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

Lecture 13: Power Method (cont.), Random Walks

Lecturer: Shayan Oveis Gharan *Feb. 27, 2017* *Scribe: Steven S. Lyubomirsky*

13.1 Aside: Graph Visualization With Eigenvectors

At the beginning of class we ran several simulations to show applications of spectral embedding of a graph; in particular we showed a 2-dimensional mapping of a graph G where vertex i is mapped to the points (x_i, y_i) where x, y represent the 2nd and 3rd eigenvectors of normalized Laplacian of G. vThere are theorems that can characterize the visualizations produced in this manner, for planar graphs. Note that while planar graphs are frequently described in terms of being those which can be drawn on a single plane with no edges intersections, it is not obvious how to produce this mapping; visualizing with the second and third eigenvectors can be a way to accomplish this.)

13.2 Continuation of Proof of Bounds on Power Method

Recall that in the last lecture, we were in the process of proving the following theorem:

Theorem 13.1. *For all $k > 0$ and $\epsilon > 0$, if y is the vector returned by the power method on PSD matrix M and λ_1 is the largest eigenvalue of matrix M, then with constant probability*

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

Recall that to generate y, we let $x \in \mathbb{R}^n$ be a random Gaussian vector, i.e., for all i, we sample $x_i \sim \mathcal{N}(0,1)$ independent of all other coordinates. Then, $y = M^k x$ in the above theorem.

Continuing the proof from before, we will argue that it will suffice to prove three claims related to the power method in order to complete the proof, which we will now state:

Claim 13.2. *Let x be a random Gaussian vector. For any unit norm vector v, with constant probability, $|\langle x, v\rangle| \geq 1/2$.*

We will invoke the above claim for $v = v_1$, the first eigenvector of M. Note that by a constant probability in the above claim we mean a number which is independent of n. That number only depends on the density function of the standard normal distribution, and it does not grow as $n \to \infty$.

We proved a very similar claim in the midterm; we have sketched the proof of this claim in the notes of last lecture.

Claim 13.3. *For any standard normal random vector $x \in \mathbb{R}^n$,*

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

We sketched the proof of this claim in the notes of the last lecture. Just note that $\mathbb{E}\left[\|x\|^2\right] = n$; so the above theorem simply follows by the concentration of sum of normal random variables.

Next, we prove the third claim.

Claim 13.4. *For any vector $x \in \mathbb{R}^n$, if $y = M^k x$ and $\epsilon > 0$, then*

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

where M is PSD, λ_1 is the largest eigenvalue of M, and v_1 is the corresponding eigenvector.

Putting the above three claims together we can prove Theorem 13.1 As the form of the third claim suggests, we just need to use the preceding two claims to lower bound $\frac{\|x\|^2}{\langle x, v_1\rangle^2}$. Namely, it suffices to show that $\frac{\|x\|^2}{\langle x, v_1\rangle^2} \leq O(n)$ with constant probability. First, of all by Claim 13.2, $|\langle x, v_1\rangle| \geq 1/2$ with constant probability and by Claim 13.3, $\|x\|^2 \leq 2n$ with a very high probability probability. Therefore, by union bound, we have

$$\frac{\|x\|^2}{\langle x, v_1\rangle^2} \leq 4n$$

with a constant probability

Substituting this bound into the statement of Claim 13.4 gives us the theorem statement; with constant probability we have,

$$\frac{y^T M y}{y^T y} \geq \frac{(1-\epsilon)\lambda_1}{1 + \frac{\|x\|^2}{\langle x, v_1\rangle^2}(1-\epsilon)^{2k}} \geq \frac{(1-\epsilon)\lambda_1}{1 + 4n(1-\epsilon)^{2k}} \geq \frac{(1-\epsilon)\lambda_1}{1 + 10n(1-\epsilon)^{2k}}.$$

This will complete the proof of Theorem 13.1.

It remains to prove Claim 13.4.

Proof of Claim 13.4. 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

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

as desired □

13.3 Application of Power Method in Spectral Partitioning

For the spectral partitioning algorithm, we require a means of quickly computing an (approximate) second-smallest eigenvector of the normalized Laplacian matrix, i.e., ideally we would like to find a vector x such

that

$$\frac{x^T \tilde{L} x}{x^T x} \leq (1+\epsilon)\lambda_2,$$

where λ_2 is the 2nd smallest eigenvalue of $\tilde{L}$. Noting Theorem 13.1, power method seems to be the method of choice as it gives multiplicative approximation in almost linear time. Note that since we are looking for the (second) smallest eigenvalue we need to apply the power method to a modified version of the normalized Laplacian. Ideally, we would like to modify the matrix such that the smallest eigenvalues become the largest ones and then apply power method.

Recall that we defined the normalized Laplacian matrix of a graph as $\tilde{L} = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$. It is not hard to see that the largest eigenvalue of $\tilde{L}$ is at most 2. So, we now consider $M = 2I - \tilde{L}$. This matrix is also a PSD matrix and the order of the eigenvalues is the opposite of the eigenvalues of $\tilde{L}$, i.e., for every eigenvalue λ_i of $\tilde{L}$, $2 - \lambda_i$ is an eigenvalue of M. So, in particular, the second largest eigenvalue of M is $2 - \lambda$; and we are trying the estimate its eigenvector. Now, all we need to do is to use the power method.

Firstly, recall that we showed $v_1 = D^{1/2}\mathbf{1}$ is the smallest eigenvector of $\tilde{L}$; so it is also the largest eigenvector of M. We can find the second-largest eigenvector of M by finding the vector with largest Rayleigh quotient among all vectors orthogonal to v_1. First, we let x to be a random Gaussian vector, and then we make it orthogonal to v_1, and we run the power method on the resulting vector. See notes of the last lecture for details.

By Theorem 13.1 we find a vector y such that

$$\frac{y^T(2I - \tilde{L})y}{y^T y} \geq (1-\epsilon)(2-\lambda_2)$$

in time $O(|E| \log n/\epsilon)$. From the above inequality we can conclude that $y^T L y \leq y^T y(\lambda_2 + 2\epsilon)$, i.e., we get a 2ϵ additive approximation to the second eigenvector of $\tilde{L}$. This bound is not ideal for our purposes, since it is *additive* error. For example, if G is a cycle, then $\lambda_2 = O(1/n^2)$, so the Rayleigh quotient can very large unless $\epsilon = O(1/n^2)$. But, in that case the power method takes cubic time $O(|E|n^2)$.

An alternative approach that would yield multiplicative approximation to the second eigenvalue/eigenvector of $\tilde{L}$ is use the inverse of $\tilde{L}$, $\tilde{L}^{-1}$. $\tilde{L}^{-1}$ which is also PSD and has eigenvectors in reverse order. For this matrix any multiplicative approximation of eigenvalues of $\tilde{L}^{-1}$ will also give a multiplicative approximation of eigenvalues of $\tilde{L}$. Note that the smallest eigenvalue of $\tilde{L}$ is 0, so $\tilde{L}^{-1}$ is not well-defined. To avoid that we will consider the pseudoinverse of $\tilde{L}$, defined and denoted as follows:

$$\tilde{L}^{+} := \sum_{i:\lambda_i \neq 0} \frac{1}{\lambda_i} v_i v_i^T$$

where the v_i are the eigenvectors corresponding to the eigenvalue λ_i of $\tilde{L}$.

Now, all we need to do is to run the power method on $\tilde{L}^{+}$. But computing $\tilde{L}^{+}$ is as hard as computing the SVD of $\tilde{L}$. The idea is that computing $\tilde{L}^{+}x$, for a given vector x, is equivalent to solving the system of linear equations $\tilde{L}y = x$. There are very efficient methods for solving systems of linear equations for a Laplacian matrix in almost linear time (see the course website for several pointers). Using that we can give a multiplicative $1+\epsilon$ approximation to the second smallest eigenvalue of $\tilde{L}$ by solving $O(\log n/\epsilon)$ systems of linear equations. So, the algorithm will run in almost linear time. Putting this together with the spectral partitioning algorithm this gives a near linear time algorithm to approximate $\phi(G)$.

13.4 Random Walks

In this section, we will introduce random walks on graphs, and we see that many of the theorems that we proved for the normalized Laplacian matrix naturally translates to the properties of the random walk matrix. Here, we will discuss random walks on undirected graphs, but much of the ideas that we discuss also translates to random walks in general directed graphs.

First, we will define the first-order Markov property. For a stochastic process, we say it has the first order Markov property if the process's state at time $t+1$, X_{t+1}, depends only at its state in the previous instant, time t. That is,

$$\mathbb{P}\left[X_{t+1}|X_t,\ldots,X_1\right] = \mathbb{P}\left[X_{t+1}|X_t\right].$$

A random walk on a graph is one such Markov process. In this case, the state at time t simply refers to the vertex the walker has reached and the probability of transiting from vertex i at time t to j at time $t+1$ is the weight of the edge from i to j, $w_{i,j}$, divided by the total weight of the outgoing edges of i, $d_w(i)$, i.e.,

$$\mathbb{P}\left[X_{t+1}=j|X_t=i\right] = \frac{w_{i,j}}{d_w(i)}.$$

We can express the transition probabilities as a matrix. We define the $n\times n$ transition probability matrix P as follows: For all valid indices i,j,

$$P_{i,j} = \begin{cases} \frac{w(i,j)}{d_w(i)} & \text{if i, j are neighbors} \\ 0 & \text{otherwise} \end{cases}.$$

where $d_w(i) = \sum_k w_{i,k}$ is the total weight of edges incident to i.

Note that the above definition means that $P = D^{-1}A$, where A is the adjacency matrix of a graph and D is the degree matrix. This has interesting implications for the eigenvalues of P, which we will now briefly explore. P may or may not be symmetric. In particular, if the graph is regular P is symmetric and otherwise it is not. However, it can be seen that P always have real eigenvalues. Its left and right eigenvectors are not necessarily the same, but its eigenvalues are equal to the eigenvalues of the normalized adjacency matrix $D^{-1/2}AD^{-1/2}$. Recall that the left eigenvector is a vector v such that $v^TM = \lambda v^T$; in a symmetric matrix, each left eigenvector will be the transpose of a right eigenvector, but they do not need to correspond in this manner for P.

Claim 13.5. *The eigenvalues of P are the same as the eigenvalues of the normalized adjacency matrix, $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$. Since $\tilde{L} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, each eigenvalue λ_i of $\tilde{L}$ corresponds to an eigenvalue $1-\lambda_i$ of P.*

Proof. Suppose we had a right eigenvector v of P, so

$$D^{-1}Av = \lambda v.$$

Let $x = D^{\frac{1}{2}}v$, so $v = D^{-\frac{1}{2}}x$.

Now we will substitute this last identity into our equation above to yield

$$D^{-1}AD^{-\frac{1}{2}}x = \lambda D^{-\frac{1}{2}}x.$$

If we multiply both sides of the equality from the left by $D^{\frac{1}{2}}$, we obtain

$$D^{-\frac{1}{2}}AD^{-\frac{1}{2}}x = \lambda x.$$

thus establishing that for each right eigenvector v of P with eigenvalue λ, $D^{\frac{1}{2}}v$ is an eigenvector of $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ with the same eigenvalue. □

Reasoning along similar lines to the above proof, we may conclude that if x is an eigenvector of $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, then $D^{\frac{1}{2}}x$ is a left eigenvector of P and, as we saw in the proof, $D^{-\frac{1}{2}}x$ is a right eigenvector of P.

What are the ramifications of this relationship between the eigenvalues of P and $\tilde{L}$? In the next lecture, we will use this information to relate the second-smallest eigenvector of $\tilde{L}$ to the *mixing time* of the random walk. The mixing time is the time at which the probability distribution of possible locations of the random walker will converge to some fixed distribution no matter the starting point of the random walk, called the *stationary distribution* (specifically, we define the mixing time to be the maximum number of steps to reach the stationary distribution, quantifying over the time to reach the stationary distribution from each possible starting vertex, but this will all be defined in greater detail in the following lecture).