MATH 5330: Computational Methods of Linear Algebra

Lecture Note 14: Eigenvalue Problem for General Matrices

Xianyi Zeng
Department of Mathematical Sciences, UTEP

1 The Complex Schur Decomposition

If we want to find the matrices whose eigenvalues are easy to compute, diagonal ones must be on the top of the list and the second one should be upper-triangular (or lower-triangular) matrices – the eigenvalues are just the diagonal elements. In the Jacobi iteration method for symmetric matrices, we try to reduce a real symmetric matrix A to a diagonal one by using Givens rotations. Extending this strategy to general matrices requires the so called *Schur decomposition*: Let $A \in \mathbb{C}^{n\times n}$, the Schur decomposition is defined as:

$$A = QUQ^*, \tag{1.1}$$

where * denotes the Hermitian of a complex matrix and Q is unitary, i.e., $Q \in \mathbb{C}^{n\times n}$ and $QQ^* = I$; and $U \in \mathbb{C}^{n\times n}$ is upper-triangular. Then the characteristic polynomials of A and U are the same; and the roots for the latter are just the diagonal elements of U.

The good news is the Schur decomposition exists for any complex matrix A so that the eigenvalue problem for A is equivalent to finding its Schur decomposition. The bad news, however, is that we cannot simply claim the same only using real matrices. In fact, if $A = QUQ^t$ where A,Q,$U \in \mathbb{R}^{n\times n}$, Q being orthogonal and U being upper-triangular would imply that all the eigenvalues of A are real, which is not true.

Now we have two possible choices. First, we can treat the real matrix A as a complex one and try to find its Schur decomposition approximately, just like what we did in Jacobi iteration method, and find complex approximations to all its eigenvalues. The problem with this approach is that replacing real arithmetics with complex ones significantly increase the number of flops; furthermore, the eigenvalue problem is usually a sub-problem of a much larger program, which may not be equipped with the complex number capability.

In the second choice, we give up on reducing A to upper-triangular form but accept an upper Hessenberg form:

$$A = QHQ^t, \tag{1.2}$$

and keep everything real. Now the eigenvalue problem for A reduces to two parts: Part one involves reducing A to H by an orthogonal matrix Q; and part two is to solve the eigenvalue problem for the upper Hessenberg matrix H. The next sections will focus on this path, and eventually lead to the method called *real Schur decomposition*.

2 Reducing to Upper Hessenberg Form

The idea to reduce a general matrix $A \in \mathbb{R}^{n\times n}$ is fairly similar to the Jacobi iteration method where we use the Givens rotation to reduce a symmetric matrix to diagonal form. However, the strategy

here is to use $G_{ij}(\theta)$, where $i>j$, to eliminate any non-zero $a_{i,j-1}$ rather than a_{ij}. Again, we denote $c=\cos(\theta)$ and $s=\sin(\theta)$ and will try to figure out what they should be to fulfill our purpose.

Because post-multiplying by G_{ij}^t will only modify the i-th and the j-th columns, we only need to make sure the $(i,j-1)$-th component of $G_{ij}A$ is zero:

$$G_{ij}A=\begin{bmatrix} & & \vdots & & \vdots & \\ \cdots & ca_{j,j-1}-sa_{i,j-1} & ca_{jj}-sa_{ij} & \cdots & ca_{ji}-sa_{ii} & \cdots \\ & & \vdots & & \vdots & \\ \cdots & sa_{j,j-1}+ca_{i,j-1} & sa_{jj}+ca_{ij} & \cdots & sa_{ji}+ca_{ii} & \cdots \\ & & \vdots & & \vdots & \end{bmatrix},$$

that is:

$$sa_{j,j-1}+ca_{i,j-1}=0\,.$$

Hence we set:

$$s=-\frac{a_{i,j-1}}{\sqrt{a_{j,j-1}^2+a_{i,j-1}^2}}\,,\quad c=\frac{a_{j,j-1}}{\sqrt{a_{j,j-1}^2+a_{i,j-1}^2}}\,.$$

The denominator will never be zero since we only perform the operation when $a_{i,j-1}\neq 0$. Once the Givens matrix G_{ij} is computed, $A\mapsto G_{ij}AG_{ij}^t$ will have zero at the $(i,j-1)$-th location. Unlike in the Jacobi iteration method, it is possible to order the Givens rotations carefully to put all (i,j)-component to zero, where $i>j+1$. The strategy is very similar to the QR decomposition that we had before. That is, we eliminate every non-zero element starting from the first column and move right towards the $(n-2)$-th column; and in each column we follow a top-down strategy. This is illustrated by the next example for a 5×5 matrix, where in each operation the entries to be changed are marked by blue except the one to be eliminated, which is marked by red.

$$\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix}\to\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix}\to\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ * & * & * & * & * \end{bmatrix}\to\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \end{bmatrix}$$

$$\to\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & * & * & * & * \end{bmatrix}\to\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{bmatrix}\to\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \end{bmatrix}.$$

Paying special attention to the moves starting from the last one in the first row, when a zero entry is encountered it is always paired with another zero entry so that previous work is not ruined.

In the end, we can use at most $(n-1)(n-2)/2$ Givens matrices to transform A to an upper Hessenberg form. Also note that if the same strategy is applied to a symmetric matrix, the result is a tridiagonal matrix. As a final note, as in the orthogonalization process, we can also use Householder transforms to achieve the same goal here.

3 The QR Method and a Special Case

The QR method dates back to 1950s and is one of the oldest methods to compute all the eigenvalues of a real or complex matrices. The method is well-known for its simplicity: Let us start with $A^{(0)} = A$, then when $A^{(k)}$, $k \geq 0$ is available we compute its QR decomposition:

$$A^{(k)} = Q_k R_k \,, \tag{3.1}$$

where Q_k is orthogonal and R_k is upper-triangular. R_k usually has a different eigenstructure from $A^{(k)}$; but we can use Q_k to find a similar transformation of $A^{(k)}$ to obtain the next iterate as:

$$A^{(k+1)} = Q_k^t A^{(k)} Q_k = R_k Q_k \,. \tag{3.2}$$

The hope is that when enough iterates are used, $A^{(k)}$ has a form whose eigenvalues are easy to compute. The following theorem shows that the convergence is guaranteed in a special situation of the general matrices; but in this case, the convergence is linear hence it is crucial in practice not to apply the QR method to the original matrix A, but to its reduced Hessenberg form (see previous section) for which the QR decomposition costs much less.

Theorem 3.1. *If A has n real eigenvalues satisfying:*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0 \,, \tag{3.3}$$

then in the QR method, $A^{(k)}$ approaches upper triangular form as $k \to +\infty$; or equivalently $a_{ij}^{(k)} \to 0$ as $k \to \infty$ for all $i > j$.

First we note that the assumption all λ_i are real is redundant, since if A has a complex eigenvalue $\lambda = a + ib$ where $a,b \in \mathbb{R}$ and $b \neq 0$, then $\overline{\lambda} = a - ib$ is also an eigenvalue of A and $|\lambda| = \left|\overline{\lambda}\right|$.

Proof. Let us look at the goal here: Because A has n distinct real eigenvalues, it is actually diagonalizable by its eigenvalues:

$$A = T^{-1} D T \,,$$

where the diagonal matrix D is arranged such that $d_{ii} = \lambda_i$ and the i-th column of T^{-1} is its non-zero eigenvector. Thus we want to write $A^{(k)}$ in a form similar to this diagonal decomposition, i.e., similar to a diagonal matrix.

By construction, we have:

$$Q_0 Q_1 \cdots Q_{k-1} A^{(k)} = A Q_0 \cdots Q_{k-1} \quad \text{or} \quad Q_0 Q_1 \cdots Q_k R_k = A Q_0 \cdots Q_{k-1} \,.$$

Now we multiply both sides by R_{k-1}, R_{k-2}, $\cdots$, and R_0 consecutively to have:

$$\begin{aligned}
Q_0 Q_1 \cdots Q_k R_k R_{k-1} &= A Q_0 \cdots Q_{k-1} R_{k-1} = A Q_0 \cdots Q_{k-2} A^{(k-1)} \,, \\
Q_0 Q_1 \cdots Q_k R_k R_{k-1} R_{k-2} &= A Q_0 \cdots (Q_{k-2} A^{(k-1)} Q_{k-2}^t) Q_{k-2} R_{k-2} = A Q_0 \cdots Q_{k-3} (A^{(k-2)})^2 \,, \\
&\vdots \\
Q_0 Q_1 \cdots Q_k R_k \cdots R_0 &= A (A^{(0)})^k = A^{k+1} = T^{-1} D^{k+1} T \,.
\end{aligned}$$

This helps us to use the similarity relation:

$$A^{(k+1)} = (Q_0 \cdots Q_k)^{-1} A (Q_0 \cdots Q_k) = (T Q_0 \cdots Q_k)^{-1} D (T Q_0 \cdots Q_k) \,,$$

and define $T_k = T Q_0 \cdots Q_k$, so that $A^{(k+1)} = T_k^{-1} D T_k$.

Because A is non-singular, so is every $A^{(k)}$ and every R_k; hence we deduce:

$$T_k = T Q_0 \cdots Q_k = D^{k+1} T R_0^{-1} \cdots R_k^{-1} \,.$$

All R_k^{-1} are upper-triangular; so let us consider the LU decomposition of T: $T = LU$ where L has ones on its diagonal. Then

$$T_k = D^{k+1} L U R_0^{-1} \cdots R_k^{-1} = (D^{k+1} L D^{-(k+1)}) K_k \,,$$

where $K_k = D^{k+1} U R_0^{-1} \cdots R_k^{-1}$ is upper-triangular. It is not difficult to see that the ji-component ($i < j$) of D^{k+1} converges to zero linearly at the rate $\left|\frac{\lambda_j}{\lambda_i}\right|$; hence $D^{k+1} \to I$ linearly at the rate $\max_{i<j} \left|\frac{\lambda_j}{\lambda_i}\right| < 1$ (see exercises). So T_k approaches an upper-triangular form as $k \to \infty$. Let $T_k = L_k + U_k$ and $T_k^{-1} = \tilde{L}_k + \tilde{U}_k$ be such that L_k and $\tilde{L}_k$ is lower-triangular with zero diagonal elements and U_k and $\tilde{U}_k$ are upper-triangular , then we already know $L_k \to 0$ and $\tilde{L}_k \to 0$. Furthermore, because $||T_k||_{\mathrm{F}} = ||T||_{\mathrm{F}}$ and $\left|\left|T_k^{-1}\right|\right|_{\mathrm{F}} = \left|\left|T^{-1}\right|\right|_{\mathrm{F}}$, we have:

$$\begin{aligned} ||T||_{\mathrm{F}}^2 &= ||T_k||_{\mathrm{F}}^2 = ||L_k||_{\mathrm{F}}^2 + ||U_k||_{\mathrm{F}}^2 \geq ||U_k||_{\mathrm{F}}^2 \,, \\ \left|\left|T^{-1}\right|\right|_{\mathrm{F}}^2 &= \left|\left|T_k^{-1}\right|\right|_{\mathrm{F}}^2 = \left|\left|\tilde{L}_k\right|\right|_{\mathrm{F}}^2 + \left|\left|\tilde{U}_k\right|\right|_{\mathrm{F}}^2 \geq \left|\left|\tilde{U}_k\right|\right|_{\mathrm{F}}^2 \,. \end{aligned}$$

Thus the sequences $\{U_k\}$ and $\{\tilde{U}_k\}$ remain bounded as $k \to \infty$. Finally, we have:

$$A^{(k+1)} = T_k^{-1} D T_k = (\tilde{L}_k + \tilde{U}_k) D (L_k + U_k) = \tilde{L}_k D L_k + \tilde{U}_k D L_k + \tilde{L}_k D U_k + \tilde{U}_k D U_k \to \tilde{U}_k D U_k \,,$$

where we used the boundedness of $\tilde{U}_k$ and U_k to deduce the diminishing of the second term and the third term. Hence the distance from $A^{(k+1)}$ from an upper triangular form $\tilde{U}_k D U_k$ is decreasing to zero as $k \to \infty$. □

The fact that we can approximate the eigenvalues of A by the diagonal elements of $A^{(k+1)}$ when the latter is sufficiently close to an upper-triangular form is due to a result by Hoffman and Wieland, of which a special version is used for establishing a similar claim regarding the Jacobi iteration method for symmetric matrices.

The major concern about the QR method is that it requires a QR decomposition ($\sim O(n^3)$) in every iteration, and its slow convergence. However, we already see that reducing A to an upper Hessenberg form H can be done in $O(n^3)$ time; now let us look at applying the QR method to this matrix H.

First, we observe that the QR iteration preserve the upper Hessenberg form. That is:

> *If H is upper Hessenberg and $H = QR$ is the QR decomposition of H, then $Q^t H Q$ is also upper Hessenberg.*

In fact, we clearly see that $Q = HR^{-1}$ is the product of an upper Hessenberg matrix and an upper triangular one, hence Q is also upper Hessenberg. Thus $Q^tHQ = RQ$ is the product of an upper triangular matrix and an upper Hessenberg matrix, and it is upper Hessenberg.

Next, we recall that at most $n-1$ Givens rotations are required to compute the QR decomposition of an upper Hessenberg matrix; thus both finding $H = QR$ and computing $Q^tHQ = RQ$ takes time $O(n^2)$ if Q is treated as the product of a sequence of Givens rotations. In the end, we reduce the computational cost of each iteration of the QR method to $O(n^2)$, and the total cost is kept down to $O(n^3)$ if $O(n)$ iterations are used. The $O(n)$ iterations assumption is reasonable due to the linear convergence of $A^{(k)}$ to an upper-triangular form.

Finally, it is remarked that we only showed that the difference between $A^{(k)}$ and **a** upper-triangular form is getting smaller, but nothing is said about the convergence of $A^{(k)}$. So in the program, the difference of two consecutive iterates should not be used to check whether we want to terminate the iterations; instead, we can use the Frobenius norm of the lower-triangular part (excluding the diagonal entries) of $A^{(k)}$.

4 The Practical QR Method and the Real Schur Decomposition

The problem with applying the QR method to general real matrices is of course, that the condition in Theorem 3.1 does not usually hold. So will the sequence $A^{(k)}$ approach something else if those conditions are not satisfied, at least with some modifications to the QR method?

The answer is affirmative and given by the *real Schur form* (c.f. upper-triangular form in Theorem 3.1). We say that a matrix S is of the real Schur form, or quasi-upper-triangular form if it has the following block structure:

$$S = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1m} \\ 0 & S_{22} & \cdots & S_{2m} \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & S_{mm} \end{bmatrix}, \tag{4.1}$$

where each S_{ii} is either 1×1 or 2×2 having complex conjugate eigenvalues. The idea is that if A is similar to a quasi-upper-triangular matrix S, then its eigenvalues is given by the collection of all eigenvalues of the diagonal blocks of S. These are either a real number if S_{ii} is 1×1, or two complex numbers that are conjugate to each other if S_{ii} is 2×2. The good news is that one can prove for every $A\in\mathbb{R}^n$, there exists an orthogonal $Q\in\mathbb{R}^n$, such that:

$$Q^tAQ = S \tag{4.2}$$

is in the real Schur form.

Theorem 4.1. *Let $A\in\mathbb{R}^{n\times n}$ be arbitrary, then there exist an orthogonal matrix $Q\in\mathbb{R}^{n\times n}$ and a quasi-upper-triangular matrix $S\in\mathbb{R}^{n\times n}$ such that $A = QSQ^t$.*

Proof. Let us use induction on the integer m, the number of pairs of complex conjugate roots of the characteristic polynomial of A. As the base case, we have $m = 0$ and A only have real eigenvalues. Thus the Jordan canonical form of A can be accomplished with all real matrices, that is: We can find non-singular $T\in\mathbb{R}^{n\times n}$ and upper triangular matrix $J\in\mathbb{R}^{n\times n}$ such that:

$$A = TJT^{-1}\,.$$

Now we find the QR decomposition of T: $T=QR$ where Q is orthogonal and R is upper-triangular; hence:

$$A=QRJ(QR)^{-1}=QSQ^t\,,$$

where $S=RJR^{-1}\in\mathbb{R}^{n\times n}$ is upper-triangular. This is clearly the Schur decomposition of A.

Now we suppose the real Schur decomposition for matrices with no more than m pair of complex conjugate eigenvalues, where $m\geq 0$ is an integer; and we consider the case when A has $m+1\geq 1$ such pairs. Let $\lambda_1=\alpha+i\beta$ ($\alpha,\beta\in\mathbb{R}$ and $\beta\neq 0$) be a complex eigenvalue of A, and $\boldsymbol{w}=\boldsymbol{u}+i\boldsymbol{v}$ be an eigenvector. Here $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^n$ and they are not simultaneously zero. Comparing the real parts and the imaginary parts of both sides of $A\boldsymbol{w}=\lambda_1\boldsymbol{w}$ we obtain:

$$\begin{aligned}A\boldsymbol{u}&=\alpha\boldsymbol{u}-\beta\boldsymbol{v}\,,\\ A\boldsymbol{v}&=\beta\boldsymbol{u}+\alpha\boldsymbol{v}\,.\end{aligned}$$

In order to show that $\boldsymbol{u}$ and $\boldsymbol{v}$ are linearly independent, let $a\boldsymbol{u}+b\boldsymbol{v}=0$ for some $a,b\in\mathbb{R}$, then:

$$\begin{aligned}&0=A(a\boldsymbol{u}+b\boldsymbol{v})=a(\alpha\boldsymbol{u}-\beta\boldsymbol{v})+b(\beta\boldsymbol{u}+\alpha\boldsymbol{v})\\ &\quad=\alpha(a\boldsymbol{u}+b\boldsymbol{v})+\beta(-a\boldsymbol{v}+b\boldsymbol{u})=\beta(-a\boldsymbol{v}+b\boldsymbol{u})\\ \Rightarrow\quad&0=-a\boldsymbol{v}+b\boldsymbol{u}\\ \Rightarrow\quad&0=a(a\boldsymbol{u}+b\boldsymbol{v})+b(-a\boldsymbol{v}+b\boldsymbol{u})=(a^2+b^2)\boldsymbol{u}\,.\end{aligned}$$

Now if $\boldsymbol{u}=0$, we have $-\beta\boldsymbol{v}=A\boldsymbol{u}-\alpha\boldsymbol{u}=0$ hence $\boldsymbol{v}=0$, contradiction. Thus we must have $a^2+b^2=0$ or $a=b=0$; or equivalently $\boldsymbol{u}$ and $\boldsymbol{v}$ are linearly independent.

Let $[\boldsymbol{u}\ \boldsymbol{v}]=Q_1R_1$ be the partial QR decomposition of the $n\times 2$ matrix $[\boldsymbol{u}\ \boldsymbol{v}]$, here $Q_1\in\mathbb{R}^{n\times 2}$ has two orthonormal column vectors and $R_1\in\mathbb{R}^{2\times 2}$ is non-singular and upper-triangular. Following:

$$A[\boldsymbol{u}\ \boldsymbol{v}]=[\boldsymbol{u}\ \boldsymbol{v}]S_1\,,\quad S_1=\begin{bmatrix}\alpha & -\beta\\ \beta & \alpha\end{bmatrix}$$

we have $AQ_1R_1=Q_1R_1S_1$ or $Q_1^tAQ_1=R_1S_1R_1^{-1}$. Let $P=[Q_1\ Q_2]\in\mathbb{R}^{n\times n}$ be an orthogonal matrix, where $Q_2\in\mathbb{R}^{n\times(n-2)}$, there is:

$$P^tAP=\begin{bmatrix}Q_1^tAQ_1 & Q_1^tAQ_2\\ Q_2^tAQ_1 & Q_2^tAQ_2\end{bmatrix}=\begin{bmatrix}Q_1^tAQ_1 & Q_1^tAQ_2\\ Q_2^tQ_1R_1S_1R_1^{-1} & Q_2^tAQ_2\end{bmatrix}=\begin{bmatrix}R_1S_1R_1^{-1} & Q_1^tAQ_2\\ 0 & Q_2^tAQ_2\end{bmatrix}$$

It is not difficult to see that $p_A(\lambda)=(\lambda-\alpha-i\beta)(\lambda-\alpha+i\beta)p_{Q_2^tAQ_2}(\lambda)$ so the $(n-2)\times(n-2)$ real matrix $Q_2^tAQ_2$ has exactly one less pair of complex conjugate eigenvalues than A, and by induction assumption, it has a real Schur decomposition:

$$Q_2^tAQ_2=\hat{Q}\hat{S}\hat{Q}^t\,.$$

Now we define:

$$Q=P\begin{bmatrix}I_2 & \\ & \hat{Q}\end{bmatrix},$$

which is clearly orthogonal, and there is:

$$\begin{aligned}Q^tAQ&=\begin{bmatrix}I_2 & \\ & \hat{Q}^t\end{bmatrix}P^tAP\begin{bmatrix}I_2 & \\ & \hat{Q}\end{bmatrix}=\begin{bmatrix}I_2 & \\ & \hat{Q}^t\end{bmatrix}\begin{bmatrix}R_1S_1R_1^{-1} & Q_1^tAQ_2\\ 0 & Q_2^tAQ_2\end{bmatrix}\begin{bmatrix}I_2 & \\ & \hat{Q}\end{bmatrix}\\ &=\begin{bmatrix}I_2 & \\ & \hat{Q}^t\end{bmatrix}\begin{bmatrix}R_1S_1R_1^{-1} & Q_1^tAQ_2\\ 0 & \hat{Q}\hat{S}\hat{Q}^t\end{bmatrix}\begin{bmatrix}I_2 & \\ & \hat{Q}\end{bmatrix}=\begin{bmatrix}R_1S_1R_1^{-1} & Q_1^tAQ_2\hat{Q}\\ 0 & \hat{S}\end{bmatrix}.\end{aligned}$$

It is not difficult to see that the last entity is of the real Schur form. □

Thus the real Schur form seems to be an adequate tool for solving eigenvalue problem for general matrices. To modify the previous QR method so that the iterates will approach the real Schur form, we need several steps that are went through briefly below.

Unreduced Hessenberg matrices. Let H be an upper Hessenberg matrix. If $h_{i+1,i}=0$ for some $1\le i\le n-1$, we say H is *reducible* since the eigenvalue problem for H reduces to two sub-problems for $H(1:i,1:i)$ and $H((i+1):n,(i+1):n)$, respectively. In practice, we check if $|h_{i+1,i}|$ is sufficiently small for some i, and if it is, we set it to zero and solve two smaller problems. For this reason, in the following analysis we always assume that H is not reducible, called *unreduced*; and our target is to let the iterates to approach a reducible form (hence to decrease the size of the problem). An important property of unreduced Hessenberg matrix is that it can only have simple eigenvalues (see Exercise 2).

The shifted QR method. Instead of computing $H^{(k)}=Q_kR_k$ and $H^{(k+1)}=R_kQ_k$; we can use the idea from the shifted power method and compute for some $\mu\in\mathbb{R}$:

$$H^{(k)}-\mu I=Q_kR_k \quad \text{and} \quad H^{(k+1)}=R_kQ_k+\mu I\,. \tag{4.3}$$

Following previous analysis we see if the eigenvalues $\{\lambda_i\}$ of A (or H) are organized such that:

$$|\lambda_1-\mu|\ge\cdots\ge|\lambda_n-\mu|\,,$$

then the $(i+1,i)$-th component of $H^{(k)}$ converges to zero linearly with the rate:

$$\left|\frac{\lambda_{i+1}-\mu}{\lambda_i-\mu}\right|.$$

If $\lambda_i=\lambda_{i+1}$ or $\lambda_i=\overline{\lambda_{i+1}}$, there is clearly no convergence at all. The first of these two special cases will not happen if H is unreduced (Exercise 2), and the second one is handled by a separate procedure.

To illustrate the idea, we assume first that neither special case occurs, then the hope is to find some μ such that it is much closer to λ_n than to the other eigenvalues, consequently the zeroing of the $(n,n-1)$-entry is very fast. This idea leads to the *single shift strategy*, where we vary μ from iteration to iteration by setting $\mu^{(k)}=h^{(k)}_{n,n}$. The rationale is to have $h^{(k)}_{n,n-1}\to 0$ fast, so that $h^{(k)}_{n,n}$ approaches quickly to an eigenvalue of H; hence $h^{(k)}_{n,n}$ provides a good estimate for this eigenvalue.

The single shift strategy will of course fail, if $h^{(k)}_{n,n-1}$ does not converge to zero and $h^{(k)}_{n,n}$ is not a good approximation to an eigenvalue. This can happen, for example if $h^{(k)}_{n-1,n-2}$ approaches zero and the trailing 2×2 sub-matrix:

$$\begin{bmatrix} h^{(k)}_{n-1,n-1} & h^{(k)}_{n-1,n} \\ h^{(k)}_{n,n-1} & h^{(k)}_{nn} \end{bmatrix}$$

has two complex conjugate eigenvalues.

A way around this difficulty is the so called *double shift strategy*, where in order to update from $H^{(k)}$ to $H^{(k+1)}$, we first compute the two eigenvalues α_1 and α_2 of the preceding trailing matrix,

and compute:

$$\begin{aligned} H^{(k)} - \alpha_1 I &= Q_k R_k \\ H^{(k+1/2)} &= R_k Q_k + \alpha_1 I \\ H^{(k+1/2)} - \alpha_2 I &= Q_{k+1/2} R_{k+1/2} \\ H^{(k+1)} &= R_{k+1/2} Q_{k+1/2} + \alpha_2 I \,. \end{aligned}$$

As a result:

$$(Q_k Q_{k+1/2})(R_{k+1/2} R_k) = (H^{(k)} - \alpha_1 I)(H^{(k)} - \alpha_2 I) \stackrel{\text{def}}{=\!=} M_k \,.$$

The net outcome of the double shift strategy is an orthogonal matrix $Z_k = Q_k Q_{k+1/2}$, such that:

$$H^{(k+1)} = Z_k^t H^{(k)} Z_k \,;$$

and a real upper-triangular matrix $U_k = R_{k+1/2} R_k$ such that:

$$Z_k R_k = M_k = (H^{(k)})^2 - (h^{(k)}_{n-1,n-1} + h^{(k)}_{nn}) H^{(k)} + (h^{(k)}_{n-1,n-1} h^{(k)}_{nn} - h^{(k)}_{n-1,n} h^{(k)}_{n,n-1}) I \,.$$

This is the QR decomposition of a real matrix given by the right hand side of the last equality, which has two non-zero sub-diagonals; thus no complex-number operations are needed.

A major grudge towards this approach is the $O(n^3)$ cost in calculating the square of $H^{(k)}$; people have developed an alternative *double implicit shift strategy* to compute Z'_k and R'_k that are almost Z_k and R_k. The theoretical foundation is the *Implicit Q Theorem*, which says that if $A = Q^t H Q$ with orthogonal Q and unreduced H, then Q and H are almost unique (up to a difference in the signs in their components) providing that the first column of Q is fixed. Due to this theorem, people only form explicit the first column of M_k, compute the first column of Z_k, and use Householder transformations to build up the subsequent columns to construct Z'_k that is almost identical to Z_k. This process does not require constructing M explicitly, and the computation cost is $O(n^2)$.

There are two possible outcomes of the double (implicit) shift strategy, that is either $h_{n,n-1}$ or $h_{n-1,n-2}$ will converge to zero, leaving us a 1×1 or 2×2 block in the final real Schur form. This is known as the *Francis QR step*.

The QR method by Golub and van Loan. Combining all together, we obtain a practical QR method as described by the following procedure:

Algorithm 4.1 QR Method

Compute the upper Hessenberg form $H = GAG^t$.
Set $q = 0$.
while $q < n$ **do**
  Set to zero all $h_{i,i-1}$ such that $|h_{i,i-1}| \le \varepsilon_0(|h_{ii}| + |h_{i-1,i-1}|)$.
  Find the largest $q \ge 0$ and the smallest $p \ge 0$ such that:

$$H = \begin{bmatrix} H_{11} & H_{12} & H_{13} \\ 0 & H_{22} & H_{23} \\ 0 & 0 & H_{33} \end{bmatrix}$$

  where $H_{33} \in \mathbb{R}^{q \times q}$ is of real Schur form and $H_{22} \in \mathbb{R}^{(n-p-q)\times(n-p-q)}$ is unreduced.
  if $q < n$ **then**
    Perform a Francis QR step on H_{22}: $H_{22} \leftarrow Z^t H_{22} Z$.
  end if
end while
Upper triangularize all 2×2 diagonal blocks in H that have real eigenvalues.

5 Single Vector Iterations

Similar to the symmetric case, when only a few special eigenvalues are desired it is not necessary to use the QR method, which will eventually provide all the eigenvalues; and we consider the analogue of the power method, the inverse iteration method, and the Rayleigh Quotient Iteration method.

Algorithm 5.1 The Power Method

1: Set $\varepsilon_0 > 0$ and $\boldsymbol{x}_0$ such that $||\boldsymbol{x}_0|| = 1$.
2: **for** $i = 1,2,\cdots$ **do**
3: Compute $\boldsymbol{y}_i = A\boldsymbol{x}_{i-1}$.
4: Compute $\boldsymbol{x}_i = \boldsymbol{y}_i / ||\boldsymbol{y}_i||$.
5: **if** $||\boldsymbol{x}_i - \boldsymbol{x}_{i-1}|| < \varepsilon_0$ **then**
6: Break.
7: **end if**
8: **end for**

For example, the power method for general matrices (Algorithm 5.1) is exactly the same as that for symmetric matrices. The condition for convergence of $\boldsymbol{x}_i$ to an eigenvector of the eigenvalue of A that has the largest modulus is also very similar to previous case, as described by the next theorem.

Theorem 5.1. *If A has one and only one eigenvalue λ_1 and λ_1 is semi-simple (that is, its algebraic multiplicity is the same as the geometric multiplicity), then either $\boldsymbol{x}_0 \in \mathcal{E}_{\lambda_1}^{\perp}$ or the sequence $\{\boldsymbol{x}_i\}$ converges to an eigenvector of λ_1 and $||\boldsymbol{y}_i||$ converges to $|\lambda_1|$.*

Note this theorem relaxes the condition on λ_1 a little bit, and the result applies also to symmetric matrices. In particular, because all eigenvalues of a symmetric matrix A are real and semi-simple,

we can always use a suitable shift to have the power method converge to either the largest or the smallest eigenvalue of A.

Proof. We consider the Jordan canonical form of A:

$$A = [P_1 \ P_2] \begin{bmatrix} \lambda_1 I_m & \\ & \hat{J} \end{bmatrix} [P_1 \ P_2]^{-1},$$

where m is the geometric multiplicity of λ_1, the columns of P_1 form an orthonormal basis of $\mathcal{E}_{\lambda_1}$ and the columns of P_2 form a basis of $\mathcal{E}_{\lambda_1}^{\perp}$. Let $\boldsymbol{x}_0 = P_1\boldsymbol{u}_1 + P_2\boldsymbol{u}_2$ where $\boldsymbol{u}_1 \in \mathbb{R}^m$ and $\boldsymbol{u}_2 \in \mathbb{R}^{n-m}$. Then the non-normalized iterate $\boldsymbol{z}_k = A^k\boldsymbol{x}_0$ is given by:

$$\boldsymbol{z}_k = A^k[P_1 \ P_2] \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} = [P_1 \ P_2] \begin{bmatrix} \lambda_1^k I_m & \\ & \hat{J}^k \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix} = \lambda_1^k \Big[P_1\boldsymbol{u}_1 + P_2(\lambda_1^{-1}\hat{J})^k\boldsymbol{u}_2 \Big]$$

Note that $P_1\boldsymbol{u}_1$ and $P_2(\lambda_1^{-1}\hat{J})^k\boldsymbol{u}_2$ are orthogonal to each other, thus the normalization gives:

$$\boldsymbol{x}_k = \frac{\boldsymbol{z}_k}{||\boldsymbol{z}_k||} = \frac{P_1\boldsymbol{u}_1 + P_2(\lambda_1^{-1}\hat{J})^k\boldsymbol{u}_2}{\sqrt{||P_1\boldsymbol{u}_1||^2 + \Big|\Big|P_2(\lambda_1^{-1}\hat{J})^k\boldsymbol{u}_2\Big|\Big|^2}}.$$

It is not difficult to see that $(\lambda_1^{-1}\hat{J})^k \to 0$ as $k \to \infty$; thus $\boldsymbol{x}_k \to (P_1\boldsymbol{u}_1)/||P_1\boldsymbol{u}_1||$ unless $P_1\boldsymbol{u}_1 = 0$, or equivalently $\boldsymbol{x}_0 \in \mathcal{E}_{\lambda_1}^{\perp}$.

Lastly, when $\boldsymbol{x}_k \to \boldsymbol{x}$, we see that $\boldsymbol{y}_k \to A\boldsymbol{x} = \lambda_1\boldsymbol{x}$, and $||\boldsymbol{y}_k|| = ||\boldsymbol{y}_k||/||\boldsymbol{x}_k|| = |\lambda_1|$. □

The power method usually finds the eigenvalue with the largest modulus, or if we combine it with the *shift* strategy, one of the vertex of the convex hull of the spectrum of A. Note that Algorithm 5.1 applies both to pure real operations and complex variable operations; in the former case it is not possible to find a complex eigenvalue of A, nor does this case satisfies the assumption of the theorem.

If solving the linear system with A is easy and A is non-singular, we can also use the *inverse iteration method* to find the eigenvalue of A that is closest to zero.

Algorithm 5.2 The Inverse Iteration Method

1: Set $\varepsilon_0 > 0$ and $\boldsymbol{x}_0$ such that $||\boldsymbol{x}_0|| = 1$.
2: **for** $i = 1,2,\cdots$ **do**
3: Solve $A\boldsymbol{y}_i = \boldsymbol{x}_{i-1}$.
4: Compute $\boldsymbol{x}_i = \boldsymbol{y}_i/||\boldsymbol{y}_i||$.
5: **if** $||\boldsymbol{x}_i - \boldsymbol{x}_{i-1}|| < \varepsilon_0$ **then**
6: Break.
7: **end if**
8: **end for**

When the conditions of Theorem 5.1 are satisfied for A^{-1}, the inverse iteration method converges: $\{\boldsymbol{x}_i\}$ converges to an eigenvector of the eigenvalue with the smallest modulus of A, and $||\boldsymbol{y}_i||^{-1}$ converges to the modulus of this eigenvalue. Combining with shifting, it is possible to use

the inverse iteration method to find any semi-simple eigenvalue that is the only one has the same modulus. Both the power method and the inverse iteration method converge at most linearly, if they converge at all.

Finally, we notice that even for general matrices A, $\boldsymbol{v}^t A\boldsymbol{v}$ is a good approximation to an eigenvalue λ if $\boldsymbol{v}$ is close to a unit eigenvector of this eigenvalue. Thus one can also extend the RQI method to general matrices without trouble. However, this method is less utilized for several reasons. First of all, global convergence is only known for diagonalizable matrices. Secondly, even when the method converges, the provable convergence rate is at best quadratic for general matrices, slower than the cubic convergence in the case of symmetric matrices. Last but not the least, unlike in the symmetric case where we can reduce A to a tridiagonal form so that the linear solve is much easier, we still need to solve a linear system with a rather full matrix (the upper Hessenberg form) with the tools at hand. For this reason, RQI is much less utilized for solving eigenvalue problems of general matrices.

At the end of this section, we briefly discuss the deflation in the case of general matrix A. The Wielandt deflation technique uses an eigenpair $(\lambda,\boldsymbol{v})$ of A to replace the eigenvalue λ by zero; we assume that $\boldsymbol{v}$ is a unit vector. Particularly, let $\boldsymbol{u}$ be any vector such that $\boldsymbol{v}^t\boldsymbol{u}=1$, then we define:

$$B = A - \lambda \boldsymbol{v}\boldsymbol{u}^t .$$

We only consider the case of a real eigenvalue λ. To show that the spectrum of B only differ from that of A by replacing λ with 0, we construct a real Schur decomposition $A=QSQ^t$, where the first diagonal block of S is 1×1 with the value λ and the first column of Q is given by $\boldsymbol{v}$. Let $Q=[\boldsymbol{v}\ \hat{Q}]$, then:

$$Q^t B Q = S - \lambda \begin{bmatrix} \boldsymbol{v}^t \\ \hat{Q}^t \end{bmatrix} \boldsymbol{v}\boldsymbol{u}^t [\boldsymbol{v}\ \hat{Q}] = S - \lambda \begin{bmatrix} 1 \\ 0 \end{bmatrix} [1\ \boldsymbol{u}^t\hat{Q}] = S - \begin{bmatrix} \lambda & \lambda\boldsymbol{u}^t\hat{Q} \\ 0 & 0 \end{bmatrix}$$

is clearly of real Schur form, with the diagonal block λ replaced by zero.

6 The Generalized Eigenvalue Problem

The generalized eigenvalue problem $A\boldsymbol{x} = \lambda B\boldsymbol{x}$ can appear in various circumstances. For example, let us consider the eigenvalue problem for the one-dimensional Poisson operator and its numerical approximations:

$$\begin{cases} u''(x) = \lambda u(x) & 0 \le x \le 1\,, \\ u(0) = u(1) = 0\,. \end{cases} \tag{6.1}$$

It is not difficult to see that the solutions are given by:

$$\lambda_k = -k^2\pi^2\,, \quad \varphi_k(x) = \sin(k\pi x)\,, \quad k = 1,2,\cdots. \tag{6.2}$$

If we use finite difference method to approximate this differential equation, say using $N+1$ uniformly distributed points $x_i = ih$, $h = 1/N$ and $i = 0,1,\cdots,N$. Let the approximation of $u(x_i)$ be u_i then the central difference leads to:

$$\begin{cases} u_0 &= 0 \\ \frac{1}{h^2}(u_0 - 2u_1 + u_2) &= \lambda u_1 \\ \frac{1}{h^2}(u_1 - 2u_2 + u_3) &= \lambda u_2 \\ \vdots & \\ \frac{1}{h^2}(u_{N-2} - 2u_{N-1} + u_N) &= \lambda u_{N-1} \\ u_N &= 0 \end{cases},$$

or equivalently

$$A\boldsymbol{u} = \lambda\boldsymbol{u} \tag{6.3}$$

where

$$A = N^2 \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{bmatrix}_{(N-1)\times(N-1)}, \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \end{bmatrix}.$$

The matrix A is in the *Toeplitz form*, that is tridiagonal with the same element along the main diagonal and the sub-diagonals; and it has the eigenvalues:

$$\lambda_k = N^2(-2 + 2\cos\frac{k\pi}{N}) = -4N^2\sin^2\frac{k\pi}{2N}\,, \quad k = 1,2,\cdots,N-1\,.$$

Using the Taylor series expansion of the sine function at zero and at $\pi/2$ we see that they are good approximations for small and large values of k.

An alternative and also extremely popular method is the finite element method. The idea is to discretize the "weak form" of the governing equation, that is if u is a solution to (6.1), then for any continuous function ϕ defined on [0,1] such that $\phi(0) = \phi(1) = 0$, we have:

$$\int_0^1 u''(x)\phi(x)dx = \lambda\int_0^1 u(x)\phi(x)dx\,,$$

or equivalently, by applying the integration-by-part to the left hand side:

$$-\int_0^1 u'(x)\phi'(x)dx = \lambda\int_0^1 u(x)\phi(x)dx\,. \tag{6.4}$$

The idea of the finite element method again uses $N+1$ uniformly distributed points, but approximate both u and ϕ by piecewise linear functions:

$$\begin{aligned} & u(x)\approx u^h(x) = u_i\frac{(i+1)h-x}{h} + u_{i+1}\frac{x-ih}{h}\,, \\ \text{and}\quad & \phi(x)\approx \phi^h(x) = \phi_i\frac{(i+1)h-x}{h} + \phi_{i+1}\frac{x-ih}{h} \\ \text{for}\quad & ih\le x\le (i+1)h\,,\ 0\le i\le N-1\,. \end{aligned}$$

Replacing u and ϕ by u^h and ϕ^h in (6.4), we can carry out both integrals exactly and obtain:

$$\begin{aligned} & -\int_0^1 \frac{du^h(x)}{dx}\frac{d\phi^h(x)}{dx}dx = \lambda\int_0^1 u^h(x)\phi^h(x)dx \\ \Leftrightarrow\quad & -\sum_{i=0}^{N-1}\int_{ih}^{(i+1)h}\frac{du^h(x)}{dx}\frac{d\phi^h(x)}{dx}dx = \lambda\sum_{i=0}^{N-1}\int_{ih}^{(i+1)h} u^h(x)\phi^h(x)dx \\ \Leftrightarrow\quad & -\sum_{i=0}^{N-1} h\cdot\frac{u_{i+1}-u_i}{h}\cdot\frac{\phi_{i+1}-\phi_i}{h} = \lambda\sum_{i=0}^{N-1} h\left(\frac{1}{3}u_i\phi_i+\frac{1}{3}u_{i+1}\phi_{i+1}+\frac{1}{6}u_i\phi_{i+1}+\frac{1}{6}u_{i+1}\phi_i\right) \\ \Leftrightarrow\quad & \boldsymbol{\phi}^t A\boldsymbol{u} = \lambda\boldsymbol{\phi}^t B\boldsymbol{u}\,, \end{aligned}$$

where A happens to be the same as that in the finite difference method, and

$$\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \end{bmatrix},\quad \boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \vdots \\ \phi_{N-1} \end{bmatrix},\quad B = \begin{bmatrix} 2/3 & 1/6 & & & \\ 1/6 & 2/3 & 1/6 & & \\ & 1/6 & 2/3 & \ddots & \\ & & \ddots & \ddots & 1/6 \\ & & & 1/6 & 2/3 \end{bmatrix}.$$

We want $\boldsymbol{\phi}^t A\boldsymbol{u} = \lambda\boldsymbol{\phi}^t B\boldsymbol{u}$ to be hold for any approximations ϕ^h, or any vector $\boldsymbol{\phi}\in\mathbb{R}^{N-1}$. Hence the only solution is given by $A\boldsymbol{u} = \lambda B\boldsymbol{u}$, a generalized eigenvalue problem for A and B. The eigenvalues for this particular problem are difficult to compute; but we will see soon at least that there are $N-1$ real eigenvalues because A is symmetric and B is symmetric and positive-definite.

Now let us look at the problem:

$$A\boldsymbol{x} = \lambda B\boldsymbol{x}\,, \tag{6.5}$$

in general. First of all, if B is invertible, this problem is equivalent to $B^{-1}A\boldsymbol{x} = \lambda\boldsymbol{x}$ and we already know how to solve it.

However, if B is not invertible, there is really very little we can say about the solution to (6.5). Let us consider some extreme cases:

- If $B = 0$, the problem has no solution at all if A is non-singular.

- If $B = A$ and A is non-singular, the problem has exactly one eigenvalue 1 and all vectors are its eigenvector; but if $B = A$ and A is singular, any number is an eigenvalue and its eigenspace contains the null space of A.

So we can no longer claim that the generalized eigenvalue problem has a solution at all, or has at most finite number of eigenvalues.

A special case, however, frequently appears in practical applications is when A is symmetric and B is symmetric and positive definite. In this case, B is non-singular so that we know (6.5) is a well-posed problem. Furthermore, we can find the Cholesky decomposition $B = LL^t$ where L is lower-triangular and non-singular and derive:

$$A\boldsymbol{x} = \lambda B\boldsymbol{x} \quad \Leftrightarrow \quad L^{-1}A(L^t)^{-1}(L^t\boldsymbol{x}) = \lambda L^t\boldsymbol{x}\,.$$

Hence $(\lambda,\boldsymbol{x})$ is an eigenpair of (6.5) if and only if $(\lambda,L^t\boldsymbol{x})$ is an eigenpair of $L^{-1}A(L^{-1})^t$, a symmetric matrix. To this end we know immediately that all the eigenvalues to the generalized problem are real.

Finally, we briefly mention that if B is invertible, almost all previous methods can be modified to solve the corresponding generalized eigenvalue problem without forming B^{-1} and computing $B^{-1}A$ explicitly. For example:

- *The power method.* Instead of computing $\boldsymbol{y}_{i+1} = B^{-1}A\boldsymbol{x}_i$ we solve $B\boldsymbol{y}_{i+1} = A\boldsymbol{x}_i$.
- *The inverse iteration method.* Instead of computing $B^{-1}A\boldsymbol{y}_{i+1} = \boldsymbol{x}_i$ we solve $A\boldsymbol{y}_{i+1} = B\boldsymbol{x}_i$.

Hence there is no significant difference between the two methods now. Despite the straightforward extension from the regular eigenvalue problem to the generalized one, there is an important difference, though. That is, in general we cannot simply use orthogonal matrices to transform A (or B) to a similar upper Hessenberg form (or tridiagonal form in the case of symmetric matrices) to reduce the cost associated with the linear solves.

Indeed, now we want to find "good" matrices Q and Z, and transform A and B according to:

$$A_1 = Q^{-1}AZ\,, \quad B_1 = Q^{-1}BZ\,,$$

then the two problems $A\boldsymbol{x} = \lambda B\boldsymbol{x}$ and $A_1\boldsymbol{x} = \lambda B_1\boldsymbol{x}$ are equivalent to each other. This relates to the *generalized Schur decomposition*, that is:

- If $A,B \in \mathbb{C}^{n\times n}$, we can always find unitary Q and Z such that Q^*AZ and Q^*BZ are upper-triangular.
- If $A,B \in \mathbb{R}^{n\times n}$, we can always find real orthogonal matrices Q and Z such that Q^tAZ is upper Hessenberg and Q^tBZ is upper-triangular. This is known as the *Hessenberg-Triangular form* of the matrix pair (A,B).

Lastly, the *Hessenberg-Triangular form* leads to the so called *QZ method*, the counterpart of the QR method for problem (6.5).

Exercises

Exercise 1. *Show that* $D^{k+1} \to I$ *linearly as* $k \to \infty$ *at the rate* $\max_{i<j} \left| \frac{\lambda_j}{\lambda_i} \right|$ *in the proof of Theorem 3.1.*

Exercise 2. *If* $H \in \mathbb{C}^n$ *is an upper Hessenberg matrix and it is unreduced, then the null space of* $H - \lambda I$ *has at least dimension* $n-1$ *for any* $\lambda \in \mathbb{C}$*. Deduce that any eigenvalue of* H *has geometric multiplicity one.*