CSE 521: Design and Analysis of Algorithms I **Fall 2020**

Lecture 15: Cheeger's Inequality cont., Spectral Clustering

Lecturer: Shayan Oveis Gharan *11/23/20*

15.1 Cheeger's Inequality (continued)

15.1.1 Review from last class

Definition 15.1 (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:

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

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.

15.1.2 Cheeger's Inequality

The following theorem is one of the fundamental inequalities in spectral graph theory.

Theorem 15.2 (Cheeger's Inequality)**.** *For any graph G,*

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

where λ_2 is the 2nd smallest eigenvalue of $\tilde{L}$.

Cheeger's inequality relates the combinatorial property of conductance to a spectral property, the 2nd smallest eigenvalue. Observe that in the extreme case where $\lambda_2 = 0$, we also have $\phi(G) = 0$ and vice versa.

A crucial fact about the above inequality is that it does not depend on the size of G, n. It implies that if G has a "small" 2nd eigenvalue, then it is partitionable, whereas if the 2nd eigenvalue is "large" the graph is similar to a complete graph and it is not partitionable.

The proof of the right side of Cheeger's inequality, $\phi(G) \leq \sqrt{2\lambda_2}$ is constructive, and it shows that the spectral partitioning algorithm always returns a set S such that $\text{vol}(S) \leq \text{vol}(V)/2$ and

$$\phi(S) \leq \sqrt{2\lambda_2} \leq \sqrt{4\phi(G)}.$$

We now discuss several consequences of the above theorem for a special family of graphs.

Definition 15.3 (Expander Graphs)**.** *Expander graphs are sparse highly connected graphs with large 2nd eigenvalues, i.e., $\lambda_2 \geq \Omega(1)$. So, the can be seen as a sparse complete graphs which have $\lambda_2 = 1$. It turns out that most of the graphs are expanders, because a random d-regular graph satisfies $\lambda_2 \geq 1 - \frac{2}{\sqrt{d}}$*

Expander graphs are the easiest instances to use for many of the optimization problems (see PS4 for applications to the max cut problem). They are frequently used in coding theory (see PS4) and in psuedorandom number generators. In Problem Set 4, we will discuss the expander mixing lemma, which states that Expander Graphs are approximately the same as random graph. In words, in a d-regular expander graphs, for every disjoint large sets S, T, $|E(S,T)|$ is very close to the expected number of edges between S, T in a random $G(n, d/n)$ graph.

Definition 15.4 (Planar Graphs)**.** *A planar graph is one where all vertices can be projected onto a plane with no crossing edges.*

We know a lot about planar graphs. For example, we know that they tend to be sparse with average degrees at most 5.

It turns out that the 2nd eigenvalue of $\tilde{L}$ of any planar graph is at most $O(1/n)$.

Theorem 15.5. *If G is a bounded degree planar graph, the*

$$\lambda_2 \leq O(\frac{1}{n}).$$

Using the Cheeger's inequality, we can show that for every bounded degree planar graph G, $\phi(G) \leq O(1/\sqrt{n})$. In fact, by repeatedly peeling off sets of small conductance in G, we can show that every planar graph with bounded degree has a sparse bisection, i.e., a set $S \subseteq V$ such that $\text{vol}(V)/3 \leq \text{vol}(S) \leq 2\,\text{vol}(V)/3$ and $\phi(S) \leq O(1/\sqrt{n})$. This means that it is very easy to break a bounded degree planar graph into two sets such that there is very small number of connections between the two. This makes planar graphs ideal candidate for divide and conquer algorithms. We can recursively solve our problem on the two sides $S, \overline{S}$ and then merge the solutions. Since there are only $O(\sqrt{n})$ edges between $S, \overline{S}$, the merge operation can be done very efficiently

Finally, we also note that we can generalize the task of partitioning a graph into two sets, into partitioning a graph into k sets. Doing so, we define the k-way conductance of a graph to be:

Definition 15.6 (k-way conductance)**.** *For an integer $k > 1$ and a graph G, let*

$$\phi_k(G) = \min_{disjoint S_1,\ldots,S_k} \max_{1 \leq i \leq k} \phi(S)$$

where the min is over all k disjoint sets $S_1, \ldots, S_k$ in G.

In other words, we are interested in finding k disjoint sets such that their maximum conductance is as small as possible. With this definition, $\phi(G)$ corresponds to the case $k = 2$.

It turns out that there is a natural generalization of Cheeger's inequality for larger values of k.

Theorem 15.7. *For any graph G and an integer $k > 2$,*

$$\begin{aligned} \lambda_k/2 \le \phi_k(G) &\le O(\sqrt{\lambda_k} \cdot k^2) \\ \phi_k(G) &\le O(\sqrt{\log k \cdot \lambda_{2k}}). \end{aligned}$$

Note that the second inequality is stronger than the first one only when λ_{2k} is not much larger than λ_k. Similar to Cheeger's inequality, the proof of the right side of this inequality is constructive and provides an algorithm to k disjoint sets with small conductance.

15.1.3 Proof of "easier side" 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,

$$\frac{\lambda_2}{4} \le \phi(G). \tag{15.1}$$

Recall that the normalized Laplacian matrix is defined as $\tilde{L} = D^{-1/2}LD^{-1/2}$, So, the first eigenvector of $\tilde{L}$ is $D^{1/2}\mathbf{1}$ with eigenvalue 0. This is because,

$$\tilde{L}(D^{1/2}\mathbf{1}) = D^{-1/2}LD^{-1/2}D^{1/2}\mathbf{1} = D^{-1/2}L\mathbf{1} = D^{1/2}\mathbf{0} = \mathbf{0}.$$

By Rayleigh quotient,

$$\begin{aligned} \lambda_2 &= \min_{x: x \perp D^{1/2}\mathbf{1}} \frac{x^T \tilde{L} x}{x^T x} \\ &= \min_{x: x \perp D^{1/2}\mathbf{1}} \frac{x^T D^{-1/2} L D^{-1/2} x}{x D^{-1/2} D D^{-1/2} x} \\ &= \min_{\substack{x: \, x \perp D^{1/2}\mathbf{1} \\ y = D^{-1/2}x}} \frac{y^T L y}{y D y} \end{aligned}$$

Note that if $x \perp D^{1/2}\mathbf{1}$ then

$$0 = \langle x, D^{1/2}\mathbf{1} \rangle = \langle D^{1/2}y, D^{1/2}\mathbf{1} \rangle = \langle y, D\mathbf{1} \rangle$$

So, we can do a change variables as follows:

$$\lambda_2 = \min_{y: y \perp D\mathbf{1}} \frac{y^T L y}{y^T D y} = \min_{y: y \perp D\mathbf{1}} \frac{\sum_{i \sim j} (y_i - y_j)^2}{\sum_i d_i y_i^2}. \tag{15.2}$$

To prove (15.1), 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,

$$\begin{aligned} \phi(S) &= \frac{|E(S, \overline{S}|}{\mathrm{vol}(S)} \\ &= \frac{\sum_{i \sim j} |\mathbb{I}[i \in S] - \mathbb{I}[j \in S]|}{\sum_{i \in S} d_i} \\ &= \frac{\sum_{i \sim j} |\mathbb{I}[i \in S] - \mathbb{I}[j \in S]|^2}{\sum_i \mathbb{I}[i \in S]^2} \end{aligned}$$

To see the last identity note that the absolute value of the difference of two indicator functions is either 0 or 1, so $|\mathbb{I}[i \in S] - \mathbb{I}[j \in S]| = |\mathbb{I}[i \in S] - \mathbb{I}[j \in S]|^2$. As usual let

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

Note that the above equation is very similar to (15.2) except that $y = \mathbf{1}^S$ is not necessarily orthogonal to $D\mathbf{1}$. In particular,

$$\langle \mathbf{1}^S, D\mathbf{1} \rangle = \sum_i 1_i^S d_i = \sum_{i \in S} d_i = \mathrm{vol}(S) \neq 0 \tag{15.3}$$

So, we have to perturb the $\mathbf{1}^S$ vector and make it orthogonal to the all-ones vector. The idea is to shift it, i.e., add a constant c to all coordinates of the vector; note that $y^T L y$ is invariant under shifts, so we just need to show that the denominator does not change so much under a shift. Here is where we use that $\mathrm{vol}(S) \leq \mathrm{vol}(V)/2$.

Define $y = \mathbf{1}^S - \frac{\mathrm{vol}(S)}{\mathrm{vol}(V)}\mathbf{1}$. Then obviously, first note

$$\langle y, D\mathbf{1} \rangle = \sum_i y_i d_i = \sum_{i \in S} d_i - \sum_i d_i \frac{\mathrm{vol}(S)}{\mathrm{vol}(V)} = \mathrm{vol}(S) - \mathrm{vol}(S) = 0.$$

As explained above,

$$y^T L y = \sum_{i \sim j} (y_i - y_j)^2 = \sum_{i \sim j} (\mathbf{1}_i^S - \mathbf{1}_j^S)^2 = |E(S, \overline{S})|.$$

Finally,

$$y^T D y = \sum_i y_i^2 d_i \geq \sum_{i \in S} (1 - \mathrm{vol}(S)/\mathrm{vol}(V))^2 d_i \geq \sum_{i \in S} (1/2)^2 d_i = \mathrm{vol}(S)/4,$$

where in the second inequality we used that $\mathrm{vol}(S) \leq \mathrm{vol}(V)/2$. It follows that

$$\lambda_2 \leq \frac{y^T L y}{y^T D y} \leq \frac{|E(S, \overline{S})|}{\mathrm{vol}(S)/4} = 4\phi(S),$$

where the first inequality follows by (15.3) and that $\langle y, D\mathbf{1} \rangle = 0$.

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 15.8. *For all y such that $\langle y, 1 \rangle_D = 0$, the spectral partitioning algorithm returns S such that $\phi(S) \leq 2\sqrt{\frac{y^T L y}{\|y\|_D^2}}$.*

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, $\frac{y^T L y}{y D y}$, we can run the spectral partitioning algorithm on the approximate vector to obtain a set S of small conductance. In the next lecture we will see how to find an approximate second eigenvector of the $\tilde{L}$ in almost linear time.

15.1.4 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 significantly larger conductance than the optimum. Consider the following example.

Suppose we have the graph shown in Figure 15.1. 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.

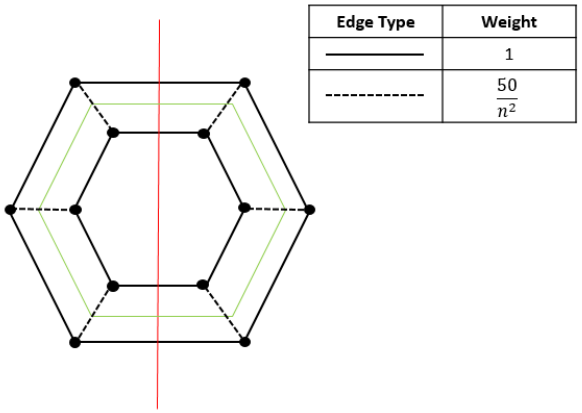

Figure 15.1: A weighted graph comprised of two cycles. The conductance of the red cut is n times the conductance of the green cut, but the spectral partitioning algorithm returns the red cut.

15.2 Spectral Clustering Algorithm

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

(a) Original data

(b) Data after spectral clustering

Figure 15.2: 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 15.2a. 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 15.2b 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.

15.3 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 15.9. *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 15.10 (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 15.11 (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 15.12 (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 $\tilde{L}$.

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

Claim 15.13. *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. Recall that by rotational invariance property of Gaussians, $\langle x, v\rangle$ is distributed as a $\mathcal{N}(0,1)$ random variable. It can be seen from the density function 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. In fact a normal is distributed almost uniformly in the interval $[-1, 1]$. Therefore, the probability that it is not in the $[-0.5, -0.5]$ is at least a constant say $1/3$. □

Claim 15.14. *For any Gaussian random vector $x \in \mathbb{R}^n$, we have*

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

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

Theorem 15.15. *Let $g_1, \ldots, g_n \sim \mathcal{N}(0,1)$ be independent normal random variables. Then,*

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

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 15.9.

Claim 15.16. *For all vectors $x \in \mathbb{R}^n$, $\epsilon > 0$ and and $y = M^k x$ we have*

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

Note that this claim holds for any vector x. Now, if x in the above claim is a Gaussian random vector, then, by Claim 15.13 and Claim 15.14, $\frac{\|x\|^2}{\langle x, v_1\rangle^2} \leq 4n$ with a constant probability.

Note that we needed randomness in the statement of Theorem 15.9 to ensure that the vector x has a significant inner product with the first eigenvector of M, with respect to its norm. So Gaussian distributions are not doing a fundamental role in this proof. For example, we could also choose coordinates of x to be uniformly and independently chosen from $\{-1, +1\}$ and almost a similar proof would follow.

Proof. By definition of y,

$$y^T M y = x^T M^k M M^k x = x^T M^{2k+1} x$$

recalling that M is PSD and thus symmetric. Similarly, $y^T y = x M^{2k} x$.

Suppose $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of M and $v_1, \ldots, v_n$ are the corresponding eigenvectors. Then $\lambda_1^{2k+1}, \ldots, \lambda_n^{2k+1}$ are eigenvalues of M^{2k+1}.

Let us divide the eigenvalues into two groups: $\lambda_1, \ldots, \lambda_j$ where all of these are greater than or equal to $(1-\epsilon)\lambda_1$, and $\lambda_{j+1}, \ldots, \lambda_n$ where all are less than $(1-\epsilon)\lambda_1$.

Let us first discuss the highlevel idea of the proof. For the sake of intuition assume that $k \gg \frac{\lg n}{\epsilon}$. Then we may note that $\lambda_{j+1}^k \leq \lambda_1^k(1-\epsilon)^k \leq \lambda_1^k(\frac{1}{n^2})$. It follows that $\sum_{i=j+1}^n \lambda_i^{2k} \leq n\lambda_j^{2k} \leq \frac{\lambda_1^{2k}}{n}$, meaning the total contribution of eigenvalues after j in the spectral decomposition of M^{2k} is very small – if k is large, then y is essentially in the span of $v_1, \ldots, v_j$, and that is all I need to prove the claim, because all of the first j eigenvalues are at least $(1-\epsilon)\lambda_1$.

Next, we do the algebra. First, let us expand the spectral decomposition of M^{2k+1}:

$$\begin{aligned} x^T M^{2k+1} x &= x^T \left(\sum_{i=1}^{n} \lambda_i^{2k+1} v_i v_i^T \right) x \\ &= \sum_{i=1}^{n} \lambda_i^{2k+1} \langle v_i, x \rangle^2 \\ &\geq \sum_{i=1}^{j} \lambda_i^{2k+1} \langle v_i, x \rangle^2 \\ &\geq \sum_{i=1}^{j} (1-\epsilon)\lambda_1 \lambda_i^{2k} \langle v_i, x \rangle^2 \end{aligned}$$

where in the last inequality we use the fact that $\lambda_1, \ldots, \lambda_j \geq (1-\epsilon)\lambda_1$. So this gives us a lower bound for $x^T M^{2k+1} x$.

Next, we derive an upper bound for $x^T M^{2k} x$. Putting these together we will lower bound the ratio $\frac{x^T M^{2k+1} x}{x^T M^{2k} x}$. Proceeding similarly to the above, we note

$$\begin{aligned} x^T M^{2k} x &= \sum_{i=1}^{n} \lambda_i^{2k} \langle v_i, x \rangle^2 \\ &= \sum_{i=1}^{j} \lambda_i^{2k} \langle v_i, x \rangle^2 + \sum_{i=j+1}^{n} \lambda_i^{2k} \langle v_i, x \rangle^2 \\ &\leq \sum_{i=1}^{j} \lambda_i^{2k} \langle v_i, x \rangle^2 + (1-\epsilon)^{2k} \lambda_1^{2k} \sum_{i=j+1}^{n} \langle v_i, x \rangle^2 \\ &\leq \sum_{i=1}^{j} \lambda_i^{2k} \langle v_i, x \rangle^2 + (1-\epsilon)^{2k} \lambda_1^{2k} \|x\|^2 \end{aligned}$$

where in the last inequality we used that $\lambda_{j+1}, \ldots, \lambda_n \leq (1-\epsilon)\lambda_1$ and that $\sum_{i=j+1}^{n} \langle v_i, x \rangle^2 \leq \|x\|^2$, since we are projecting x onto a set of at most n orthonormal vectors.

Now, substituting these bounds, we get

$$\begin{aligned} \frac{y^T M^k y}{y^T y} = \frac{x^T M^{2k+1} x}{x^T M^{2k} x} &\geq \frac{(1-\epsilon)\lambda_1 \sum_{i=1}^{j} \lambda_i^{2k} \langle v_i, x \rangle^2}{\sum_{i=1}^{j} \lambda_i^{2k} \langle v_i, x \rangle^2 + (1-\epsilon)^{2k} \lambda_1^{2k} \|x\|^2} \\ &= \frac{(1-\epsilon)\lambda_1}{1 + \frac{(1-\epsilon)^{2k} \lambda_1^{2k} \|x\|^2}{\sum_{i=1}^{j} \lambda_i^{2k} \langle v_i, x \rangle^2}} \\ &\geq \frac{(1-\epsilon)\lambda_1}{1 + \frac{(1-\epsilon)^{2k} \lambda_1^{2k} \|x\|^2}{\lambda_1^{2k} \langle v_1, x \rangle^2}} \\ &\geq \frac{(1-\epsilon)\lambda_1}{1 + (1-\epsilon)^{2k} \frac{\|x\|^2}{\langle x, v_1 \rangle^2}}. \end{aligned}$$

as desired □

References

[NJW02] A. Ng, M. Jordan, and Y. Weiss. "On spectral clustering: Analysis and an algorithm". In: *NIPS*. 2002 (cit. on p. 15-5).