CSE 521: Design and Analysis of Algorithms I **Winter 2017**

Lecture 12: Cheeger's Inequality cont., Spectral Clustering, Power Method

Lecturer: Shayan Oveis Gharan *02/22/17* *Scribe: Elizabeth Clark*

12.1 Cheeger's Inequality (continued)

12.1.1 Review from last class

Definition 12.1 (Conductance)**.** *Given a graph $G = (V, E)$ with V partitioned into S and $\bar{S}$, the conductance of S is defined as:*

$$\phi(S) = \frac{|E(S, \bar{S})|}{\text{vol}(S)}$$

The conductance of G is defined as:

$$\phi(G) = \min_{\text{vol}(S) \leq \frac{\text{vol}(V)}{2}} \phi(S)$$

Theorem 12.2 (Cheeger's Inequality)**.** *For any graph G,*

$$\frac{\lambda_2}{2} \leq \phi(G) \leq \sqrt{2\lambda_2}$$

Question: Why aren't we looking at the mincut as a measure of conductance? Why are we normalizing by the size/volume of S in $\frac{|E(S,\bar{S})|}{\text{vol}(S)}$?

Answer: Consider a network of roads. One road in this network is a highway that connects two major cities. Another is your driveway that connects your house to the rest of the network. If cut either of these roads, it will divide the network into two disconnected sub-networks, so these are two mincuts of our graph each with size 1. However, the two roads have very different levels of conductance. The numbers of drivers inconvenienced by the shutdown of a highway is much greater than those inconvenienced by the shutdown of your driveway.

12.1.2 Continuation of proof of Cheeger's Inequality

In this lecture we prove the easy direction of Cheeger's inequality, i.e., we show that, for any graph G,

$$\frac{\lambda_2}{4} \leq \phi(G). \tag{12.1}$$

Note that one can also show $\phi(G) \geq \lambda_2/2$, but here for the sake of simplcity we show the above weaker version.

For simplicity of the argument, we assume that G is d-regular. Recall that the normalized Laplacian matrix is defined as

$$\tilde{L} = D^{-1/2} L D^{-1/2} = L/d,$$

where the last equality holds because G is d-regular. So, the first eigenvector of $\tilde{L}$ is the all ones vector, $\mathbf{1}$ with eigenvalue 0.

So by Rayleigh quotient,

$$\lambda_2 = \min_{x:x\perp\mathbf{1}} \frac{x^T\tilde{L}x}{x^Tx} = \min_{x:x\perp 1} \frac{\sum_{i\sim j}(x_i - x_j)^2}{d\sum x_i^2}. \tag{12.2}$$

To prove (12.1), we need to relate this value to

$$\phi(G) = \min_{S:\mathrm{vol}(S)\le \mathrm{vol}(V)/2} \phi(S).$$

Let S be the best set in the RHS of above, i.e., assume $\phi(S) = \phi(G)$ and $\mathrm{vol}(S) \le \mathrm{vol}(V)/2$. We can write,

$$\begin{aligned}\phi(S) &= \frac{|E(S,\bar{(S)}|}{\mathrm{vol}(S)} \\ &= \frac{\sum_{i\sim j}|\mathbb{I}[i\in S] - \mathbb{I}[j\in S]|}{d\cdot|S|}.\end{aligned}$$

Note that $|\mathbb{I}[i\in S] - \mathbb{I}[j\in S]| = 1$ if and only if i,j lie on two different sides of the cut $(S,\overline{S})$. In the denominator of the above we used that G is d-regular, so $\mathrm{vol}(S) = d\cdot|S|$.

Recall that as usual

$$1_i^S = \begin{cases} 1 & i\in S \\ 0 & \text{otherwise.}\end{cases}$$

Therefore, we can write

$$\phi(S) = \frac{\sum_{i\sim j}|1_i^S - 1_j^S|}{d\sum 1_i^S} = \frac{\sum_{i\sim j}|\mathbf{1}_i^S - \mathbf{1}_j^S|^2}{d\cdot\sum_{i=1}^n \mathbf{1}_i^{S^2}}.$$

In the last equality we used the fact that $|\mathbf{1}_i^S - \mathbf{1}_j^S|$ is always 0 or 1; so its square is the same.

Note that the above equation is very similar to (12.2). Roughly speaking, in the above we are looking at the Rayleigh quotient for a specific vector $\mathbf{1}^S$ whereas the RHS of (12.2) is the minimum posible value of Rayleigh quotient over all vectors in $\mathbb{R}^n$. So, it seems that we should get $\phi(G) = \phi(S) \ge \lambda_2$.

This is however is not quite right: The minimum in the RHS of (12.2) is taken over all vectors orthogonal to the all ones vector, $\mathbf{1}$, and the vector $\mathbf{1}^S$ is not orthogonal to $\mathbf{1}$. In general, if we to make a given vector x orthogonal to a vector v we let

$$x = x - \langle x, \frac{v}{\|v\|}\rangle \frac{v}{\|v\|}.$$

So, in this case we want to $\mathbf{1}^S$ orthogonal to $\mathbf{1}$ we need to let

$$x = \mathbf{1}^S - \langle \mathbf{1}^S, \mathbf{1}/\|\mathbf{1}\|\rangle \mathbf{1}/\|\mathbf{1}\| = \mathbf{1}^S - (|S|/\sqrt{n})\mathbf{1}/\sqrt{n} = \mathbf{1}^S - \frac{|S|}{n}.$$

So, to prove (12.1), we need to show that

$$\frac{\sum_{i\sim j}|\mathbf{1}_i^S - \mathbf{1}_j^S|^2}{d\sum \mathbf{1}_i^{S^2}} \ge \frac{1}{4}\frac{\sum_{i\sim j}(x_i - x_j)^2}{d\sum x_i^2} \tag{12.3}$$

where $x = \mathbf{1}^S - \frac{|S|}{n}$. Note that the numerator is shift-invariant, i.e.,

$$\sum_{i\sim j}(x_i - x_j)^2 = \sum_{i\sim j}((x_i - c) - (x_j - c))^2$$

for any $c \in \mathbb{R}$. Therefore, the numerators of the left and right hand sides of (12.3) are equal. So, it is enough to show that

$$\|x\|^2 \geq \frac{1}{4}\|\mathbf{1}^S\|^2.$$

Note that the above inequality is not true if $S = V$. In fact, this is the only place in the proof that we use that $\text{vol}(S) \leq \text{vol}(V)/2$, i.e., $|S| \leq n/2$ in regular graphs. Here, we prove a weaker version of the above inequality.

Claim 12.3. *Let $|S| \leq n/2$ and $x = \mathbf{1}^S - |S|/n$. Then,*

$$\|x\|^2 \geq \frac{1}{4}\|\mathbf{1}^S\|.$$

Proof. First note that since $|S| \leq n/2$, we have $|S|/n \leq 1/2$. We can write,

$$\|x\|^2 = \sum x_i^2 \geq \sum_{i \in S} \left(1 - \frac{|S|}{n}\right)^2 \geq |S|(\frac{1}{2})^2 = \frac{|S|}{4},$$

where in the second inequality we used $|S|/n \leq 1/2$. □

This completes the proof of (12.3) which completes the proof of the easy direction of Cheeger's inequality (12.1).

We do not prove the following lemma; interested reader can see lecture notes of more advanced courses linked in the course website for the proof of the harder direction of Cheeger's inequality.

Lemma 12.4. *For all $x \perp 1$, the spectral partitioning algorithm returns S such that $\phi(S) \leq 2\sqrt{\frac{x^T \tilde{L} x}{x^T x}}$.*

The importance of the above lemma is that we don't need to find the actual eigenvector of λ_2 to use the spectral partitioning algorithm. As long as we can approximately minimize the Rayleigh quotient, $\frac{x^T \tilde{L} x}{x^T x}$, we can run the spectral partitioning algorithm on the approximate vector to obtain a set S of small conductance. In section 12.3 we will see how to find an approximate second eigenvector of the $\tilde{L}$ in almost linear time.

12.1.3 A Bad example for Spectral Partitioning Algorithm

Spectral Partitioning Algorithm does not always return the optimal solution, in fact it may return a set of a singificantly larger conductance than the optimum. Consider the following example.

Suppose we have the graph shown in Figure 12.1. Consider 2 possible cuts of this graph. Cut 1 (shown in red) will give a conductance value $\frac{4}{2n}$, or $O(\frac{1}{n})$. Cut 2 (shown in green) will give a conductance value of $\frac{n\frac{50}{n^2}}{2n}$, or $O(\frac{1}{n^2})$. While Cut 2 is much better than Cut 1, SPA will return Cut 1. This is because the 2nd smallest eigenvector of this graph is the same as the 2nd smallest eigenvector of a cycle, i.e., it maps the endpoints of each dashed edge to the same value. Because of that the algorithm indeed returns a cut whose conductance is n times the optimum.

12.2 Spectral Clustering Algorithm

This is a brief discussion of Ng, Jordan, and Weiss [**NJW02**] paper on spectral clustering.

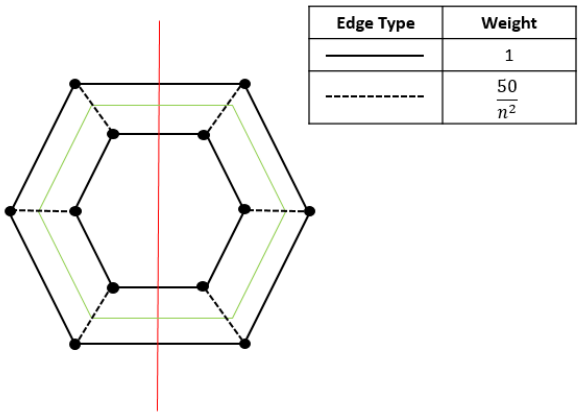

Figure 12.1: A weighted graph comprised of two cycles. The conductance of the red cut is n times the conductance of the green cut, but the spectral partitioning algorithm returns the red cut.

(a) Original data

(b) Data after spectral clustering

Figure 12.2: Spectral clustering: before and after

Motivating Example: Suppose you want to cluster a set of points, but your points look something like those depicted in Figure 12.2a. In this case, you want to find the green and black clusters. If you run k-means on this data, you won't find these clusters.

Instead, we can use SPA by creating a graph from this data by connecting points with an edge of weight

$$e^{\frac{-\|x_i - x_j\|^2}{\sigma^2}},$$

where x_i, x_j represents any two datapoints in $\mathbb{R}^d$. Note that the above Gaussian kernel is maximized if x_i is very close to x_j. The parameter σ must be tuned based on the particular application in mind.

After constructing this graph, we compute the normalized Laplacian matrix and the first k eigenvectors $v_1, v_2, \ldots, v_k$ of the matrix (since we want a k-partition of the graph).

Then we build the spectral embedding of graph, i.e., a matrix

$$F = \begin{bmatrix} D^{-\frac{1}{2}} v_1 \\ \vdots \\ D^{-\frac{1}{2}} v_k \end{bmatrix} \in \mathbb{R}^{k \times n},$$

which has a column for every vertex in the graph. Now, we map each vertex of graph (or each data point) i

to a point in k dimensions corresponding to the i-th column of the above matrix. It turns out that in this new mapping the each cluster of points will be mapped close to one another, see Figure 12.2b and we can use k-means to find the k partition. In **??** we give a rigorous analysis of (a varaint of) this algorithm; we show that for any graph G we can find k disjoint sets $S_1, \ldots, S_k$ each of conductance $O(\sqrt{lambda_k}k^2)$. In other words, this shows that if the graph that we construct from the data points has k small eigenvalues then we can use k means to find a k partitioning of the graph. Also, conversely, if the first k eigenvalues of G are not small, then there is no "good" k partitionings of G.

12.3 Power Method

We discussed in previous lectures that computing SVD takes cubic time in the size of the matrix. So, one in general is interested in faster algorithms for computing (approximating) eigenvalues/eigenvectors of a matrix. The Power Method is a method to approximate the largest eigenvalue of a PSD matrix M within a multiplicative $1 \pm \epsilon$ factor in time linear in the number of nonzero entries of M.

Recall that a Guassian vector $x \in \mathbb{R}^n$, is a vector of n independently chosen $\mathcal{N}(0,1)$ random variable, i.e., for all $1 \leq i \leq n$, $x_i \sim \mathcal{N}(0,1)$.

Algorithm 1 Power Method

Input: Given a PSD matrix $M \succeq 0$.

Choose a random Gaussian vector $x \in \mathbb{R}^n$.

for $j = 1 \to k$ **do**

$x \leftarrow Mx$ ▷ For numerical stability, set $x \leftarrow \frac{x}{\|x\|}$; we don't add it here to get a simpler proof.

end for

return x, $\frac{x^T M x}{x^T x}$

Let y be the output vector of Algorithm 1. In our main theorem we show that y is an approximate largest eigenvector of M.

Theorem 12.5. *Given a matrix $M \succeq 0$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$, for any $\epsilon > 0$ and integer $k > 1$ with constant probability,*

$$\frac{y^T M y}{yTy} \geq \frac{\lambda_1(1-\epsilon)}{1 + 10n(1-\epsilon)^{2k}}.$$

Note that ϵ is a parameter of choice in the above theorem (it has nothing to do with the algortihm). We should choose it based on the error that we can tolerate in our application. For a given ϵ, letting $k = \frac{\lg n}{\epsilon}$ in Algorithm 1 the RHS of the above theorem becomes $\frac{\lambda_1(1-\epsilon)}{1+\frac{1}{n}}$.

Also, observe that the algorithm runs a loop for k iterations; each iteration is just a matrix vector product which can be implemented in time $O(\text{nnz}(M)$. It follows that for any PSD matrix M we can use the above theorem to find a vector y such that the Rayleigh quotient of y is at least $(1-\epsilon)\lambda_1$. The algorithm will run in time $O(\frac{1}{\epsilon}\,\text{nnz}(M)\log n)$.

Before discussing the proof of the above theorem, let us discuss two remarks:

Remark 12.6 (2nd largest eigenvalue:)**.** *Suppose we want to estimate the 2nd largest eigenvalue of M. Then, we can first run the above algorithm to find an approximate largest eigenvector y. Then, we choose*

another random Gaussian vector x. First we make x orthogonal to y by letting:

$$x = x - \langle x, \frac{y}{\|y\|}\rangle \frac{y}{\|y\|}.$$

In other words, if $\|y\| = 1$, we let

$$x = x - \langle x, y\rangle y.$$

Remark 12.7 (Eigenvalues of Symmetric Matrices)**.** *Suppose that M is not PSD but it is a symmetric matrix. Then, we can run the above algorithm on M^2 which is a PSD matrix. The algorithm gives a $1 \pm \epsilon$ approximation of the largest eigenvalue of M in* absolute value*.*

Remark 12.8 (2nd Smallest eigenvalue of $\tilde{L}$)**.** *First of all, it turns out that the largest eigenvalue of $\tilde{L}$ is at most 2. Therefore, we can turn the smallest eigenvalues of $\tilde{L}$ into the largest ones by working with $2I - \tilde{L}$. Note that $2I - \tilde{L}$ is PSD, and the 2nd smallest eigenvalue of $\tilde{L}$ is the 2nd largest eigenvalue of $2I - \tilde{L}$. Now, all we need to do is to choose a Gaussian random vector x and make it orthogonal to the largest eigenvector of $2I - \tilde{L}$, and then use the power method*

Recall that the smallest eigenvector of $\tilde{L}$ is $v_1 = D^{1/2}\mathbf{1}$. This is because

$$v_1^T \tilde{L} v_1 = \mathbf{1}^T D^{1/2}(D^{-1/2}LD^{-1/2})D^{1/2}\mathbf{1} = \mathbf{1}^T L \mathbf{1} = \sum_{i\sim j}(\mathbf{1}_i - \mathbf{1}_j)^2 = 0.$$

Therefore, to find the 2nd smallest eigenvector of $\tilde{L}$ we do the following: Choose a random Gaussian vector x. Then, let

$$y = x - \langle x, v_1/\|v_1\|\rangle v/\|v_1\|,$$

where $v_1 = D^{1/2}\mathbf{1}$. Then calculate $(2I - \tilde{L})^k y$ as an approximation of the 2nd smallest eigenvalue of $\tilde{L}$. We will analyze this algorithm in the next lecture.

To prove the above theorem, we use the following 3 claims:

Claim 12.9. *For any Gaussian random vector $x \in R^n$ and any unit-norm vector $v \in \mathbb{R}^n$, we have*

$$\mathbb{P}\left[|\langle x, v\rangle| \geq \frac{1}{2}\right] \geq \Omega(1)$$

Proof. First of all observe that $\mathbb{E}\left[\langle x, v\rangle\right] = 0$ and

$$\mathbb{E}\left[\langle x, v\rangle^2\right] = \sum_{i,j} x_i v_i x_j v_j = \sum_{i,j} v_i v_j \mathbb{E}\left[x_i x_j\right] = \sum_i v_i^2 \mathbb{E}\left[x_i^2\right] = \sum_{i=1}^n v_i^2 = \|v\|^2 = 1.$$

Since $\langle x, v\rangle$ is a linear combination of independent normal random variables, it is also a normal random variable. So, from the above equations, we have $\langle x, v\rangle \sim \mathcal{N}(0, 1)$. But, it can be seen from the CDF of the standard normal random variable that if $g \sim \mathcal{N}(0, 1)$, then

$$\mathbb{P}\left[|g| \geq 1/2\right] \geq \Omega(1)$$

as desired. □

Letting $v = v_1$, the eigenvector of λ_1, we get that

$$\mathbb{P}\left[|\langle x, v_1\rangle| \geq \frac{1}{2}\right] \geq \Omega(1). \tag{12.4}$$

Claim 12.10. *For any Gaussian random vector $x \in \mathbb{R}^n$, we have*

$$\mathbb{P}\left[\|x\|^2 \leq 2n\right] \geq 1 - e^{\frac{-n}{8}}.$$

$\mathbb{P}\left[|\sum {x_i}^2 - n| > \epsilon\right] \leq e^{\frac{-\epsilon^2}{8n}}$. *In other words, square of the norm of a Gaussian random vector is at most $2n$ with high probability.)*

Proof. The proof follows from strong concentration bounds on sum of independent normal random variables. We use the following theorem withouth proof:

Theorem 12.11. *Let $g_1, \ldots, g_n \sim \mathcal{N}(0,1)$ be independent normal random variables. Then,*

$$\mathbb{P}\left[\left|\frac{1}{n}(g_1^2 + \cdots + g_n^2) - 1\right| \geq \epsilon\right] \leq e^{-n\epsilon^2/8}.$$

The proof of this is very similar to the proof of Chernoff/Hoeffding concentration inequalities and we are not going into the details. So, we can write

$$\mathbb{P}\left[|x_1^2 + \cdots + x_n^2 - n| \geq \epsilon\right] \leq e^{\epsilon^2/8n}.$$

Letting $\epsilon = n$ in the above proves the claim. □

Our last claim which finishes the proof of Theorem 12.5 is not probabilistic anymore.

Claim 12.12. *For all vectors $x \in \mathbb{R}^n$, $\epsilon > 0$ and and $y = M^k x$ we have*

$$\frac{y^T M y}{y^T y} \geq \frac{\lambda_1(1-\epsilon)}{1 + \frac{\|x\|^2}{\langle x, v_1\rangle^2}(1-\epsilon)^{2k-1}}.$$

Now, if x in the above claim is a Gaussian random vector, then with a constant probability $\frac{\|x\|^2}{\langle x, v_1\rangle^2} \leq 4n$ with a constant probability. This simply follows from (12.4) and Claim 12.10.

We will prove the last claim in the next lecture.