CSE 521: Design and Analysis of Algorithms I **Fall 2019**

Lecture 15: Power Method, Spectral Sparsification

Lecturer: Shayan Oveis Gharan *11/21/18*

15.1 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 15.1. *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 15.2 (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 15.3 (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 15.4 (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 15.5. *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 15.6. *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 15.7. *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 15.1.

Claim 15.8. *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 15.5 and Claim 15.6, $\frac{\|x\|^2}{\langle x, v_1\rangle^2} \leq 4n$ with a constant probability.

Note that we needed randomness in the statement of Theorem 15.1 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 □

15.2 Spectral Sparsifiers

For two symmetric matrix $A, B \in \mathbb{R}^{n \times n}$ we write

$$A \preceq B$$

iff $B - A \succeq 0$, i.e., $B - A$ is a PSD matrix. In other words, $A \preceq B$ iff for any vector $x \in \mathbb{R}^n$,

$$x^T A x \leq x^T B x$$

Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of A and $\tilde{\lambda}_1 \leq \cdots \leq \tilde{\lambda}_n$ be the eigenvalues of B. It follows that if $A \preceq B$, then for all i, $\lambda_i \leq \tilde{\lambda}_i$.

Definition 15.9. *Given a graph $G = (V, E)$ and $\epsilon > 0$, we say a (weighted) graph $H = (V, E')$ is a $1 \pm \epsilon$-spectral sparisifier of G if*

$$(1-\epsilon) L_G \preceq L_H \preceq (1+\epsilon) L_G.$$

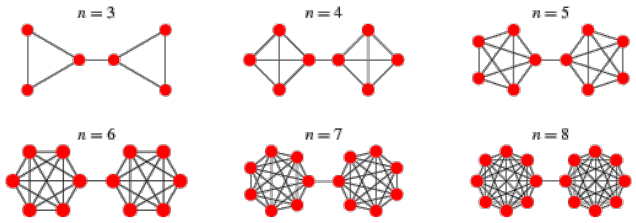

Figure 15.1: Barbell Graph

Ideally, we want H to be a subgraph of G which has much fewer edges than G. An immediate consequence of the above definition is that all eigenvalues of H approximate eigenvalues of H up to multiplicative $1 \pm \epsilon$ error.

It is also not hard to see that if H is a $1 \pm \epsilon$-spectral sparisifer of G then it preserves the size of all cuts of G. In particular, for a set $S \subseteq V$, recall $\mathbf{1}^S$ is the indicator vector of the set S. It follows that for a graph G,

$$\mathbf{1}^S L_G \mathbf{1}^S = \sum_{i \sim j} (\mathbf{1}^S_i - \mathbf{1}^S_j)^2 = \sum_{i \sim j} \mathbb{I}\left[|\{i,j\} \cap S| = 1\right] = 2|E(S, \overline{S})|$$

So, if H is a $1 \pm \epsilon$-spectral sparsifier of G we have

$$(1-\epsilon)\mathbf{1}^S L_G \mathbf{1}^S \leq \mathbf{1}^S L_H \mathbf{1}^S \leq (1+\epsilon)\mathbf{1}^S L_G \mathbf{1}^S,$$

so the (weighted) size of every cut in H is within $1 \pm \epsilon$ multiplicative factor of the same cut in G.

Theorem 15.10 (Speilman-Srivastava)**.** *For every graph $G = (V, E)$ and $\epsilon > 0$, there is a weighted graph H that is a subgraph of G such that H is a $1 \pm \epsilon$-spectral sparsifier of G and that H has at most $O(n \log n/\epsilon^2)$ many edges.*

The first idea that come to mind is to construct an unbiased estimator: Let X be a random matrix defined as follows: For every edge $e \in E$, $X = L_e/p_e$ with probability p_e, Then, observe that

$$\mathbb{E}[X] = \sum_e p_e \frac{L_e}{p_e} = \sum_e L_e = L_G.$$

So, X is an unbiased estimator. And, the main question is how to choose the probabilities such that concentration bounds can kick in and imply $X \approx \mathbb{E}[X]$.

Let us start with a simple case of a complete graph. If G is a complete graph, we can simply let $p_e = 1/\binom{n}{2}$ for all edges. It then follows that $O(n \log n/\epsilon^2)$ many samples are enough to approximate the complete graph. However, it turns out that a uniform distribution does not necessarily work out in a general graph. For example, if G is a *Barbell graph*, i.e., union of two K_n connected by an edge (see Figure 15.1), then, if we want to down-size G to $O(n \log n)$ edges we need to let $p_e = O(\log n)/n$ for all edges, but then the single edge connecting the two complete graphs won't be chosen with high probability. So, H is disconnected with high probability and it cannot be a spectral sparsifier of G for any $\epsilon < 1$. In the rest of this section we will see how to choose the edge probabilities p_e.

15.2.1 Reduction to Isotropic Case

First, it turns out that we can reduce the graph sparsification problem to a linear algebraic problem. First, let us recall the generalized eigenvalue problem. In the generalized eigenvalue problem we are given a symmetric

matrix A and a PSD matrix B and we want to find

$$\max_x \frac{x^T A x}{x^T B x}$$

In the special case that B is the identity matrix, the solution of the above problem is exactly the largest eigenvector of A. We can solve the above problem by reducing it to an eigenvalue problem.

$$\max_x \frac{x^T A x}{x^T B x} = \max_x \frac{x^T B^{1/2} B^{-1/2} A B^{-1/2} B^{1/2} x}{x^T B^{1/2} B^{1/2} x} = \max_{x: y = B^{1/2} x} \frac{y^T B^{-1/2} A B^{-1/2} y}{y^T y} = \max_y \frac{y^T B^{-1/2} A B^{-1/2} y}{y^T y}$$

So, to find the solution to the generalized eigenvalue problem it is enough to find the largest eigenvector y of the matrix $B^{-1/2} A B^{-1/2}$ and then let $x = B^{-1/2} y$. Note that, here we are using the fact that B is PSD; otherwise $B^{-1/2}$ is not well defined.

Now, let us go back to the spectral sparsifier problem. Suppose H is a $1 \pm \epsilon$-spectral sparsifier of G. It follows that for all $x \in \mathbb{R}^n$.

$$1 - \epsilon \leq \frac{x^T L_H x}{x^T L_G x} \leq 1 + \epsilon$$

By a similar analogy, it follows that for all y,

$$1 - \epsilon \leq \frac{y^T L_G^{-1/2} L_H L_G^{-1/2} y}{y^T y} \leq 1 + \epsilon$$

So, the above inequality implies that the matrix $L_G^{-1/2} L_H L_G^{-1/2}$ is approximately equal to the identity matrix.

Remark 15.11. *There is a technical problem here: since L_G has a zero eigenvalue the inverse of L_G is not well-defined. In the above calculation, we take the inverse with respect to positive eigenvalues of G; in particular if $L_G = \sum_i \lambda_i v_i v_i^T$, we let $L_G^{-1/2} = \sum_{i:\lambda_i > 0} \frac{1}{\sqrt{\lambda_i}} v_i v_i^T$. We ignore this fact in the rest of our calculations for the simplicity of the argument.*

Now, we reformulate the spectral sparsification problem as follows:

Theorem 15.12. *Given $n \times n$ PSD matrices, $E_1, \ldots, E_m$ such that*

$$\sum_{i=1}^m E_i = I,$$

For any $\epsilon > 0$, there is a subset S of them of size $O(n \log n / \epsilon^2)$ and a set of weights w_i for each $i \in S$ such that

$$(1 - \epsilon) I \preceq \sum_{i \in S} w_i E_i \preceq (1 + \epsilon) I$$

Let us discuss how we can reduce the sparsification problem to the above theorem. Say our graph G has m edges. For edge e_i define

$$E_i = L_G^{-1/2} L_{e_i} L_G^{-1/2}.$$

First, observe that each E_i is a PSD matrix, and furthermore,

$$\sum_{i=1}^m E_i = \sum_{i=1}^m L_G^{-1/2} L_{e_i} L_G^{-1/2} = L_G^{-1/2} \left(\sum_{i=1}^m L_{e_i} \right) L_G^{-1/2} = L_G^{-1/2} L_G L_G^{-1/2} = I.$$

So, roughly speaking by multiplying the Laplacians of the edges of G by $L_G^{-1/2}$ on both sides we are normalizing the space such that every direction look the same. We are reducing the graph spectral sparsification problem to a linear algebraic problem of finding a sparsifier of the sum of PSD matrices that add up to the identity matrix.

15.2.2 Finding the Spectral Sparsifier

Now, as before, let

$$X = \frac{E_i}{p_i}$$

with probability p_i. Similar to before, $\mathbb{E}[X] = I$. So, X is an unbiased estimator. To prove the concentration we used the following generalization of the Chernoff bound which is known as matrix Chernoff bound

Theorem 15.13. *Let X be a random $n \times n$ PSD matrix. Suppose that $X \preceq \alpha \mathbb{E}[X]$ with probability 1. Let $X_1, \ldots, X_k$ be independent copies of X, then for any $\epsilon > 0$,*

$$\mathbb{P}\left[(1-\epsilon)\mathbb{E}[X] \preceq \frac{1}{k}(X_1 + \cdots + X_k) \preceq (1+\epsilon)\right] \geq 1 - 2ne^{-\epsilon^2 k/4\alpha}.$$

So, this says that to prove Theorem 15.12 it is enough to choose $k = O(\alpha \log n/\epsilon^2)$ many copies of X. Finally, to finish the proof we need to choose the probabilities p_i such that $\alpha \leq O(n)$.

First, suppose we let p_i be uniform, i.e., $p_i = 1/m$ for all i. Then, we need to choose α such that for all i,

$$\frac{E_i}{1/m} \preceq \alpha I.$$

But it turns out that in the worst case we have to let $\alpha = m$.

The idea is to let $p_i \propto \operatorname{Tr}(E_i)$. Let us first find the normalizing constant: Suppose $p_i = \beta \operatorname{Tr}(E_i)$. Then,

$$\sum_i p_i = \beta \sum_i \operatorname{Tr}(E_i) = \beta \operatorname{Tr}\left(\sum_i E_i\right) = \beta n$$

So, we should let $\beta = 1/n$. It follows that $p_i = \beta \operatorname{Tr}(E_i) = \operatorname{Tr}(E_i)/n$.

Now, we claim that for all i,

$$\frac{E_i}{\operatorname{Tr}(E_i)/n} \preceq \alpha I$$

for $\alpha = n$. This will complete the proof of Theorem 15.12. To show the above it is enough to show

$$\frac{E_i}{\operatorname{Tr}(E_i)} \preceq I$$

To show this we only use the fact that all eigenvalues of E_i are in the range $[0, 1]$ (this is true because E_i is PSD, and $\sum_j E_j = I$). So, it remains to prove the above inequality. Say $E_i = \sum_j \lambda_j v_j v_j^T$. For any arbitrary vector $x \in \mathbb{R}^n$,

$$x^T \frac{E_i}{\operatorname{Tr}(E_i)} x = \frac{\sum_j \lambda_j \langle x, v_j\rangle^2}{\sum_j \lambda_j} \leq \max_j \langle x, v_j\rangle^2 \leq \|x\|^2 = x^T I x.$$

15.2.3 Back to Spectral Sparsifiation

In the previous section we saw that we should choose each E_i with probability $\operatorname{Tr}(E_i)/n$. Translating this back to the setting of graph sparsification; recall that for edge e_i, $E_i = L_G^{-1/2} L_{e_i} L_G^{-1/2}$. So, we should sample every edge e of G with probability

$$p_e = \frac{\operatorname{Tr}(L_G^{-1/2} L_e L_G^{-1/2})}{n}$$

The quantity

$$\mathrm{Tr}(L_G^{-1/2} L_e L_G^{-1/2}) = b_e^T L_G^{-1} b_e$$

is called the *effective resistance* of the edge e; here for an edge $e = \{u, v\}$, $b_e = \mathbf{1}_u - \mathbf{1}_v$ is the vector which is $+1$ at one endpoint of e and -1 at the other endpoint and 0 everywhere else. It is very well understood and there are fast algorithms to estimate it; one can also compute the inverse of the Laplacian and compute the effective resistance of all edges immediately.

The following simple algorithm can be used to construct a $1 \pm \epsilon$-spectral sparsifier of G:

1. For $i = 1$ to $O(n \log n/\epsilon^2)$
2. Sample each edge e of G with probability $p_e = \mathrm{Tr}(L_G^{-1/2} L_e L_G^{-1/2})/n$. If the edge e is sampled weight it by $1/p_e$.