iff $z_u \le t$ and $z_v \ge t$. This event occurs with probability

$$\int_{z_u}^{z_v} 2|t|\mathrm{d}t = \mathrm{sgn}(z_v) \cdot z_v^2 - \mathrm{sgn}(z_u) \cdot z_u^2,$$

which equals to $|z_u^2 - z_v^2|$ if $\mathrm{sgn}(z_u) = \mathrm{sgn}(z_v)$, and $z_u^2 + z_v^2$ otherwise. We upper bound both terms by the inequality

$$|z_u^2 - z_v^2| \le |z_u - z_v| \cdot \left(|z_u| + |z_v|\right),$$

and

$$z_u^2 + z_v^2 \le (z_u - z_v)^2 \le |z_u - z_v| \cdot \left(|z_u| + |z_v|\right).$$

Then, it holds that

$$\begin{aligned}
\mathbb{E}\left[|\partial V_t|\right] &= \sum_{\{u,v\} \in E} \mathbb{P}\left[z_u \le t \text{ and } z_v > t\right] \\
&\le \sum_{u \sim v} |z_u - z_v| \cdot \left(|z_u| + |z_v|\right) \\
&\le \sqrt{\sum_{u \sim v} |z_u - z_v|^2} \cdot \sqrt{\sum_{u \sim v} \left(|z_u| + |z_v|\right)^2} \\
&\le \sqrt{z^\intercal L z} \cdot \sqrt{2 \cdot z^\intercal D z},
\end{aligned}$$

where the second inequality follows by the Cauchy-Schwarz inequality. Therefore, we have that

$$\frac{\mathbb{E}\left[|\partial V_t|\right]}{\mathbb{E}\left[\min\{\mathrm{vol}(V_t), \mathrm{vol}(V \setminus V_t)\}\right]} \le \sqrt{2 \cdot \frac{z^\intercal L z}{z^\intercal D z}} \le \sqrt{2\rho}.$$

Therefore, there is a set V_t such that $h_G(V_t) \le \sqrt{2\rho}$. □

2 Further discussions

Are these inequalities tight? Combining the Cheeger inequality with Lemma 2, we have that

$$\lambda_2/2 \le h_G \le \sqrt{2 \cdot \lambda_2}. \tag{2}$$

The following two examples show that both sides of (2) are tight up to a constant factor.

- For a path graph P_n, the Cheeger constant is $\frac{1}{\lceil (n-1)/2 \rceil}$, and

  $$\lambda_2 = 1 - \cos\left(\frac{\pi}{n-1}\right) \approx \frac{\pi^2}{2(n-1)^2}.$$

  This shows that the Cheeger inequality is tight up to a constant factor.

- For an n-cube on 2^n vertices, the Cheeger constant is $2/n$ which is equal to λ_2. Hence, the first inequality in (2) is tight within a constant factor as well.

Why is it called the Cheeger inequality? Theorem 3 was originally proven by Cheeger in the setting of manifolds. It was shown about 20 years later that the same inequality by Cheeger holds for graphs as well, and the proof for graphs essentially follows exactly from the proof for manifolds. However, it is worth pointing out that, the easier direction of (2), i.e., $\lambda_2/2 \le h_G$ does not hold for manifolds.

Graphs in which the sweep cut failed to find a sparse cut. There have been extensive studies about the graphs in which a sweep set algorithm based on f_2 fails to find a sparse cut. As an example, we define a grid graph as follows:

- There are $\sqrt{n}$ rows and $3\sqrt{n}$ columns in the grid, and there is a vertex at every crossing "point" between a horizontal line segment and a vertical line segment.
- The weight of every edge, except the edges sitting in the middle row, has weight 1.
- The weight of every edge sitting in the middle row has weight $1/\sqrt{n}$.

See Figure 1 for example. it is easy to see that the "horizontal cut" crossing the "thin" edges is the sparsest cut, while the output of a sweep set algorithm is the "vertical cut".

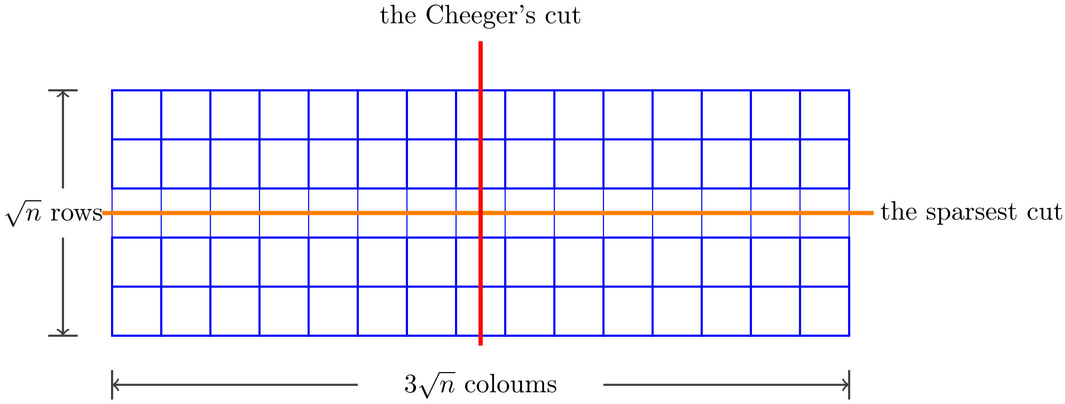

Figure 1: A grid graph with $\sqrt{n}$ rows, and $3\sqrt{n}$ columns.

3 Higher-order Cheeger Inequality

So far we have discussed the relations between λ_2 and h_G. Based on this, one can ask if the structure of multi-clusters in a graph relates to the other eigenvalues of $\mathcal{L}$. To build this connection, we generalise the Cheeger constant and define the k**-way expansion constant** as

$$\rho(k) \triangleq \min_{\text{partition } A_1,\ldots,A_k} \max_{1\leq i\leq k} h_G(A_i). \tag{3}$$

Here, we call subsets of vertices (i.e. **clusters**) $A_1,\ldots,A_k$ a k**-way partition** of G if $A_i \cap A_j = \emptyset$ for different i and j, and $\bigcup_{i=1}^{k} A_i = V$. Usually we say a graph G occurring in practice has k clusters if we can partition the vertex set of G into k subsets $A_1,\ldots,A_k$, such that different clusters are loosely connected, i.e., the value of $\rho(k)$ is small. It is known that $\rho(k)$ is related to λ_k by the following higher-order Cheeger inequality:

$$\frac{\lambda_k}{2} \leq \rho(k) \leq O(k^2)\sqrt{\lambda_k}. \tag{4}$$

At a very high level, the proof of the higher-order Cheeger inequality is to apply the eigenvectors associated with $\lambda_2,\ldots,\lambda_k$ to embed every vertex into a point in $\mathbb{R}^k$. We will discuss more about this approach when we discuss spectral clustering algorithms in later lectures.