Approximating the Largest Eigenvalue via Powering and the Lanczos method

Jeffrey Lai

April 18, 2015

Notation: $\lambda_1(A)$ = largest eigenvalue of A.

Motivation: The largest eigenvalue tells us the spectrum of a matrix and thus can be somewhat useful. More crucially for the adjacency matrix of a graph, we know exactly that the all one's vector is the largest eigenvector, and thus by working orthogonal to this vector we can approximate the second largest eigenvalue of the graph, which tells us how well connected the graph is and how nodes are clustered.

1 Power Method

Theorem 1.1. *Given a symmetric matrix $A \in R^{n\times n}$ an error parameter $\delta > 0$ and $k > \frac{1}{2\delta}\log(9n/4)$, the following holds with probability at least 1/2 over unit vectors v chosen uniformly at random.*

$$\frac{\left\|A^{k+1}v\right\|}{\|A^k v\|} \geq (1-\varepsilon)|\lambda_1(A)|$$

Proof. Let $u_1...u_n$ be the orthonormal eigenbasis of A. We can write v in this eigenbasis as $v = \sum \alpha_i u_i$. Fact: with probability at least 1/2 $|\alpha_1| \geq \frac{2}{3\sqrt{n}}$.

$$\begin{aligned}\|A^k v\|^2 &= \left\|\left(\sum_i \lambda_i u_i u_i^\top\right)\left(\sum_i \alpha_i u_i\right)\right\|^2 \\ &= \sum_i \alpha_i^2 \lambda_i^{2k}\end{aligned}$$

We now use Hölder's inequality. Recall this is $|w \cdot x| \leq \|w\|_p \|x\|_q$ for $\frac{1}{p} + \frac{1}{q} = 1$. Choosing the values

$$w_i = \alpha_i^{2k/k+1}\lambda_i^{2k}$$
$$x_i = \alpha_i^{2/k+1}$$
$$p = k+1, q = \frac{k+1}{k}$$

We have then that

$$\begin{aligned}\left\|A^k v\right\|^2 &= \sum_i \alpha_i^2 \lambda_i^{2k} \\ &\leq \left(\sum_i \alpha_i^2 \lambda_i^{2k+2}\right)^{k/k+1} \left(\sum_i \alpha_i^2\right)^{1/k+1} \\ &= \left(\sum_i \alpha_i^2 \lambda_i^{2k+2}\right)^{k/k+1} \\ &= \left\|A^{k+1} v\right\|^{2k/k+1}\end{aligned}$$

Note that $\left\|A^{k+1} v\right\|^{2/k+1} \geq \alpha_1^{2/k+1} \lambda_1^2$ as the right hand side is just a single term from the left hand side. Thus the inequality holds when we multiply by the quotient of these values (as it is greater than 1).

$$\left\|A^k v\right\|^2 \leq \left\|A^{k+1} v\right\|^{2k/k+1} \cdot \frac{\left\|A^{k+1} v\right\|^{2/k+1}}{\alpha_1^{2/k+1} \lambda_1^2} = \frac{\left\|A^{k+1} v\right\|^2}{\alpha_1^{2/k+1} \lambda_1^2}$$

Rearrange to

$$\frac{\left\|A^{k+1} v\right\|}{\|A^k v\|} \geq \alpha_1^{2/k+1} \lambda_1^2$$

Substitute in $k+1 \geq \frac{1}{2\delta} \log(9n/4)$ to show that $|\alpha_1|^{1/k+1} \geq e^{-\delta} \geq 1-\delta$. [1]

2 Lanczos method

Theorem 2.1. *Given a symmetric PSD matrix $A \in R^{n \times n}$, and a parameter $\delta > 0$, the Lanczos method after k iterations for $k = O(1/\sqrt{\delta} \cdot \log n/\delta)$, outputs a value $\mu \in [(1-\delta)\lambda_1(A), \lambda_1(A)]$ with constant probability. The running time is $O((t_A + n)k + k^2)$*

This bound was first proven by [3, Kucźynski, Woźniakowski 1992]. The proof below is from [2, Sachdeva, Vishnoi 2014] and uses Chebyshev polynomials like in Conjugate Gradient.

Summary: We compute an orthonormal basis for our Krylov subspace, and use that to calculate T. We can then calculate the largest eigenvalue of T and this turns out to be close to $\lambda_1 = \lambda_1(A)$ for random v and appropriate k.

Fact: For a symmetric matrix A it's largest eigenvalue is characterized as follows

$$\lambda_1(A) = \max_{w \neq 0} \frac{w^\top A w}{w^\top w}$$

Similar to conjugate gradient we will work in the Krylov subspace $\mathcal{K} = Span\{v, Av, ...vA^k v\}$. The intuition is that within this subspace we can work with A restricted to this subspace, and the largest eigenvalue of that matrix is not far from λ_1. By working only in the subspace we can do this relatively quickly.

Now unlike conjugate gradient we will actually compute an orthonormal basis for $\mathcal{K}$. We will later show this can be done in $O((t_A+n)k)$ operations. For now say we have such a basis $\{v_0...v_k\}$. Let V be the $n \times k+1$ matrix with column $i = v_i$. Note $VV^\top$ is the orthogonal projection onto $\mathcal{K}$. Let $T \stackrel{\text{def}}{=} V^\top AV$. This is the operator A restricted to $\|$. Note that for all $w \in \mathcal{K}, w = VV^\top w$. Thus we have

$$w^\top Aw = w^\top VV^\top AVV^\top w = (w^\top V)T(V^\top w)$$

Thus the Rayleigh quotient of w with respect to A is the same as the Rayleigh quotient of $V^\top w$ with respect to T. Noting the as w ranges over $\mathcal{K}$, $V^\top w$ ranges over $\mathbb{R}^{k+1}$ we have $\lambda_1(T) \leq \lambda_1$.

Note: the analysis below is purely for the showing that our algorithm produces a good approximation. After computing V,T, and $\lambda_1(T)$ the algorithm just returns $\lambda_1(T)$. By our Rayleigh quotient definition we have

$$\lambda_1(T) = \max_{z \in \mathbb{R}^{k+1}} \frac{z^\top Tz}{z^\top z} = \max_{w \in \mathcal{K}} \frac{w^\top VTV^\top z}{z^\top z} = \max_{w \in \mathcal{K}} \frac{w^\top Aw}{w^\top w}$$

Because $w \in \mathcal{K} = Span\{v, ..., A^k v\}$ we have $w = p(A)v$ for some $p \in \sum_k$.

$$\lambda_1(T) = \max_{p \in \sum_k} \frac{v^\top p(A)Ap(A)v}{v^\top p(A)^2 v}$$

Writing $v = \sum \alpha_i u_i$, u_i eigenvectors of A (and p(A) etc.)

$$\lambda_1(T) = \max_{p \in \sum_k} \frac{\sum_i \lambda_i p(\lambda_i)^2 \alpha_i^2}{\sum_i p(\lambda_i)^2 \alpha_i^2}$$

Now want to bound how well $\lambda_1(T)$ approximates λ_1. We have

$$\begin{aligned} 1 - \frac{\lambda_1(T)}{\lambda_1} &= 1 - \max_{p \in \sum_k} \frac{\sum_i \lambda_i/\lambda_1 p(\lambda_i)^2 \alpha_i^2}{\sum_i p(\lambda_i)^2 \alpha_i^2} \\ &\leq 1 - \frac{\sum_i \lambda_i/\lambda_1 p(\lambda_i)^2 \alpha_i^2}{\sum_i p(\lambda_i)^2 \alpha_i^2} \\ &= \frac{\sum_i (1 - \lambda_i/\lambda_1) p(\lambda_i)^2 \alpha_i^2}{\sum_i p(\lambda_i)^2 \alpha_i^2} \end{aligned}$$

Like in the power method we have $\alpha_1^2 \geq 1/9n$ with probability 1/2. Split sum into $\lambda \geq (1-\delta)\lambda_1$ and less.

$$\begin{aligned} \frac{\sum_i (1 - \lambda_i/\lambda_1) p(\lambda_i)^2 \alpha_i^2}{\sum_i p(\lambda_i)^2 \alpha_i^2} &\leq \delta + \frac{\sum_{\lambda_i < (1-\delta)\lambda_1} (1 - \lambda_i/\lambda_1) p(\lambda_i)^2 \alpha_i^2}{\sum_i p(\lambda_i)^2 \alpha_i^2} \\ &\leq \delta + 9n \sup_{\lambda \in [0,(1-\delta)\lambda_1]} \frac{p(\lambda)^2}{p(\lambda_1)^2} \end{aligned}$$

Note that for the choice of polynomial $p(\lambda) = (\lambda/\lambda_1)^s$ for $s = \lceil 1/2\delta \cdot \log 9n/\delta \rceil$ we get our relative error is

$$\delta + 9n \sup_{\lambda \in [0,(1-\delta)\lambda_1]} \left(\frac{\lambda}{\lambda_1}\right)^{2s} = \delta + 9n \cdot (1-\delta)^{2s} \leq 2\delta$$

This gives us the same guarantee as the power method of $O(1/\delta \log n/\delta)$. We will show with Chebyshev polynomials (much like in Conjugate Gradient) that Lanczos actually does better.

2.1 Chebyshev polynomials

Recall that: $D_s = \sum_{i=1}^{s} Y_i$, $Y_i = \pm 1$.

$$p_{s,d}(x) \stackrel{\text{def}}{=} \mathbf{E}_{Y_1,\ldots,Y_s}\left[T_{D_s}(x) \cdot \mathbb{1}_{|D|\leq d}\right]$$

Fact:

$$\sup_{x\in[-1,1]} |p_{s,d}(x) - x^s| \leq 2e^{-d^2/2s}$$

Thus for $p(\lambda) = p_{s,d}(\lambda/\lambda_1)$, $s = \lceil 1/2\delta \cdot \log 9n/\delta \rceil$ and $d = \lceil \sqrt{2s \log 2n/\delta} \rceil$ we have

$$|p(\lambda) - (\lambda/\lambda_1)^s| \leq \delta/n$$

Thus we get

$$\sup p(\lambda)^2 \leq \sup(\lambda/\lambda_1)^{2s} + \delta/n = O(\delta/n)$$

Plugging in $x = \lambda_1$ we have $p(\lambda_1) \geq 1 - \delta/n$. Plugging back into our above bound we get

$$\frac{\lambda_1 - \lambda_1(T)}{\lambda_1} = \delta + 9n\frac{O(\delta/n)}{1-\delta/n} = O(\delta)$$

thus completing the proof of our bound. □

2.2 Running Time

Note that for all i we have $Span\{v, Av, ..., A^i v\} = Span\{v_0, v_1, ..., v_i\}$, thus for any $v_j \in Span\{v, Av, ..., A^i v\}$ we have $Av_j \in Span\{Av, A^2 v, ..., A^{i+1} v\} = Span\{v_0, v_1, ..., v_i, v_{i+1}\}$. Thus for any $j > i+1$, v_j is not in the span of these vectors, hence we have

$$v_i^\top A v_j = 0$$

Crucially now we use the fact that A is symmetric to take the transpose to get

$$v_j^\top A v_i = 0$$

This gives us then for $j > i+1$ v_j is orthogonal to Av_i, thus when we are calculating our orthonormal basis, we need only orthogonalize with respect to at most two vectors, as the vector is already orthogonal to the rest of the vectors. Furthermore this also implies that $T = V^\top AV$ is tridiagonal. Due to the sparsity and structure of T we can compute the largest eigenvalue in $O(k^2)$ time. [4]

References

[1] Nisheeth Vishnoi, "Lx = b: Laplacian Solvers and Their Algorithmic Applications." *Foundations and Trends in Theoretical Computer Science* 8 (2013): 1-141. Web. <http://research.microsoft.com/en-us/um/people/nvishno/site/Lxb-Web.pdf>

[2] J. Kuczyński and H. Woźniakowski. Estimating the largest eigenvalues by the power and Lanczos algorithms with a random start. *SIAM J. Matrix Anal. Appl.*, 13(4):1094-1122, October 1992.

[3] Sachdeva, Sushant, and Nisheeth Vishnoi. "Faster Algorithms via Approximation Theory." *Foundations and Trends in Theoretical Computer Science* 9.2 (2014): 125-210. Web. <http://www.cs.yale.edu/homes/sachdeva/pubs/fast-algos-via-approx-theory.pdf>.

[4] V. Y. Pan and Z. Q. Chen. The complexity of the matrix eigenproblem. In *STOC'99*, pages 507-516, 1999.