Jane Kim

Numerical Linear Algebra

Final Project

25 April 2019

The objective of this project is to solve the 1D constant-coefficient Sturm-Liouville (SL) spectrum problem,

$$\begin{aligned} -pu'' + qu &= \lambda u, \quad 0 < x < \pi \\ u(0) &= 0, \quad u'(\pi) = 0, \end{aligned}$$

using two different discretization schemes and three different eigenvalue solvers. The analytical eigenvalues are given by

$$\lambda_l = p\left(l - \frac{1}{2}\right)^2 + q, \quad l = 1, 2, \cdots,$$

so that we have an infinite sequence of real eigenvalues with $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_l \leq \cdots \to \infty$.

First, we partition the domain $[0, \pi]$ into $N$ subintervals of equal width $h = \frac{\pi}{N}$. Let $x_i = ih$ denote the grid points and $U_i = U(x_i)$ denote the numerical estimation of the eigenfunction $u(x)$ at the point $x_i$. Then we can obtain two different ways of approximating the SL problem using the finite difference method. The discretization of Scheme A can be written as

$$A\vec{U} = \Lambda_A^h \vec{U}. \tag{1}$$

where

$$A = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -2 & 2 \end{bmatrix}, \quad \text{and } \vec{U} = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-1} \\ U_N \end{bmatrix},$$

Here, $\Lambda_A^h = \frac{h^2}{p}(\lambda^h - q)$ is the eigenvalue of the matrix $A$ and $\lambda^h$ is the eigenvalue of the Sturm-Liouville problem. Then after solving for $\Lambda_A^h$ numerically, the original eigenvalue of the SL problem can be obtained with $\lambda^h = \frac{p}{h^2}\Lambda_A^h + q$. The solutions for the SL eigenvalues in this scheme are given by

$$\hat{\lambda}_l^h = pk_h(l) + q, \quad k_h(l) = \frac{2}{h^2}\left(1 - \cos\left(\left(l - \frac{1}{2}\right)h\right)\right),$$

for $l = 1, 2, \cdots$. To show that $\hat{\lambda}_l^h$ converges to $\lambda_l$, we can use the Taylor expansion of $\cos(x)$ at $x = 0$ to

expand $k_h(l)$ as

$$k_h(l) \approx \left(l - \frac{1}{2}\right)^2 - \frac{1}{12}\left(l - \frac{1}{2}\right)^4 h^2.$$

Then we have that

$$\hat{\lambda}_l^h \approx p\left(l - \frac{1}{2}\right)^2 - \frac{p}{12}\left(l - \frac{1}{2}\right)^4 h^2 + q \approx \lambda_l - \frac{p}{12}\left(l - \frac{1}{2}\right)^4 h^2,$$
$$|\lambda_l - \hat{\lambda}_l^h| \approx \frac{p}{12}\left(l - \frac{1}{2}\right)^4 h^2 = O(h^2).$$

Therefore, $\hat{\lambda}_l^h$ converges to $\lambda_l$ with $O(h^2)$ accuracy.

The discretization of Scheme $B$ can be written as a general eigenvalue problem

$$A\vec{U} = \Lambda_B^h B\vec{U}, \tag{2}$$

where

$$B = \begin{bmatrix} 4 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 2 & 4 \end{bmatrix}, \quad \Lambda_B^h = \frac{h^2}{6p}(\lambda^h - q),$$

and $A$ and $\vec{U}$ have the same definitions as before. Then the original eigenvalue of the Sturm-Liouville problem is given by $\lambda^h = \frac{6p}{h^2}\Lambda_B^h + q$. We can similarly show that the solutions to this generalized eigenvalue problem,

$$\lambda_l^h = \frac{pk_h(l) + qm_h(l)}{m_h(l)}, \quad m_h(l) = \frac{1}{3}\left(2 + \cos\left(\left(l - \frac{1}{2}\right)h\right)\right),$$

where $k_h(l)$ is has the same definition as in Scheme A, converges to $\lambda_l$. First we expand

$$m_h(l) \approx 1 - \frac{1}{6}\left(l - \frac{1}{2}\right)^2 h^2.$$

Then we have that

$$pk_h(l) + qm_h(l) \approx \lambda_l m_h(l) + \frac{p}{12}\left(l - \frac{1}{2}\right)^4 h^2,$$
$$\lambda_l^h = \lambda_l + \frac{p}{12}\left(l - \frac{1}{2}\right)^4 \frac{h^2}{1 - \frac{1}{6}(l - \frac{1}{2})^2 h^2} \approx \lambda_l + \frac{p}{12}\left(l - \frac{1}{2}\right)^4 h^2,$$

where we have used the Taylor series expansion $\dfrac{x^2}{1 - \alpha x^2} = x^2 + \alpha x^4 + O(x^6)$ in the last line. Therefore,

both $\lambda_l^h$ and $\hat{\lambda}_l^h$ converge to $\lambda_l$ in the same $O(h^2)$ accuracy:

$$|\lambda_l - \lambda_l^h| \approx \frac{p}{12}\left(l - \frac{1}{2}\right)^4 h^2 = O(h^2).$$

Three different methods were used to compute the smallest eigenvalue of the standard eigenvalue problem (1): the inverse power method, the shifted power method, and the QR iteration with deflation. Since we are interested in computing the smallest eigenvalue, we implemented the inverse power method with no shift. To increase the speed of the program slightly, we chose to solve the linear system in each iteration instead of computing the inverse. In addition, we opted out of storing all of the iterates as suggested in the pseudocode in Algorithm 27.2. The iteration was terminated once the change in the eigenvalue fell below some given tolerance.

The standard power method returns the largest eigenvalue, so to implement the shifted power method, we must choose the shift $\mu$ wisely to allow for proper convergence to the smallest eigenvalue. In Scheme $A$, the matrix in question is matrix $A$ which has only 2's along the diagonal. A common choice for $\mu$ is the last diagonal element of the matrix, but this choice does not always guarantee convergence. So we chose $\mu$ as the last diagonal element plus a small perturbation, $\mu = 2.001$, which should ensure that $|\lambda_1 - \mu| > |\lambda_N - \mu|$. The smaller eigenvalues are closer together than the larger ones, so we expect that the convergence rate of this method to be quite poor. So it is important to again opt out of storing the eigenvalues/eigenvectors at each iteration.

The QR iteration was applied to the matrix $A$ until the element $a_{N,N-1}$ fell below the tolerance. This element was then declared to be zero and its neighbor $a_{N,N}$ was stored as an eigenvalue. This process was repeated to the upper-left $(N-1) \times (N-1)$ submatrix, then to the $(N-2) \times (N-2)$ submatrix, and so on, until all of the eigenvalues were computed. Finally, the eigenvalues were sorted to obtain the smallest eigenvalue.

The results of these three methods applied to scheme $A$ are presented in Table 1. In addition, the difference $|\lambda_1 - \hat{\lambda}_1^h|$ was plotted as function of step size $h$ for each of the three methods in Figures 1-3. A linear regression was performed for each method to obtain the order of the convergence. We can see from Table 1 that the inverse power method has a clear advantage over the other two methods. It requires two to five orders of magnitude less iterations than the shifted power method and the QR iteration. And in fact, the number of iterations required decreases as the size of the matrix increases. From Figure 1, the convergence order of the method using a tolerance of $1e-12$ was 1.99729, which is very close to the theoretical value of of 2.

Meanwhile, the shifted power method has some obvious drawbacks. First, the optimal shift $\mu$ is difficult to determine universally, since larger matrices may need a slightly different shift than the smaller ones. Since a fixed shift $\mu = 2.001$ was used for all $N = 16, 32, 64, 128, 256, 512$, it is easy to see that this choice of shift was best suited for the $N = 128$ case. In Figure 2, we can see another problem with the shifted power method. It appears the method fails when the matrix size becomes too large (or when

| method | $N$ | iterations | smallest eigenvalue |
|---|---|---|---|
| inverse power | 16 | 11 | 5.249799 |
| | 32 | 10 | 5.249950 |
| | 64 | 9 | 5.249987 |
| | 128 | 8 | 5.249997 |
| | 256 | 7 | 5.249999 |
| | 512 | 6 | 5.250000 |
| shifted power | 16 | 20746 | 5.249799 |
| | 32 | 18777 | 5.249950 |
| | 64 | 16731 | 5.249987 |
| | 128 | 14386 | 5.249996 |
| | 256 | 28701 | 5.250004 |
| | 512 | 84668 | 5.250326 |
| QR iteration | 16 | 1351 | 5.249799 |
| | 32 | 5136 | 5.249950 |
| | 64 | 19412 | 5.249987 |
| | 128 | 73064 | 5.249997 |
| | 256 | 273861 | 5.249999 |
| | 512 | 1021818 | 5.250326 |

Table 1: The results for scheme A discretization. Here, we have taken $p = 1$ and $q = 5$ for the constant-coefficient Sturm-Liouville problem. For the shifted power method, the shift was fixed at 2.001, the last diagonal element of matrix $A$ plus a small perturbation.

the step size becomes too small), since $N = 256$ and $N = 512$ are clearly off the line of convergence. As a result, taking a linear regression of all six points results in a very poor overall convergence order of 0.25944. However, excluding the outliers results in a convergence order of 1.89211.

Although, the QR iteration with deflation requires the computation of all $N$ eigenvalues to obtain the smallest one, it is still more stable than the shifted power method for larger matrices. This method resulted in a convergence order of 1.99992, the closest convergence order to the theoretical value in scheme A. However, the number of iterations, and hence computation time, seems to grow exponentially with matrix size.

Let $B = QR$ be the QR decomposition of the matrix $B$ in the general eigenvalue problem (2). Then we can transform (2) into a standard eigenvalue problem:

$$A\vec{U} = \Lambda_B^h QR\vec{U} \implies Q^{-1}A(R^{-1}R)\vec{U} = \Lambda_B^h R\vec{U} \implies (Q^{-1}AR^{-1})(R\vec{U}) = \Lambda_B^h(R\vec{U}).$$

This type of transformation can be done using any kind of decomposition of $B$. Note that using the QR decomposition is particularly convenient, because one of the matrices we need to invert is orthogonal so we can just use that $Q^{-1} = Q^T$. Then all there is left to do is solve for the smallest eigenvalue of $Q^T AR^{-1}$ using a method of choice. The code was written in a way that allows one to choose any of the three methods used to solve scheme A. We will show the results using the inverse power method only,

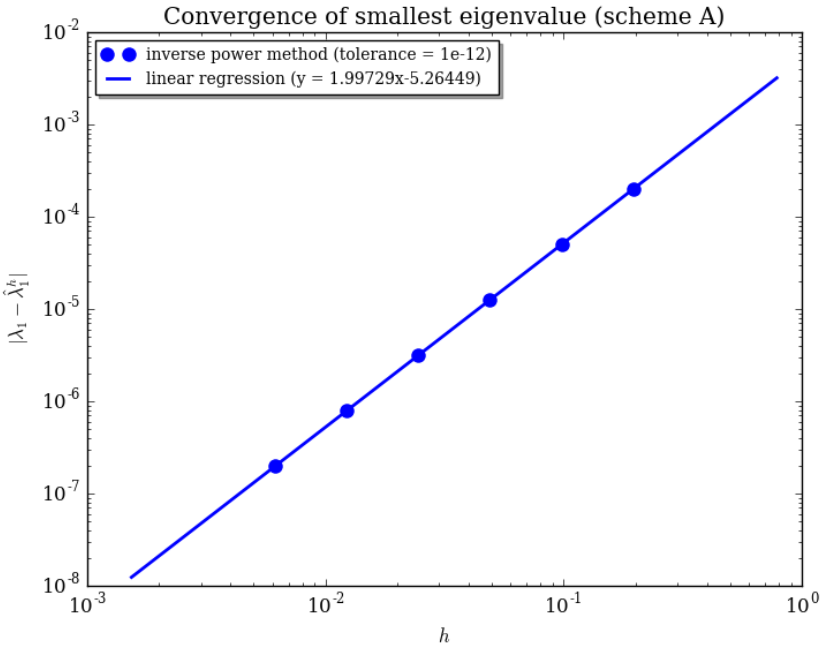


Figure 1: Inverse power method applied to scheme A.

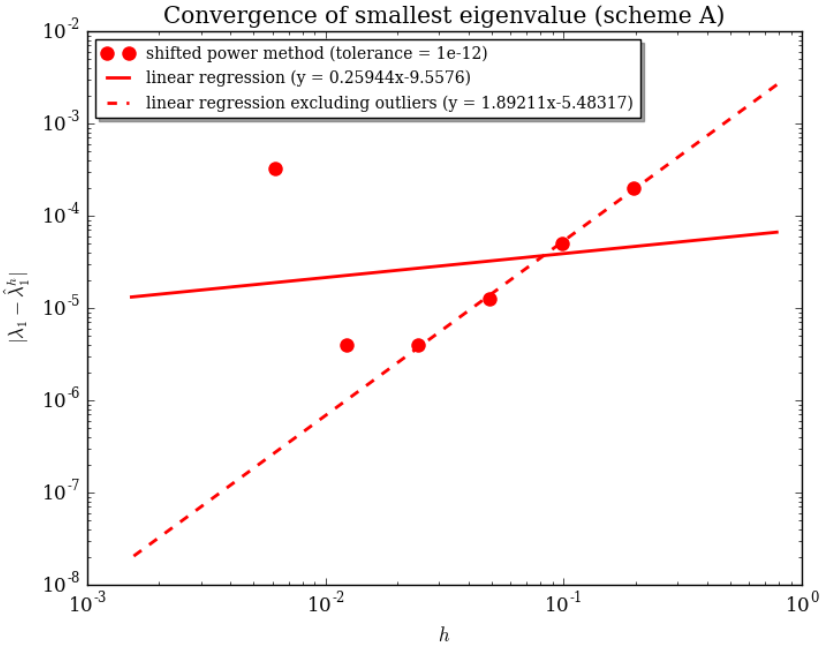


Figure 2: Shifted power method applied to scheme A.

since the results from scheme A suggest that this method is the best compromise between computation time and accuracy. The results are presented in Table 2 and in Figure 4. The convergence order for scheme $B$ was determined to be 2.00344 using the inverse power method.

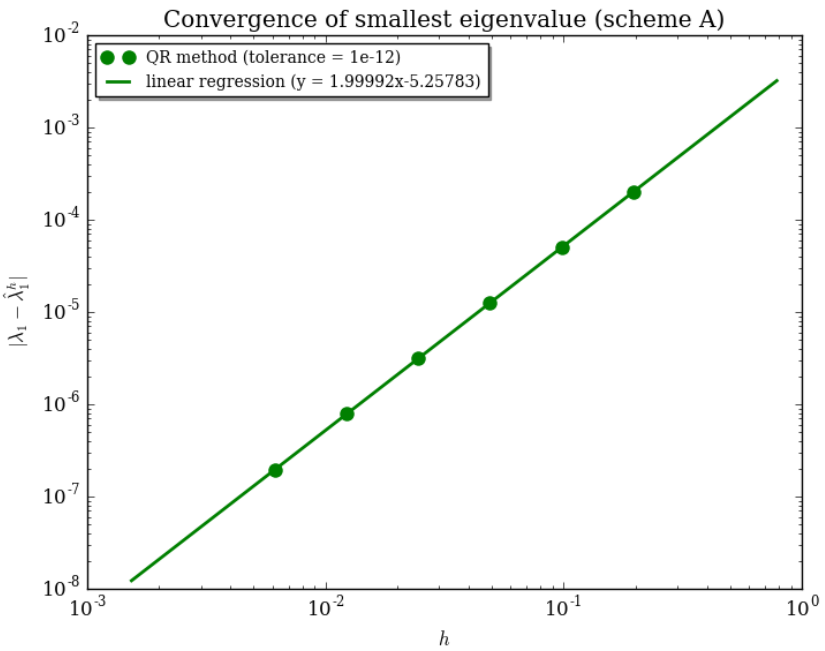


Figure 3: QR iteration with deflation applied to scheme A.

| method | $N$ | iterations | smallest eigenvalue |
|---|---|---|---|
| inverse power | 16 | 10 | 5.250201 |
| | 32 | 9 | 5.250050 |
| | 64 | 8 | 5.250013 |
| | 128 | 7 | 5.250003 |
| | 256 | 6 | 5.250001 |
| | 512 | 5 | 5.250000 |

Table 2: The results for scheme B discretization.

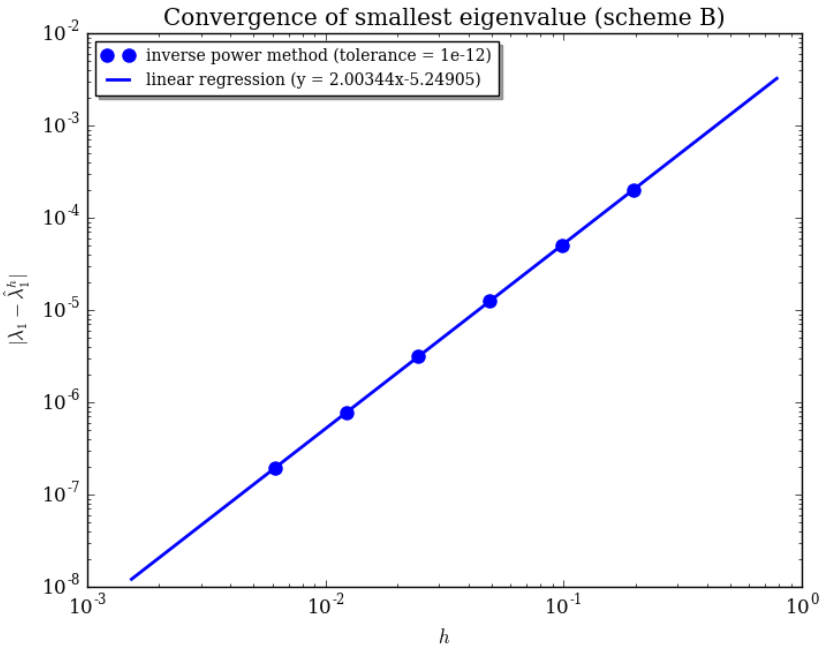


Figure 4: Inverse power method applied to scheme B.