Chapter 13
Approximation of Eigenvalues

13.1 General Considerations

The computation of the eigenvalues of a square matrix is a problem of considerable difficulty. The naive idea, according to which it is enough to compute the characteristic polynomial and then find its roots, turns out to be hopeless because of Abel's theorem, which states that the general equation $P(x) = 0$, where P is a polynomial of degree $d \geq 5$, is not solvable using algebraic operations and roots of any order. For this reason, there exists no direct method, even an expensive one, for the computation of $\mathrm{Sp}(M)$.

Dropping half of that program, one could compute the characteristic polynomial exactly, then compute an approximation of its roots. But the cost and the instability of the computation are prohibitive. Amazingly, the opposite strategy is often used: a standard algorithm for computing the roots of a polynomial $P \in \mathbb{C}[X]$ of high degree consists in forming its companion matrix[1] B_p and then applying to this matrix the *QR* algorithm to compute its eigenvalues with good accuracy.

Hence, all the methods are iterative and use the matrices directly. We need a notion of convergence, thus we limit ourselves to the cases $K = \mathbb{R}$ or $\mathbb{C}$. The general strategy consists in constructing a sequence of matrices

$$M^{(0)}, M^{(1)}, \dots, M^{(m)}, \dots,$$

pairwise similar. Each method is conceived in such a way that the sequence converges to a simple form, triangular or diagonal, because then the eigenvalues can be read on the diagonal. Such a convergence is not always possible. For example, an algorithm in $\mathbf{M}_n(\mathbb{R})$ cannot converge to a triangular form when the matrix under consideration possesses a pair of nonreal eigenvalues.

[1] Fortunately, the companion matrix is a Hessenberg matrix; see below for this notion and its practical aspects.

D. Serre, *Matrices,* Graduate Texts in Mathematics 216,
DOI 10.1007/978-1-4419-7683-3_13, © Springer Science+Business Media, LLC 2010

13.1.1 Stability

In the course of the calculations, it is fundamental that the sequence

$$M^{(0)}, M^{(1)}, \dots, M^{(m)}, \dots$$

remain bounded, in order to keep away from overflow, as well as to be allowed to apply Theorem 5.2. This is not guaranteed a priori, because the set of matrices similar to M is unbounded in general. For instance, the following matrices are pairwise similar for all values of $a \in K^*$:

$$\begin{pmatrix} 0 & a \\ a^{-1} & 0 \end{pmatrix}.$$

This boundedness is one important issue among others. When passing from $M^{(k)}$ to $M^{(k+1)}$, the conjugation by a matrix Q yields to an amplification of the roundoff errors by a factor that can be estimated as the *condition number* of Q, namely $\kappa(Q) := \|Q\|_2 \|Q^{-1}\|_2$. We recall that $\kappa(Q) \geq 1$, with equality if and only if Q is the matrix of a similitude. In order to keep control of the roundoff error, it thus seems necessary that the *product* of the numbers $\kappa(Q^{(k))})$ remain bounded. Because this is an infinite product, we need that $\kappa(Q^{(k)}) \to 1$ as $k \to +\infty$. In other words, the distance from $Q^{(k)}$ to $\mathbb{C} \cdot \mathbf{U}_n$ must tend to zero. Notice that a scalar factor in Q is harmless inasmuch as it cancels with the inverse factor in Q^{-1}. For the sake of simplicity, we thus ask that each iteration be a unitary conjugation: each $M^{(k)}$ is unitary similar to M, thus remains bounded because $\mathbf{U}_n$ is compact. When dealing with matrices in $\mathbf{M}_n(\mathbb{R})$, we employ orthogonal conjugation instead.

13.1.2 Expected Convergence

We thus assume that $M^{(k+1)} = Q_k^{-1} M^{(k)} Q_k$ for a unitary Q_k. Set $P_j := Q_0 \cdots Q_{j-1}$, which is unitary too. We have $M^{(j)} = P_j^{-1} M^{(0)} P_j$. Because $\mathbf{U}_n$ is compact, the sequence $(P_j)_{j \in \mathbb{N}}$ possesses cluster values. Let P be one of them. Then $M' := P^{-1} M^{(0)} P = P^* M^{(0)} P$ is a cluster point of $(M^{(j)})_{j \in \mathbb{N}}$ and is conjugated to M. If the sequence $(M^{(j)})_j$ converges, its limit is therefore (unitarily) similar to M, and hence has the same spectrum.

This argument shows that in general, the sequence $(M^{(j)})_j$ does not converge to a diagonal matrix, because then the eigenvectors of M would be the columns of P. In other words, M would have an orthonormal eigenbasis: M would be normal. Except in this special case, one expects merely that the sequence $(M^{(j)})_j$ converges to a triangular matrix, an expectation that is compatible with Theorem 5.1. But even this hope is too optimistic in general.

13.1.3 Initialization

Given $M \in \mathbf{M}_n(\mathbb{C})$, there are two strategies for the choice of $M^{(0)}$. One can naively take $M^{(0)} = M$. But because an iteration on a generic matrix is rather costly, one often uses a preliminary reduction to a simple form (e.g., the Hessenberg form, in the *QR* algorithm), which is preserved throughout the iterations. With a few such tricks, certain methods can be astonishingly efficient.

13.2 Hessenberg Matrices

We recall the notion of *Hessenberg* matrices.

Definition 13.1 *A square matrix $M \in \mathbf{M}_n(K)$ is called* upper Hessenberg *(one speaks simply of a Hessenberg matrix) if $m_{jk} = 0$ for every pair (j,k) such that $j-k \geq 2$.*

A Hessenberg matrix thus has the form

$$\begin{pmatrix} x & \cdots & \cdots & & \\ y & \ddots & & & \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & z & t \end{pmatrix}.$$

In particular, an upper-triangular matrix is a Hessenberg matrix.

From the point of view of matrix reduction by conjugation, one can attribute two advantages to the Hessenberg class, compared with the class of triangular matrices. First of all, if $K = \mathbb{R}$, many matrices are not trigonalizable in $\mathbb{R}$, although all are trigonalizable in $\mathbb{C}$. Even within complex numbers, the trigonalization cannot be done in practice, because it would require the computation of the eigenvalues. On the contrary, we show that every square matrix with real or complex entries is similar to a Hessenberg matrix over the real or complex numbers, respectively. This is obtained after a finite number of operations.

13.2.1 Stability of the Hessenberg Form

If M is Hessenberg and T upper-triangular, the products TM and MT are still Hessenberg.[2] For example, if M admits an LU factorization, then L is Hessenberg, and

[2] But the product of two Hessenberg matrices is not Hessenberg in general.

thus has only two nonzero diagonals, because $L = MU^{-1}$. Likewise, if $M \in \mathbf{GL}_n(\mathbb{C})$ is Hessenberg, then the factor Q in the factorization $M = QR$ is again Hessenberg, because $Q = MR^{-1}$. An elementary compactness and continuity argument shows that the same fact holds true for every $M \in \mathbf{M}_n(\mathbb{C})$.

13.2.2 Hessenberg Form versus Irreducibility

We have seen in Proposition 3.26 that a Hessenberg matrix such that the $m_{j+1,j}$s are nonzero has geometrically simple eigenvalues. The algebraic multiplicity can, however, be arbitrary, as shown in the following example

$$M = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}.$$

13.2.3 Transforming a Matrix into a Hessenberg One

Theorem 13.1 *For every matrix $M \in \mathbf{M}_n(\mathbb{C})$ there exists a unitary transformation U such that $U^{-1}MU$ is a Hessenberg matrix. If $M \in \mathbf{M}_n(\mathbb{R})$, one may take $U \in \mathbf{O}_n$.*

Moreover, the matrix U is computable in $4n^3/3 + O(n^2)$ multiplications and $4n^3/3 + O(n^2)$ additions.

Proof. Let $X \in \mathbb{C}^m$ be a unit vector: $X^*X = 1$. The matrix of the unitary (orthogonal) symmetry with respect to the hyperplane $X^\perp$ is $S = I_m - 2XX^*$. In fact, $SX = X - 2X = -X$, and $Y \in X^\perp$ (i.e., $X^*Y = 0$) implies $SY = Y$.

We construct a sequence $M_1 = M, \dots, M_{n-1}$ of unitarily similar matrices. The matrix M_{n-r} is of the form

$$\begin{pmatrix} H & & B \\ 0_{r,n-r-1} & Z & N \end{pmatrix},$$

where $H \in \mathbf{M}_{n-r}(\mathbb{C})$ is Hessenberg and Z is a vector in $\mathbb{C}^r$. Hence, M_{n-1} is Hessenberg.

One passes from M_{n-r} to M_{n-r+1}, that is, from r to $r-1$ as follows. Let $\mathbf{e}^1$ be the first vector of the canonical basis of $\mathbb{C}^r$. If Z is already colinear to $\mathbf{e}^1$, one does nothing besides defining $M_{n-r+1} = M_{n-r}$. Otherwise, one chooses $X \in \mathbb{C}^r$ so that SZ is parallel to $\mathbf{e}^1$ (we discuss below the possible choices for X). Then one sets

$$V = \begin{pmatrix} I_{n-r} & 0_{n-r,r} \\ 0_{r,n-r} & S \end{pmatrix},$$

which is a unitary matrix, with $V^* = V^{-1} = V$ (such a matrix is called a *Householder* matrix). We then have

$$V^{-1}M_{n-r}V = \begin{pmatrix} H & BS \\ 0_{n,n-r-1} & SZ\ SNS \end{pmatrix}.$$

We thus define $M_{n-r+1} = V^{-1}M_{n-r}V$.

There are two possible choices for S, given by

$$X_{\pm} := \frac{1}{\|Z \pm \|Z\|_2\mathbf{q}\|_2}(Z \pm \|Z\|_2\mathbf{q}), \quad \mathbf{q} = \frac{z_1}{|z_1|}\mathbf{e}^1.$$

It is always advantageous to choose the sign that gives the largest denominator, namely the positive sign. One thus optimizes the roundoff errors in the case where Z is almost aligned with $\mathbf{e}^1$.

13.2.4 Complexity

Let us consider now the complexity of the $(n-r)$th step. Only the terms of order r^2 and $r(n-r)$ are meaningful. The computation of X, in $O(r)$ operations, is thus negligible, like that of X^* and of $2X$. The computation of $BS = B-(BX)(2X^*)$ needs about $4r(n-r)$ operations. Then $2NX$ needs $2r^2$ operations, as does $2X^*N$. We next compute $4X^*NX$, and then form the vector $T := 4(X^*NX)X - 2NX$ at the cost $O(r)$. The product TX^* takes r^2 operations, as $2X(X^*N)$. Then $N + TX^* - X(2X^*N)$ needs $2r^2$ additions. The complete step is thus accomplished in $2r^2 + 4rn + O(n)$ operations. A sum from $r = 1$ to $n-2$ yields a complexity of $\frac{8}{3}n^3 + O(n^2)$, in which one recognizes $\frac{4}{3}n^3 + O(n^2)$ multiplications, $\frac{4}{3}n^3 + O(n^2)$ additions, and $O(n)$ square roots. □

13.2.5 The Hermitian Case

When M is Hermitian, the matrix $U^{-1}MU$ is still Hermitian. Because it is Hessenberg, it is tridiagonal, with $a_{j,j+1} = \bar{a}_{j+1,j}$ and $a_{jj} \in \mathbb{R}$. The symmetry reduces the complexity to $2n^3/3 + O(n^2)$ multiplications. One can then use the Hessenberg form of M in order to localize its eigenvalues.

Proposition 13.1 *If M is tridiagonal Hermitian and if the entries $m_{j+1,j}$ are nonzero (i.e., if M is irreducible), then the eigenvalues of M are real and simple. Furthermore, if M_j is the (Hermitian, tridiagonal, irreducible) matrix obtained by keeping only the j last rows and columns of M, the eigenvalues of M_j strictly separate those of M_{j+1}.*

The separation, not necessarily strict, of the eigenvalues of M_{j+1} by those of M_j has already been proved, in a more general framework, in Theorem 6.5.

Proof. The geometric simplicity of the eigenvalues has been stated in Proposition 3.26. Because M is Hermitian, it is diagonalizable: geometric multiplicity equals the algebraic one. Thus the eigenvalues are simple. In addition, an Hermitian matrix has a real spectrum.

We proceed by induction on j. If $j \geq 1$, we decompose the matrix M_{j+1} blockwise:

$$\begin{pmatrix} m & \bar{a} & 0 & \cdots & 0 \\ a & & & & \\ 0 & & M_j & & \\ \vdots & & & & \\ 0 & & & & \end{pmatrix},$$

where $a \neq 0$ and $m \in \mathbb{R}$. Let P_ℓ be the characteristic polynomial of M_ℓ. We compute that of M_{j+1} by expanding the determinant with respect to the first column:

$$P_{j+1}(X) = (X-m)P_j(X) - |a|^2 P_{j-1}(X), \tag{13.1}$$

where $P_0 \equiv 1$ by convention.

The induction hypothesis is as follows. The polynomials P_j and P_{j-1} have real coefficients and have, respectively, j and $j-1$ real roots $\mu_1, \dots, \mu_j$ and $\sigma_1, \dots, \sigma_{j-1}$, with

$$\mu_1 < \sigma_1 < \mu_2 < \cdots < \sigma_{j-1} < \mu_j.$$

In particular, they have no other roots, and their roots are simple. The signs of the values of P_{j-1} at points μ_j thus alternate. Because P_{j-1} is positive over $(\sigma_{j-1}, +\infty)$, we have $(-1)^{j-k} P_{j-1}(\mu_k) > 0$.

This hypothesis clearly holds at step $j = 1$. If $j \geq 2$ and if it holds at step j, then (13.1) shows that $P_{j+1} \in \mathbb{R}[X]$. Furthermore,

$$(-1)^{j-k} P_{j+1}(\mu_k) = -|a|^2 (-1)^{j-k} P_{j-1}(\mu_k) < 0.$$

From the intermediate value theorem, P_{j+1} possesses a root λ_k in (μ_{k-1}, μ_k). Furthermore, $P_{j+1}(\mu_j) < 0$, and $P_{j+1}(x)$ is positive for $x \gg 1$; hence there is another root in $(\mu_j, +\infty)$. Likewise, P_{j+1} has a root in $(-\infty, \mu_1)$. Hence, P_{j+1} possesses $j+1$ distinct real roots λ_k, with

$$\lambda_1 < \mu_1 < \lambda_2 < \cdots < \mu_j < \lambda_{j+1}.$$

Because P_{j+1} has degree $j+1$, it has no root other than the λ_ks, and these are simple.
□

The sequence of polynomials P_j is a *Sturm sequence*, which allows us to compute the number of roots of P_n in a given interval (a, b). A Sturm sequence is a finite sequence of real polynomials $Q_0, \dots, Q_n$, with Q_0 a nonzero constant such that

- If $Q_j(x) = 0$ and $0 < j < n$, then $Q_{j+1}(x) Q_{j-1}(x) < 0$. In particular, Q_j and Q_{j+1} do not share a common root.

- Likewise, if $Q_0(c) = 0$ for some $c \in (a,b)$, then

$$\frac{Q_0(x)Q_1(x)}{x-c} < 0, \qquad \forall x \in (c-\varepsilon, c+\varepsilon)$$

for some $\varepsilon > 0$.

If $a \in \mathbb{R}$ is not a root of Q_n, we denote by $V(a)$ the number of sign changes in the sequence $(Q_0(a), \dots, Q_n(a))$, in which the zeroes play no role and can be ignored.

Proposition 13.2 *If $Q_n(a) \neq 0$ and $Q_n(b) \neq 0$, and if $a < b$, then the number of roots of Q_n in (a,b) is equal to $V(a) - V(b)$.*

Let us remark that it is not necessary to compute the polynomials P_j in order to apply them to this proposition. Given $a \in \mathbb{R}$, it is enough to compute the sequence of values $P_j(a)$.

Once an interval (a,b) is known to contain an eigenvalue λ and only that one (by means of Proposition 13.2 or Theorem 5.7), one can compute an approximate value of λ, either by dichotomy, or by computing the numbers $V((a+b)/2), \dots$, or by the secant or Newton method. In the latter case, one must compute P_n itself. The last two methods are convergent, provided that we have a good initial approximation at our disposal, because $P_n'(\lambda) \neq 0$.

13.3 The *QR* Method

The *QR* method is considered the most efficient one for the approximate computation of the whole spectrum of a general square matrix $M \in \mathbf{M}_n(\mathbb{C})$. One employs it only after having reduced M to Hessenberg form, because this form is preserved throughout the algorithm, whereas each iteration is much cheaper than it would be for an arbitrary matrix.

13.3.1 Description of the QR *Method*

Let $A \in \mathbf{M}_n(K)$ be given, with $K = \mathbb{R}$ or $\mathbb{C}$. We construct a sequence of matrices $(A_j)_{j\in\mathbb{N}}$, with $A_0 = A$. The induction $A_j \mapsto A_{j+1}$ consists in performing the *QR* factorization of A_j, $A_j = Q_jR_j$, and then defining $A_{j+1} := R_jQ_j$. We have

$$A_{j+1} = Q_j^{-1}A_jQ_j,$$

which shows that A_{j+1} is unitarily similar to A_j. Hence,

$$A_j = (Q_0 \cdots Q_{j-1})^{-1}A(Q_0 \cdots Q_{j-1}) \tag{13.2}$$

is conjugate to A by a unitary transformation.

13.3.1.1 Obstructions

- If A is unitary, then $A_j = A$ for every j, with $Q_j = A$ and $R_j = I_n$. The convergence occurs but is useless, because the limit A is not simpler than the data. We show later on that the reason for this bad behavior is that the eigenvalues of a unitary matrix have the same modulus. The QR method does not do a good job of separating the eigenvalues of close modulus.
- Another bad situation is when our matrix has at least two eigenvalues of the same modulus. This happens in particular if A has real entries. In the latter case, then each Q_j is real orthogonal, R_j is real, and A_j is real. This is seen by induction on j. A limit A' is not triangular if some eigenvalues of A are nonreal, namely if A possesses a pair of complex conjugate eigenvalues.

Let us sum up what can be expected in a brave new world. If all the eigenvalues of $A \in \mathbf{M}_n(\mathbb{C})$ have distinct moduli, the sequence $(A_j)_j$ might converge to a triangular matrix, or at least its lower-triangular part might converge to

$$\begin{pmatrix} \lambda_1 & & & \\ 0 & \lambda_2 & & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}.$$

When $A \in \mathbf{M}_n(\mathbb{R})$, let us make the following assumption. Let p be the number of real eigenvalues and $2q$ that of nonreal eigenvalues ; then there are $p+q$ distinct eigenvalue moduli. In that case, $(A_j)_j$ might converge to a block-triangular form, the diagonal blocks being 2×2 or 1×1. The limits of the diagonal blocks trivially provide the eigenvalues of A.

Herebelow, we treat the complex case with eigenvalues of pairwise distinct moduli. The case with real entries and pairs of complex conjugate eigenvalues has been treated in [23].

13.3.2 The Case of a Singular Matrix

When A is not invertible, the QR factorization is not unique, raising a difficulty in the definition of the algorithm. The computation of the determinant would immediately detect the case of noninvertibility, but would not provide any cure. However, if the matrix has been first reduced to the Hessenberg form, then a single QR iteration makes a diagnosis and does provide a cure. Indeed, if A is Hessenberg and singular, then in $A = QR$, Q is Hessenberg and R is singular. If $a_{21} = 0$, the matrix A is block-triangular and we may reduce our calculations to the case of a matrix of size $(n-1)\times(n-1)$ by deleting the first row and the first column. Otherwise, there exists $j \geq 2$ such that $r_{jj} = 0$. The matrix $A_1 = RQ$ is then block-triangular, because it is

Hessenberg and $(A_1)_{j,j-1} = r_{jj}q_{j,j-1} = 0$. Again, we may reduce our calculations to that of the spectra of two matrices of sizes $j \times j$ and $(n-j) \times (n-j)$, the diagonal blocks of A_1. After finitely many such steps (not larger than the multiplicity of the null eigenvalue), there remain only Hessenberg invertible matrices to deal with. We assume therefore from now on that $A \in \mathbf{GL}_n(K)$.

13.3.3 Complexity of an Iteration

An iteration of the *QR* method requires the factorization $A_j = Q_jR_j$ and the computation of $A_{j+1} = R_jQ_j$. Each part costs $O(n^3)$ operations if it is done on a generic matrix (using the naive way of multiplying matrices). The reduction to the Hessenberg form has a comparable cost, therefore we loose nothing by reducing A to this form. Actually, we make considerable gains in two aspects. First of all, the cost of each *QR* iteration is reduced to $O(n^2)$. Secondly, the cluster values of the sequence $(A_j)_j$ must have the Hessenberg form too.

Let us first examine the Householder method of *QR* factorization for a generic matrix A. In practice, one computes only the factor R and matrices of unitary symmetries whose product is Q. One then multiplies these unitary matrices by R on the left to obtain $A' = RQ$.

Let $\mathbf{a}_1 \in \mathbb{C}^n$ be the first column vector of A. We begin by determining a unit vector $v_1 \in \mathbb{C}^n$ such that the hyperplane symmetry $H_1 := I_n - 2v_1v_1^*$ sends $\mathbf{a}_1$ to $\|\mathbf{a}_1\|_2\mathbf{e}^1$. The matrix H_1A has the form

$$\tilde{A} = \begin{pmatrix} \|\mathbf{a}_1\|_2 & x \cdots \\ 0 & \vdots \\ \vdots & \vdots \\ 0 & y \cdots \end{pmatrix}.$$

We then perform these operations again on the matrix extracted from $\tilde{A}$ by deleting the first rows and columns, and so on. At the kth step, H_k is a matrix of the form

$$\begin{pmatrix} I_k & 0 \\ 0 & I_{n-k} - 2v_kv_k^* \end{pmatrix},$$

where $v_k \in \mathbb{C}^{n-k}$ is a unit vector. The computation of v_k requires $O(n-k)$ operations. The product $H_kA^{(k)}$, where $A^{(k)}$ is block-triangular, amounts to that of two square matrices of size $n-k$, one of them $I_{n-k} - 2v_kv_k^*$. We thus compute a matrix $N - 2vv^*N$ from v and N, which costs about $4(n-k)^2$ operations. Summing from $k = 1$ to $k = n-1$, we find that the complexity of the computation of R alone is $4n^3/3 + O(n^2)$. As indicated above, we do not compute the factor Q, but compute all the matrices $RH_{n-1}\cdots H_k$. That necessitates $2n^3 + O(n)$ operations. The complexity of one step of the *QR* method on a generic matrix is thus $10n^3/3 + O(n^2)$.

Let us now analyze the situation when A is a Hessenberg matrix. By induction on k, we see that v_k belongs to the plane spanned by $\mathbf{e}^k$ and $\mathbf{e}^{k+1}$. Its computation needs $O(1)$ operations. Then the product of H_k and $A^{(k)}$ can be obtained by simply recomputing the rows of indices k and $k+1$, about $6(n-k)$ operations. Summing from $k=1$ to $n-1$, we find that the complexity of the computation of R alone is $3n^2+O(n)$. The computation of the product $(RH_{n-1}\cdots H_{k+1})H_k$ needs about $6k$ operations. Finally, the complexity of the QR iteration on a Hessenberg matrix is $6n^2+O(n)$, in which there are $4n^2+O(n)$ multiplications.

To sum up, the cost of the preliminary reduction of a matrix to Hessenberg form is less than or equal to what is saved during the first iteration of the QR method.

13.3.4 Convergence of the QR *Method*

As explained above, the best convergence statement assumes that the eigenvalues have distinct moduli.

Let us recall that the sequence A_k is not always convergent. For example, if A is already triangular, its QR factorization is $Q=D$, $R=D^{-1}A$, with $d_j = a_{jj}/|a_{jj}|$. Hence, $A_1 = D^{-1}AD$ is triangular, with the same diagonal as that of A. By induction, A_k is triangular, with the same diagonal as that of A. We have thus $Q_k = D$ for every k, so that $A_k = D^{-k}AD^k$. The entry of index (ℓ,m) is thus multiplied at each step by a unit number $z_{\ell m}$, which is not necessarily equal to one if $\ell < m$. Hence, the part above the diagonal of A_k may not converge.

Summing up, a convergence theorem may concern only the diagonal of A_k and what lies below it.

Lemma 24. *Let $A \in \mathbf{GL}_n(K)$ be given, with $K=\mathbb{R}$ or $\mathbb{C}$. Let $A_k = Q_kR_k$ be the sequence of matrices given by the* QR *algorithm. Let us define $P_k = Q_0\cdots Q_{k-1}$ and $U_k = R_{k-1}\cdots R_0$. Then P_kU_k is the* QR *factorization of the kth power of A:*

$$A^k = P_kU_k.$$

Proof. From (13.2), we have $A_k = P_k^{-1}AP_k$; that is, $P_kA_k = AP_k$. Then

$$P_{k+1}U_{k+1} = P_kQ_kR_kU_k = P_kA_kU_k = AP_kU_k.$$

By induction, $P_kU_k = A^k$. However, $P_k \in \mathbf{U}_n$ and U_k is triangular, with a positive real diagonal, as a product of such matrices. □

Theorem 13.2 *Let $A \in \mathbf{GL}_n(\mathbb{C})$ be given. Assume that the moduli of the eigenvalues of A are distinct:*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| \qquad (>0).$$

In particular, the eigenvalues are simple, and thus A is diagonalizable:

$$A = Y^{-1}\mathrm{diag}(\lambda_1,\ldots,\lambda_n)Y.$$

Assume also that Y admits an LU factorization. Then the strictly lower-triangular part of A_k converges to zero, and the diagonal of A_k converges to

$$D := \operatorname{diag}(\lambda_1,\dots,\lambda_n).$$

Proof. Let $Y = LU$ be the factorization of Y. We also make use of the QR factorization of Y^{-1} : $Y^{-1} = QR$. Because $A^k = Y^{-1}D^kY$, we have $P_kU_k = Y^{-1}D^kY = QRD^kLU$.

The matrix D^kLD^{-k} is lower-triangular with unit numbers on its diagonal. Each term is multiplied by $(\lambda_i/\lambda_j)^k$, therefore its strictly lower-triangular part tends to zero, because $|\lambda_i/\lambda_j| < 1$ for $i > j$. Therefore, $D^kLD^{-k} = I_n + E_k$ with $E_k \to 0_n$ as $k \to +\infty$. Hence, $P_kU_k = QR(I_n + E_k)D^kU = Q(I_n + RE_kR^{-1})RD^kU = Q(I_n + F_k)RD^kU$, where $F_k \to 0_n$. Let $O_kT_k = I_n + F_k$ be the QR factorization of $I_n + F_k$. By continuity, O_k and T_k both tend to I_n. Then

$$P_kU_k = (QO_k)(T_kRD^kU).$$

The first product is a unitary matrix, whereas the second is a triangular one. Let $|D|$ be the "modulus" matrix of D (whose entries are the moduli of those of D), and let D_1 be $|D|^{-1}D$, which is unitary. We also define $D_2 = \operatorname{diag}(u_{jj}/|u_{jj}|)$ and $U' = D_2^{-1}U$. Then D_2 is unitary and the diagonal of U' is positive real. From the uniqueness of the QR factorization of an invertible matrix we obtain

$$P_k = QO_kD_1^kD_2, \quad U_k = (D_1^kD_2)^{-1}T_kRD_1^kD_2|D|^kU',$$

which yields

$$\begin{aligned} Q_k &= P_k^{-1}P_{k+1} = D_2^{-1}D_1^{-k}O_k^{-1}O_{k+1}D_1^{k+1}D_2,\\ R_k &= U_{k+1}U_k^{-1} = D_2^{-1}D_1^{-k-1}T_{k+1}RDR^{-1}T_k^{-1}D_1^kD_2. \end{aligned}$$

Because D_1^{-k} and D_1^{k+1} are bounded, we deduce that Q_k converges, to D_1. Likewise, $R_k - R'_k \to 0_n$, where

$$R'_k = D_2^{-1}D_1^{-k}RDR^{-1}D_1^{k-1}D_2. \tag{13.3}$$

The fact that the matrix R'_k is upper-triangular shows that the strictly lower-triangular part of $A_k = Q_kR_k$ tends to zero (observe that the sequence $(R_k)_{k\in\mathbb{N}}$ is bounded, because the set of matrices unitarily conjugate to A is bounded). Likewise, the diagonal of R'_k is $|D|$, which shows that the diagonal of A_k converges to $D_1|D| = D$. □

Remark

Formula (13.3) shows that the sequence A_k does not converge, at least when the eigenvalues have distinct complex arguments. However, if the eigenvalues have equal complex arguments, for example, if they are real and positive, then $D_1 = \alpha I_n$

and $R_k \to T := D_2^{-1} R|D|R^{-1} D_2$; hence A_k converges. Note that the limit αT is not diagonal in this case.

The odd assumption about Y (LU factorization) in Theorem 13.2 is fulfilled in most practical situations:

Theorem 13.3 *Let $A \in \mathbf{GL}_n(\mathbb{C})$ be an irreducible Hessenberg matrix whose eigenvalues are of distinct moduli:*

$$|\lambda_1| > \cdots > |\lambda_n| \quad (>0).$$

Then the QR *method converges; that is, the lower-triangular part of A_k converges to*

$$\begin{pmatrix} \lambda_1 & & & \\ 0 & \lambda_2 & & \\ \vdots & \ddots & \ddots & \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}.$$

Proof. In the light of Theorem 13.2, it is enough to show that the matrix Y in the previous proof admits an LU factorization. We have $YA = \operatorname{diag}(\lambda_1, \dots, \lambda_n)Y$. The rows of Y are thus the left eigenvectors: $\ell_j A = \lambda_j \ell_j$.

If $x \in \mathbb{C}^n$ is nonzero, there exists a unique index r such that $x_r \neq 0$, and $j > r$ implies $x_j = 0$. By induction, quoting the Hessenberg form and the irreducibility of A, we obtain $(A^m x)_{r+m} \neq 0$, while $j > r+m$ implies $(A^m x)_j = 0$. Hence, the vectors $x, Ax, \dots, A^{n-r}x$ are linearly independent. A linear subspace, invariant for A and containing x, is thus of dimension greater than or equal to $n-r+1$.

Let F be a linear subspace, invariant for A, of dimension $p \geq 1$. Let r be the smallest integer such that F contains a nonzero vector x with $x_{r+1} = \cdots = x_n = 0$. The minimality of r implies that $x_r \neq 0$. Hence, we have $p \geq n-r+1$. By construction, the intersection of F and of linear subspace $[\mathbf{e}^1, \dots, \mathbf{e}^{r-1}]$ spanned by $\mathbf{e}^1, \dots, \mathbf{e}^{r-1}$ reduces to $\{0\}$. Thus we also have $p + (r-1) \leq n$. Finally, $r = n-p+1$, and we see that

$$F \oplus [\mathbf{e}^1, \dots, \mathbf{e}^{n-p}] = \mathbb{C}^n.$$

Let us choose $F = [\ell_1, \dots, \ell_q]^\perp$, which is invariant for A. Then $p = n-q$, and we have

$$[\ell_1, \dots, \ell_q]^\perp \oplus [\mathbf{e}^1, \dots, \mathbf{e}^q] = \mathbb{C}^n.$$

This amounts to saying that $\det(\ell_j \mathbf{e}^k)_{1 \leq j,k \leq q} \neq 0$. In other words, the leading principal minor of order q of Y is nonzero. By Theorem 11.1, Y admits an LU factorization. □

13.3.5 The Case of Hermitian Matrices

The situation is especially favorable for tridiagonal Hermitian matrices. To begin with, we may assume that A is positive-definite, up to the change of A into $A + \mu I_n$

with $\mu > -\rho(A)$. Next, we can write A in block-diagonal form, where the diagonal blocks are tridiagonal irreducible Hermitian matrices. The *QR* method then treats each block separately. We are thus reduced to the case of an Hermitian positive-definite, tridiagonal, and irreducible matrix. Its eigenvalues are real, strictly positive, and simple, from Proposition 13.1: we have $\lambda_1 > \cdots > \lambda_n > 0$. Theorems 13.2 and 13.3 can then be applied.

Corollary 13.1 *If $A \in$ **HPD**$_n$ and if A_0 is a Hessenberg matrix, unitarily similar to A (e.g., a matrix obtained by Householder's method), then the sequence A_k defined by the* QR *method converges to a diagonal matrix whose diagonal entries are the eigenvalues of A.*

Indeed, the lower-triangular part converges, hence the whole matrix, because it is Hermitian.

13.3.6 Implementing the QR *Method*

The *QR* method converges faster as λ_n, or merely λ_n/λ_{n-1}, becomes smaller. We can obtain this situation by translating $A_k \mapsto A_k - \alpha_k I_n$. The strategies for the choice of α_k are described in [27]. This procedure is called *Rayleigh translation*. It yields a significant improvement of the convergence of the *QR* method. If the eigenvalues of A are simple, a suitable translation places us into the case of eigenvalues of distinct moduli. This trick has a nonnegligible cost if A is a real matrix with a pair of complex conjugate eigenvalues, inasmuch as it requires a translation by a nonreal number α. As mentioned above, the computations become much more costly in $\mathbb{C}$ than they are in $\mathbb{R}$.

As k increases, the triangular form of A_k shows up first at the last row. As a by-product, the sequence $(A_k)_{nn}$ converges more rapidly than other sequences $(A_k)_{jj}$. When the last row is sufficiently close to $(0,\dots,0,\lambda_n)$, the Rayleigh translation must be selected in such a way as to bring λ_{n-1}, instead of λ_n, to the origin; and so on.

With a clever choice of Rayleigh translations, the *QR* method, when it converges, is of order two for a generic matrix, and is of order three for an Hermitian matrix.

13.4 The Jacobi Method

The Jacobi method gives an approximate value of the whole spectrum of a real symmetric matrix $A \in$ **Sym**$_n$. As in the *QR* method, one constructs a sequence of matrices, unitarily similar to A. In particular, the roundoff errors are not amplified. Each iteration is cheap ($O(n)$ operations), and the convergence may be quadratic or even faster when the eigenvalues are distinct. It is thus a rather efficient method.

13.4.1 Conjugating by a Rotation Matrix

Let $1 \le p,q \le n$ be two distinct indices and $\theta \in [-\pi,\pi)$ an angle. We denote by $R_{p,q}(\theta)$ the matrix of rotation of angle θ in the plane spanned by $\mathbf{e}^p$ and $\mathbf{e}^q$. For example, if $p<q$, then

$$R = R_{p,q}(\theta) := \begin{pmatrix} I_{p-1} & \vdots & 0 & \vdots & 0 \\ \cdots & \cos\theta & \cdots & \sin\theta & \cdots \\ 0 & \vdots & I_{q-p-1} & \vdots & 0 \\ \cdots & -\sin\theta & \cdots & \cos\theta & \cdots \\ 0 & \vdots & 0 & \vdots & I_{n-q} \end{pmatrix}.$$

If H is a symmetric matrix, we compute $K := R^{-1}HR = R^THR$, which is also symmetric, with the same spectrum. Setting $c=\cos\theta$, $s=\sin\theta$ the following formulæ hold.

$$\begin{aligned} k_{ij} &= h_{ij} && \text{if } i,j \neq p,q, \\ k_{ip} &= ch_{ip} - sh_{iq} && \text{if } i \neq p,q, \\ k_{iq} &= ch_{iq} + sh_{ip} && \text{if } i \neq p,q, \\ k_{pp} &= c^2h_{pp} + s^2h_{qq} - 2csh_{pq}, \\ k_{qq} &= c^2h_{qq} + s^2h_{pp} + 2csh_{pq}, \\ k_{pq} &= cs(h_{pp}-h_{qq}) + (c^2-s^2)h_{pq}. \end{aligned}$$

The cost of the computation of entries k_{ij} for $i,j \neq p,q$ is zero; that of k_{pp}, k_{qq}, and k_{pq} is $O(1)$. The cost of this conjugation is thus $6n+O(1)$ operations, keeping in mind the symmetry $K^T=K$.

Let us remark that the conjugation by the rotation of angle $\theta \pm \pi$ yields the same matrix K, up to signs. For this reason, we limit ourselves to angles $\theta \in [-\pi/2, \pi/2)$.

13.4.2 Description of the Method

One constructs a sequence $A^{(0)} = A, A^{(1)}, \dots$ of symmetric matrices, each one conjugate to the previous one by a rotation as above: $A^{(k+1)} = (R^{(k)})^T A^{(k)} R^{(k)}$. At step k, we choose two distinct indices p and q (in fact, p_k, q_k) in such a way that $a_{pq}^{(k)} \neq 0$ (if it is not possible, $A^{(k)}$ is already a diagonal matrix similar to A). We then choose θ (in fact θ_k) in such a way that $a_{pq}^{(k+1)} = 0$. From the formulæ above, this is equivalent to

$$cs(a_{pp}^{(k)} - a_{qq}^{(k)}) + (c^2-s^2)a_{pq}^{(k)} = 0.$$

This amounts to solving the equation

$$\cot 2\theta = \frac{a_{qq}^{(k)} - a_{pp}^{(k)}}{2a_{pq}^{(k)}} =: \sigma_k. \tag{13.4}$$

This equation possesses two solutions in $[-\pi/2, \pi/2)$, namely $\theta_k \in [-\pi/4, \pi/4)$ and $\theta_k \pm \pi/2$. There are thus two possible rotation matrices, which yield to two distinct results. Once the angle has been selected, its computation is useless (it would actually be rather expensive). In fact, $t := \tan\theta_k$ solves

$$\frac{2t}{1-t^2} = \tan 2\theta;$$

that is,

$$t^2 + 2t\sigma_k - 1 = 0.$$

The two angles correspond to the two possible roots of this quadratic equation. We then obtain

$$c = \frac{1}{\sqrt{1+t^2}}, \quad s = tc.$$

We show below that the stablest choice is the angle $\theta_k \in [-\pi/4, \pi/4)$, which corresponds to the unique root t in $[-1, 1)$.

The computation of c, s needs only $O(1)$ operations, so that the cost of an iteration of the Jacobi method is still $6n + O(1)$. Observe that an entry that has vanished at a previous iteration becomes in general nonzero after a few more iterations.

13.4.3 The Choice of the Pair $(\mathbf{p}, \mathbf{q})$

We use here the Schur norm $\|M\| = (\mathrm{Tr} M^T M)^{1/2}$, also called the Frobenius norm, denoted elsewhere by $\|M\|_F$. We wish to show that $A^{(k)}$ converges to a diagonal matrix, therefore we decompose $A^{(k)} = D_k + E_k$, where $D_k = \mathrm{diag}(a_{11}^{(k)}, \dots, a_{nn}^{(k)})$. To begin with, because the sequence is formed of unitarily similar matrices, we have $\|A^{(k)}\| = \|A\|$.

Lemma 25. *We have*

$$\|E_{k+1}\|^2 = \|E_k\|^2 - 2\left(a_{pq}^{(k)}\right)^2.$$

Proof. It suffices to redo the calculations of Section 13.4.1, noting that

$$k_{ip}^2 + k_{iq}^2 = h_{ip}^2 + h_{iq}^2$$

whenever $i \neq p, q$, whereas $k_{pq}^2 = 0$. □

We deduce from the lemma that $\|D_{k+1}\|^2 = \|D_k\|^2 + 2\left(a_{pq}^{(k)}\right)^2$. The convergence of the Jacobi method then depends on the choice of the pair (p, q) at each step. Notice that the choice of the same pair at two consecutive iterations is inadvisable, inasmuch as it yields $A^{(k+1)} = A^{(k)}$.

There are essentially three strategies for chosing the pair (p, q) at a given step.

Optimal choice. One chooses a pair (p,q) for which the modulus of a_{pq} is maximal among off-diagonal entries of $A^{(k)}$. At first glance, this looks to be the most efficient choice, but needs a comparison procedure whose cost is about $n^2 \log n$. If a careful storage of the order of moduli at previous steps is made, the comparison reduces to about n^2 operations, still costly enough, compared to the $6n$ operations needed in the conjugation.

Sequential choice. Here the pair is a periodic function of k. Typically, one chooses first $(1,2)$ then $(2,3), \dots, (n-1,n), (1,3), (2,4), \dots, (1,n)$. Variant: because the position $(2,3)$ was affected by the operations made around $(1,2)$, it might be better to find an order beginning with $(1,2), (3,4), \dots$, in such a way that an index p is not present in two consecutive pairs, in order to treat all the entries as fast as possible.

Random choice. The set of pairs (p,q) with $1 \le p < q \le n$ is equipped with the uniform probability. The pair (p,q) is taken at random at step k, and independently of the previous choices. Some variants of the random choice can be elaborated.

13.4.4 Convergence with the Optimal Choice

Theorem 13.4 *With the "optimal choice" of (p_k,q_k) and with the choice $\theta_k \in [-\pi/4, \pi/4)$, the Jacobi method converges in the following sense. There exists a diagonal matrix D such that*

$$\|A^{(k)} - D\| \le \frac{\sqrt{2}\|E_0\|}{1-\rho}\rho^k, \qquad \rho := \sqrt{1 - \frac{2}{n^2-n}}.$$

In particular, the spectrum of A consists of the diagonal terms of D and the limit of D_k; the Jacobi method is of order one at least.

This kind of convergence is called *linear*, because it is typical of methods in which the error obeys a linear inequality $\varepsilon_{k+1} \le \rho\varepsilon_k$, with $\rho < 1$. We also say that the convergence is of order one at least. This is a rather slow convergence that we already encountered in iterative methods for linear systems (Chapter 12).

Proof. With the optimal choice of (p,q), we have

$$(n^2 - n)\left(a_{pq}^{(k)}\right)^2 \ge \|E_k\|^2.$$

Hence,

$$\|E_{k+1}\|^2 \le \left(1 - \frac{2}{n^2-n}\right)\|E_k\|^2 = \rho^2\|E_k\|^2.$$

It follows that $\|E_k\| \le \rho^k\|E_0\|$. In particular, E_k tends to zero as $k \to +\infty$.

It remains to show that D_k converges too. A calculation using the notation of Section 13.4.1 and the fact that $k_{pq} = 0$ yields

$$k_{pp} - h_{pp} = -th_{pq}.$$

Because $|\theta_k| \leq \pi/4$, we have $|t| \leq 1$, so that $|a_{pp}^{(k+1)} - a_{pp}^{(k)}| \leq |a_{pq}^{(k)}|$. Likewise, $|a_{qq}^{(k+1)} - a_{qq}^{(k)}| \leq |a_{pq}^{(k)}|$. The other diagonal entries are unchanged, thus we have $\|D_{k+1} - D_k\| \leq \|E_k\|$.

We therefore have

$$\|D_\ell - D_k\| \leq \|E_0\|(\rho^{\ell-1} + \cdots + \rho^k) \leq \|E_0\| \frac{\rho^k}{1-\rho}, \quad \ell > k.$$

The sequence $(D_k)_{k\in\mathbb{N}}$ is thus Cauchy, hence convergent. Because E_k tends to zero, $A^{(k)}$ converges to the same limit D. This matrix is diagonal, with the same spectrum as A, because this is true for each $A^{(k)}$. Finally, we obtain

$$\|A^{(k)} - D\|^2 = \|D_k - D\|^2 + \|E_k\|^2 \leq \frac{2}{(1-\rho)^2} \|E_k\|^2.$$

□

We analyze, in Exercise 10, the (bad) behavior of D_k when we make the opposite choice $\pi/4 \leq |\theta_k| \leq \pi/2$.

13.4.5 Optimal Choice: Super-Linear Convergence

The following statement shows that the Jacobi method compares rather well with other methods.

Theorem 13.5 *The Jacobi method with optimal choice of (p,q) converges super-linearly when the eigenvalues of A are simple, in the following sense. Let $N = n(n-1)/2$ be the number of elements under the diagonal. Then there exists a number $c > 0$ such that*

$$\|E_{k+N}\| \leq c\|E_k\|^2,$$

for every $k \in \mathbb{N}$.

In the present setting, the order of the Jacobi method can be estimated at least as $\nu := 2^{1/N}$, which is slightly larger than 1. This is typical of a method where the error obeys an inequality of the form $\varepsilon_{k+1} < \text{cst} \cdot (\varepsilon_k)^\nu$ (mind, however, that the inequality given in the theorem is not exactly of this form). The convergence is much faster than a linear one. We expect in practice that the order be even larger than ν. For instance, Exercise 15 gives the order $(1+\sqrt{5})/2$ when $n = 3$. The exact order for a general n is still unknown.

Proof. We first remark that if $i \neq j$ with $\{i,j\} \neq \{p_\ell, q_\ell\}$, then

$$|a_{ij}^{(\ell+1)} - a_{ij}^{(\ell)}| \leq |t_\ell|\sqrt{2}\|E_\ell\|, \tag{13.5}$$

where $t_\ell = \tan\theta_\ell$. To see this, observe that $1 - c \leq t$ and $|s| \leq t$ whenever $|t| \leq 1$. However, Theorem 13.4 ensures that D_k converges to $\mathrm{diag}(\lambda_1, \dots, \lambda_n)$, where the λ_js are the eigenvalues of A. Because these are distinct, there exist $K \in \mathbb{N}$ and $\delta > 0$ such that, if $k \geq K$, then

$$\min_{i \neq j} |a_{ii}^{(k)} - a_{jj}^{(k)}| \geq \delta$$

for $k \geq K$. We have therefore

$$|\sigma_k| \geq \frac{\delta}{\sqrt{2}\|E_k\|} \overset{k \to +\infty}{\longrightarrow} +\infty.$$

It follows that t_k tends to zero and, more precisely, that

$$t_k \approx -\frac{1}{2\sigma_k}.$$

Finally, there exists a constant c_1 such that

$$|t_k| \leq c_1 \|E_k\|.$$

Let us then fix k larger than K, and let us denote by J the set of pairs (p_ℓ, q_ℓ) when $k \leq \ell \leq k + N - 1$. For such an index, we have $\|E_\ell\| \leq \rho^{\ell-k}\|E_k\| \leq \|E_k\|$. In particular, $|t_\ell| \leq c_1\|E_k\|$.

If $(p,q) \in J$ and if $\ell < k + N$ is the largest index such that $(p,q) = (p_\ell, q_\ell)$, a repeated application of (13.5) shows that

$$|a_{pq}^{(k+N)}| \leq c_1 N\sqrt{2}\|E_k\|^2.$$

If J is equal to the whole set of pairs (i,j) such that $i < j$, these inequalities ensure that $\|E_{k+N}\| \leq c_2\|E_k\|^2$. Otherwise, there exists a pair (p,q) that one sets to zero twice: $(p,q) = (p_\ell, q_\ell) = (p_m, q_m)$ with $k \leq \ell < m < k + N$. In that case, the same argument as above shows that

$$\|E_{k+N}\| \leq \|E_m\| \leq \sqrt{2N}|a_{pq}^{(m)}| \leq 2\sqrt{N}c_1(m - \ell)\|E_k\|^2.$$

□

Remarks

We show in Exercise 13 that when the eigenvalues of A are simple, the distance between the diagonal and the spectrum of A is $O(\|E_k\|^2)$, and not $O(\|E_k\|)$ as expected from Theorem 5.7.

13.4.6 Convergence with the Random Choice

Recall that we choose the pair (p,q) independently of those chosen at previous steps, according to the uniform distribution. The matrix $A^{(k)}$ is therefore a function of $A^{(k)}$ and of the random variable (p_k,q_k); as such, it is a random variable.

We are interested in the *expectation* of the norm of the error $\|E_{k+1}\|^2$. To begin with, we consider the *conditional* expectation, knowing $\|E_k\|^2$. We have

$$\mathrm{e}\left[\|E_{k+1}\|^2 \,|\, \|E_k\|^2\right] = \frac{2}{n^2-n}\sum_{1\le p<q\le n}\|E_{k+1}(p,q)\|^2.$$

Because of Lemma 25, we obtain

$$\begin{aligned}\mathrm{e}\left[\|E_{k+1}\|^2 \,|\, \|E_k\|^2\right] &= \frac{2}{n^2-n}\sum_{1\le p<q\le n}\left(\|E_k\|^2-|a_{pq}^{(k)}|^2\right)\\ &= \|E_k\|^2-\frac{2}{n^2-n}\sum_{1\le p<q\le n}|a_{pq}^{(k)}|^2\\ &= \left(1-\frac{2}{n^2-n}\right)\|E_k\|^2=\rho^2\|E_k\|^2.\end{aligned}$$

Taking now the expectation with respect to the previous choices, we obtain

$$\mathrm{e}\left[\|E_{k+1}\|^2\right]=\rho^2\mathrm{e}\left[\|E_k\|^2\right].$$

By induction, this yields

$$\mathrm{e}\left[\|E_k\|^2\right]=\rho^{2k}\mathrm{e}\left[\|E_0\|^2\right]. \tag{13.6}$$

Let β be a number given in the interval $(\rho,1)$. Let us denote $c_0:=\mathrm{e}\left[\|E_0\|^2\right]$. Then the probability that $\|E_k\|$ is larger than β^k is less than $c_0(\rho/\beta)^{2k}$, according to (13.6). We therefore have

$$\sum_{k=0}^{\infty}\mathbb{P}\left(\|E_k\|>\beta^k\right)<\infty. \tag{13.7}$$

Thanks to the theorem of Borel–Cantelli, this implies that for almost every choice of the sequence $(p_k,q_k)_{k\in\mathbb{N}}$, the inequality $\|E_k\|\le\beta^k$ is true for all but finitely many indices k ; in other words, $\|E_k\|\le\beta^k$ is true for large enough k. When this happens, we may apply the same analysis of the diagonal part D_k as that made in Section 13.4.4. Finally, we have the following theorem.

Theorem 13.6 *Consider the Jacobi method with random choice, the pairs (p_k,q_k) being independent and chosen according to the uniform distribution.*

For every $\varepsilon>0$, the error $\|E_k\|$ decays almost surely as an $O\left((\rho+\varepsilon)^k\right)$, with

$$\rho := \sqrt{1 - \frac{2}{n^2 - n}}.$$

Provided the angle θ_k is chosen in the interval $(-\pi/4, \pi/4]$, the diagonal converges as soon as the error tends to zero, and the diagonal entries of its limit are the eigenvalues of A.

13.5 The Power Methods

The power methods are designed for the approximation of a single eigenvalue. Consequently, their cost is significantly lower than that of the QR or the Jacobi methods. The standard power method is used in particular when searching for the optimal parameter in the SOR method for a tridiagonal matrix, where we have to compute the spectral radius of the Jacobi iteration matrix (Theorem 12.2).

13.5.1 The Standard Method

Let $M \in \mathbf{M}_n(\mathbb{C})$ be a matrix. We search for an approximation of its eigenvalue of maximum modulus, whenever only one such exists. The standard method consists in choosing a norm on $\mathbb{C}^n$, a unit vector $x^0 \in \mathbb{C}^n$, and then successively computing the vectors x^k by the formula

$$x^{k+1} := \frac{1}{\|Mx^k\|} Mx^k.$$

The justification of this method is given in the following theorem.

Theorem 13.7 *One assumes that* $\mathrm{Sp}M$ *contains only one element λ of maximal modulus (that modulus is thus equal to $\rho(M)$).*

If $\rho(M) = 0$, the method stops because $Mx^k = 0$ for some $k < n$.

Otherwise, let $\mathbb{C}^n = E \oplus F$ be the decomposition of $\mathbb{C}^n$, where E, F are invariant linear subspaces under M, with $\mathrm{Sp}(M|_E) = \{\lambda\}$ *and* $\lambda \notin \mathrm{Sp}(M|_F)$. *Assume that $x^0 \notin F$. Then $Mx^k \neq 0$ for every $k \in \mathbb{N}$ and*

$$\lim_{k \to +\infty} \|Mx^k\| = \rho(M). \tag{13.8}$$

In addition,

$$V := \lim_{k \to +\infty} \left(\frac{\bar{\lambda}}{\rho(M)} \right)^k x^k$$

is a unit eigenvector of M, associated with the eigenvalue λ. If $V_j \neq 0$, then

$$\lim_{k\to+\infty} \frac{(Mx^k)_j}{x_j^k} = \lambda .$$

Proof. The case $\rho(M) = 0$ is obvious because M is then nilpotent.

Assume otherwise that $\rho(M) > 0$. Let $x^0 = y^0 + z^0$ be the decomposition of x^0 with $y^0 \in E$ and $z^0 \in F$. By assumption, $y^0 \neq 0$. Because $M|_E$ is invertible, $M^k y^0 \neq 0$. Because $M^k x^0 = M^k y^0 + M^k z^0$, $M^k y^0 \in E$, and $M^k z^0 \in F$, we have $M^k x^0 \neq 0$. The algorithm may be rewritten as[3]

$$x^k = \frac{1}{\|M^k x^0\|} M^k x^0 .$$

We therefore have $x^k \neq 0$.

If $F \neq \{0\}$, then $\rho(M|_F) < \rho(M)$ by construction. Hence there exist (from Theorem 7.1) $\eta < \rho(M)$ and $C > 0$ such that $\|(M|_F)^k\| \leq C\eta^k$ for every k. Then $\|(M|_F)^k z^0\| \leq C_1 \eta^k$. On the other hand, $\rho((M|_E)^{-1}) = 1/\rho(M)$, and the same argument as above ensures that $\|(M|_E)^{-k}\| \leq 1/C_2\mu^k$, for some $\mu \in (\eta, \rho(M))$, so that $\|M^k y^0\| \geq C_3 \mu^k$. Hence,

$$\|M^k z^0\| \ll \|M^k y^0\|,$$

so that

$$x^k \approx \frac{1}{\|M^k y^0\|} M^k y^0 .$$

We are thus led to the analysis of the case where $z^0 = 0$, namely when M has no eigenvalue but λ. That is assumed from now on.

Let r be the degree of the minimal polynomial of M. The vector space spanned by the vectors $x^0, Mx^0, \ldots, M^{r-1}x^0$ contains all the x^ks. Up to the replacement of $\mathbb{C}^n$ by this linear subspace, we may assume that it equals $\mathbb{C}^n$. Then we have $r = n$. Furthermore, because $\ker(M-\lambda)^{n-1}$, a nontrivial linear subspace, is invariant under M, we see that $x^0 \notin \ker(M-\lambda)^{n-1}$.

The vector space $\mathbb{C}^n$ then admits the basis

$$\{v^1 = x^0, v^2 = (M-\lambda)x^0, \ldots, v^n = (M-\lambda)^{n-1}x^0\}.$$

With respect to this basis, M becomes the Jordan matrix

$$\tilde{M} = \begin{pmatrix} \lambda & 0 & \ldots & \ldots & \\ 1 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ & \ldots & 0 & 1 & \lambda \end{pmatrix}.$$

[3] One could normalize x^k at the end of the computation, but we prefer doing it at each step in order to avoid overflows, and also to ensure (13.8).

The matrix $\lambda^{-k}\tilde{M}^k$ depends polynomially on k. The coefficient of highest degree, as $k \to +\infty$, is at the intersection of the first column and the last row. It equals

$$\binom{k}{n-1}\lambda^{1-n},$$

which is equivalent to $\dfrac{(k/\lambda)^{n-1}}{(n-1)!}$. We deduce that

$$M^k x^0 \sim \frac{k^{n-1}\lambda^{k-n+1}}{(n-1)!}v^n.$$

Hence,

$$x^k \sim \left(\frac{\lambda}{|\lambda|}\right)^{k-n+1}\frac{v^n}{\|v^n\|}.$$

Because v^n is an eigenvector of M, the claims of the theorem have been proved. □

The case where the algebraic and geometric multiplicities of λ are equal (i.e., $M|_E = \lambda I_E$), for example, if λ is a simple eigenvalue, is especially favorable. Indeed, $M^k y^0 = \lambda^k y^0$, and therefore

$$x^k = \left(\frac{\lambda}{|\lambda|}\right)^k \frac{1}{\|y^0\|}y^0 + O\left(\frac{\|M^k z^0\|}{|\lambda|^k}\right).$$

Theorem 7.1 thus shows that the error

$$x^k - \left(\frac{\lambda}{|\lambda|}\right)^k \frac{1}{\|y^0\|}y^0$$

tends to zero faster than

$$\left(\frac{\rho(M|_F)+\varepsilon}{\rho(M)}\right)^k,$$

for every $\varepsilon > 0$. The convergence is thus of order one, and becomes faster as the ratio $|\lambda_2|/|\lambda_1|$ becomes smaller (arranging the eigenvalues by nonincreasing moduli). However, the convergence is much slower when the Jordan blocks of M relative to λ are nontrivial. The error then behaves like $1/k$ in general.

The situation is more delicate when $\rho(M)$ is the modulus of several distinct eigenvalues. The vector x^k, suitably normalized, does not converge in general but "spins" closer and closer to the sum of the corresponding eigenspaces. The observation of the asymptotic behavior of x^k allows us to identify the eigendirections associated with the eigenvalues of maximal modulus. The sequence $\|Mx^k\|$ does not converge and depends strongly on the choice of the norm. However, $\log\|Mx^k\|$ converges in the Cesaro sense, that is, in the mean, to $\log\rho(M)$ (Exercise 12).

Remark

The hypothesis on x_0 is generic, in the sense that it is satisfied for every choice of x_0 in an open dense subset of $\mathbb{C}^n$. If by chance x^0 belongs to F, the power method theoretically furnishes another eigenvalue, of smaller modulus. In practice, a large enough number of iterations always leads to the convergence to λ. In fact, the number λ is rarely exactly representable in a computer. When it is not, the linear subspace F does not contain any nonzero representable vector. Thus the vector x^0, or its computer representation, does not belong to F, and Theorem 13.7 applies.

13.5.2 The Inverse Power Method

Let us assume that M is invertible. The standard power method, applied to M^{-1}, furnishes the eigenvalue of least modulus, whenever it is unique, or at least produces its modulus in the general case. The inversion of a matrix is a costly operation, therefore we involve ourselves with that idea only if M has already been inverted, for example if we had previously had to make an *LU* or a *QR* factorization. That is typically the situation when one begins to implement the *QR* algorithm for M. It might look strange to involve a method giving only one eigenvalue in the course of a method that is expected to compute the whole spectrum.

The inverse power method is thus subtle. Here is how it works. One begins by implementing the *QR* method until one gets coarse approximations $\mu_1,\dots,\mu_n$ of the eigenvalues $\lambda_1,\dots,\lambda_n$. If one persists in the *QR* method, the proof of Theorem 13.2 shows that the error is at best of order σ^k with $\sigma = \max_j |\lambda_{j+1}/\lambda_j|$. When n is large, σ is in general close to 1 and this convergence is rather slow. Likewise, the method with Rayleigh translations, for which σ is replaced by $\sigma(\eta) := \max_j |(\lambda_{j+1}-\eta)/(\lambda_j-\eta)|$, is not satisfactory. However, if one wishes to compute a *single* eigenvalue, say λ_p, with full accuracy, the power method, applied to $M-\mu_p I_n$, produces an error on the order of θ^k, where $\theta := |\lambda_p-\mu_p|/\min_{j\neq p}|\lambda_j-\mu_p|$ is a small number, since $\lambda_p-\mu_p$ is small.

In practice, the inverse power method is used mainly to compute an approximate eigenvector, associated with an eigenvalue for which one already had a good approximate value.

Exercises

1. Given a polynomial $P \in \mathbb{R}[X]$, use Euclidean division in order to define a sequence of nonzero polynomials P_j in the following way. Set $P_0 = P$, $P_1 = P'$. If P_j is not constant, $-P_{j+1}$ is the remainder of the division of P_{j-1} by P_j: $P_{j-1} = Q_j P_j - P_{j+1}$, $\deg P_{j+1} < \deg P_j$.

a. Assume that P has only simple roots. Show that the sequence $(P_j)_j$ is well defined, that it has only finitely many terms, and that it is a Sturm sequence.
b. Use Proposition 13.2 to compute the number of real roots of the real polynomials $X^2 + aX + b$ or $X^3 + pX + q$ in terms of their discriminants.

2. (Wilkinson [40], Section 5.45.) Let $n = 2p - 1$ be an odd number and $W_n \in \mathbf{M}_n(\mathbb{R})$ be the symmetric tridiagonal matrix

$$\begin{pmatrix} p & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & 1 & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & p \end{pmatrix}.$$

The diagonal entries are thus $p, p-1, \dots, 2, 1, 2, \dots, p-1, p$, and the subdiagonal entries are equal to 1.

a. Show that the linear subspace

$$E' = \{X \in \mathbb{R}^n \,|\, x_{p+j} = x_{p-j}, 1 \le j < p\}$$

is invariant under W_n. Likewise, show that the linear subspace

$$E'' = \{X \in \mathbb{R}^n \,|\, x_{p+j} = -x_{p-j}, 0 \le j < p\}$$

is stable under W_n.

b. Deduce that the spectrum of W_n is the union of the spectra of the matrices

$$W_n' = \begin{pmatrix} p & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 2 & 1 \\ & & & 2 & 1 \end{pmatrix}, \quad W_n'' = \begin{pmatrix} p & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 3 & 1 \\ & & & 1 & 2 \end{pmatrix}$$

(we have $W_n' \in \mathbf{M}_p(\mathbb{R})$ and $W_n'' \in \mathbf{M}_{p-1}(\mathbb{R})$).

c. Show that the eigenvalues of W_n'' strictly separate those of W_n'.

3. For $a_1, \dots, a_n \in \mathbb{R}$, with $\sum_j a_j = 1$, form the matrix

$$M(a) := \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & & a_n \\ a_2 & b_2 & a_3 & \vdots & \vdots & \vdots \\ a_3 & a_3 & b_3 & & \vdots & \vdots \\ a_4 & \cdots & & & & \vdots \\ & \cdots & \cdots & & & a_n \\ a_n & \cdots & \cdots & \cdots & a_n & b_n \end{pmatrix},$$

where $b_j := a_1 + \cdots + a_{j-1} - (j-2)a_j$.

a. Compute the eigenvalues and the eigenvectors of $M(a)$.
b. We limit ourselves to n-uplets a that belong to the simplex S defined by $0 \leq a_n \leq \cdots \leq a_1$ and $\sum_j a_j = 1$. Show that for $a \in S$, $M(a)$ is bistochastic and $b_2 - a_2 \leq \cdots \leq b_n - a_n \leq 1$.
c. Let $\mu_1, \ldots, \mu_n$ be an n-uplet of elements in $[0,1]$ with $\mu_n = 1$. Show that there exists a unique a in S such that $\{\mu_1, \ldots, \mu_n\}$ is equal to the spectrum of $M(a)$ (counting with multiplicity).

4. Show that the cost of an iteration of the QR method for an Hermitian tridiagonal matrix is $20n + O(1)$.
5. Show that the reduction to the Hessenberg form (in this case, tridiagonal form) of an Hermitian matrix costs $7n^3/6 + O(n^2)$ operations.
6. (Invariants of the algorithm QR.) For $M \in \mathbf{M}_n(\mathbb{R})$ and $1 \leq k \leq n-1$, let us denote by $(M)_k$ the matrix of size $(n-k) \times (n-k)$ obtained by deleting the first k rows and the last k columns. For example, $(I)_1$ is the Jordan matrix $J(0; n-1)$. We also denote by $K \in \mathbf{M}_n(\mathbb{R})$ the matrix defined by $k_{1n} = 1$ and $k_{ij} = 0$ otherwise.

a. For an upper-triangular matrix T, explicitly compute KT and TK.
b. Let $M \in \mathbf{M}_n(\mathbb{R})$. Prove the equality

$$\det(M - \lambda I - \mu K) = (-1)^n \mu \det(M - \lambda I)_1 + \det(M - \lambda I).$$

c. Let $A \in \mathbf{GL}_n(\mathbb{R})$ be given, with factorization $A = QR$. Prove that

$$\det(A - \lambda I)_1 = \frac{\det R}{r_{nn}} \det(Q - \lambda R^{-1})_1.$$

d. Let $A' := RQ$. Show that

$$r_{nn} \det(A' - \lambda I)_1 = r_{11} \det(A - \lambda I)_1.$$

e. Generalize the previous calculation by replacing the index 1 by k. Deduce that the roots of the polynomial $\det(A - \lambda I)_k$ are conserved throughout the QR algorithm. How many such roots do we have for a general matrix? How many for a Hessenberg matrix?

7. (Invariants; continuing.) For $M \in \mathbf{M}_n(\mathbb{R})$, let us define $P_M(h;z) := \det((1-h)M + hM^T - zI_n)$.

   a. Show that $P_M(h;z) = P_M(1-h;z)$. Deduce that there exists a polynomial Q_M such that $P_M(h;z) = Q_M(h(1-h);z)$.
   b. Show that Q_M remains constant throughout the QR algorithm: if $Q \in \mathbf{O}_n(\mathbb{R})$, R is upper-triangular, and $M = QR$, $N = RQ$, then $Q_M = Q_N$.
   c. Deduce that there exist polynomial functions J_{rk} on $\mathbf{M}_n(\mathbb{R})$, defined by

   $$P_M(h;z) = \sum_{r=0}^{n} \sum_{k=0}^{[r/2]} (h(1-h))^k z^{n-r} J_{rk}(M),$$

   that are invariant throughout the QR algorithm. Verify that the J_{r0}s can be expressed in terms of invariants that we already know.
   d. Compute explicitly J_{21} when $n = 2$. Deduce that in the case where Theorem 13.2 applies and $\det A > 0$, the matrix A_k converges.
   e. Show that for $n \geq 2$,

   $$J_{21}(M) = -\frac{1}{2} \operatorname{Tr}\left((M - M^T)^2\right).$$

   Deduce that if A_k converges to a diagonal matrix, then A is symmetric.

8. In the Jacobi method, show that if the eigenvalues are simple, then the product $R^1 \cdots R^m$ converges to an orthogonal matrix R such that $R^* AR$ is diagonal.
9. Extend the Jacobi method to Hermitian matrices. **Hint:** Replace the rotation matrices

   $$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

   by unitary matrices

   $$\begin{pmatrix} z_1 & z_2 \\ z_3 & z_4 \end{pmatrix}.$$

10. Let $A \in \mathbf{Sym}_n(\mathbb{R})$ be a matrix whose eigenvalues, of course real, are simple. Apply the Jacobi method, but selecting the angle θ_k so that $\pi/4 \leq |\theta_k| \leq \pi/2$.

    a. Show that E_k tends to zero, that the sequence D_k is relatively compact, and that its cluster values are diagonal matrices whose diagonal terms are the eigenvalues of A.
    b. Show that an iteration has the effect of permuting, asymptotically, $a_{pp}^{(k)}$ and $a_{qq}^{(k)}$, where $(p,q) = (p_k, q_k)$. In other words

    $$\lim_{k \to +\infty} |a_{pp}^{(k+1)} - a_{qq}^{(k)}| = 0,$$

    and vice versa, permuting p and q.

11. The Bernoulli method computes an approximation of the root of largest modulus for a polynomial $a_0X^n + \cdots + a_n$, when that root is unique. To do so, one defines a sequence by a linear induction of order n:

$$z_k = -\frac{1}{a_0}(a_1 z_{k-1} + \cdots + a_n z_{k-n}).$$

Compare this method with the power method for a suitable matrix.

12. Consider the power method for a matrix $M \in \mathbf{M}_n(\mathbb{C})$ of which several eigenvalues are of modulus $\rho(M) \neq 0$. Again, $\mathbb{C}^n = E \oplus F$ is the decomposition of $\mathbb{C}^n$ into linear subspaces stable under M, such that $\rho(M|_F) < \rho(M)$ and the eigenvalues of $M|_E$ are of modulus $\rho(M)$. Finally, $x^0 = y^0 + z^0$ with $y^0 \in E$, $z^0 \in F$, and $y^0 \neq 0$.

    a. Express
    $$\frac{1}{m}\sum_{k=0}^{m-1} \log \|Mx^k\|$$
    in terms of $\|M^m x^0\|$.

    b. Show that if $0 < \mu < \rho(M) < \eta$, then there exist constants C, C' such that
    $$C\mu^k \leq \|M^k x^0\| \leq C'\eta^k, \quad \forall k \in \mathbb{N}.$$

    c. Deduce that $\log \|Mx^k\|$ converges in the mean to $\log \rho(M)$.

13. Let $M \in \mathbf{M}_n(\mathbb{C})$ be given. Assume that the Gershgorin disk D_ℓ is disjoint from the other disks D_m, $m \neq \ell$. Show that the inverse power method, applied to $M - m_{\ell\ell}I_n$, provides an approximate computation of the unique eigenvalue of M that belongs to D_ℓ.

14. The ground field is $\mathbb{R}$.

    a. Let P and Q be two monic polynomials of respective degrees n and $n-1$ ($n \geq 2$). We assume that P has n real and distinct roots, strictly separated by the $n-1$ real and distinct roots of Q. Show that there exist two real numbers d and c, and a monic polynomial R of degree $n-2$, such that
    $$P(X) = (X-d)Q(X) - c^2 R(X).$$

    b. Let P be a monic polynomial of degree n ($n \geq 2$). We assume that P has n real and distinct roots. Build sequences $(d_j, P_j)_{1 \leq j \leq n}$ and $(c_j)_{1 \leq j \leq n-1}$, where d_j, c_j are real numbers and P_j is a monic polynomial of degree j, with
    $$P_n = P, \quad P_j(X) = (X - d_j)P_{j-1}(X) - c_{j-1}^2 P_{j-2}(X), \quad (2 \leq j \leq n).$$
    Deduce that there exists a tridiagonal matrix A, which we can obtain by algebraic calculations (involving square roots), whose characteristic polynomial is P.

c. Let P be a monic polynomial. We assume that P has n real roots. Prove that one can factorize $P = Q_1 \cdots Q_r$, where each Q_j has simple roots, and the factorization requires only finitely many operations. Deduce that there is a finite algorithm, involving no more than square roots calculations, which provides a tridiagonal symmetric matrix A, whose characteristic polynomial is P (a *tridiagonal symmetric companion matrix*).

15. We apply the Jacobi method to a real 3×3 matrix A. Our strategy is one that we have called "optimal choice".

a. Let (p_1,q_1), (p_2,q_2), …, (p_k,q_k), … be the sequence of index pairs that are chosen at consecutive steps (recall that one vanishes the off-diagonal entry of largest modulus). Prove that this sequence is cyclic of order three: it is either the sequence

$$\dots,(1,2),(2,3),(3,1),(1,2),\dots,$$

or

$$\dots,(1,3),(3,2),(2,1),(1,3),\dots.$$

b. Assume now that A has simple eigenvalues. At each step, one of the three off-diagonal entries is null, and the two other ones are small, because the method converges. Say that they are $0, x_k, y_k$ with $0 < |x_k| \le |y_k|$ (if x_k vanishes then one diagonal entry is an eigenvalue and the method ends one step further). Show that $y_{k+1} \sim x_k$ and $x_{k+1} \sim 2x_k y_k/\delta$, where δ is a gap between two eigenvalues. Deduce that the method is of order $\omega = (1+\sqrt{5})/2$, the golden ratio, meaning that the error ε_k at step k satisfies

$$\varepsilon_{k+1} = O(\varepsilon_k \varepsilon_{k-1}).$$

c. Among the class of Hessenberg matrices, we distinguish the *unit* ones, which have 1s below the diagonal:

$$M = \begin{pmatrix} * & \cdots & & \cdots & * \\ 1 & \ddots & & & \vdots \\ 0 & \ddots & & & \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & * \end{pmatrix}.$$

i. Let $M \in \mathbf{M}_n(k)$ be a unit Hessenberg matrix. We denote by M_k the submatrix obtained by retaining the first k rows and columns. For instance, $M_n = M$ and $M_1 = (m_{11})$. We set P_k the characteristic polynomial of M_k. Show that

$$P_n(X) = (X - m_{nn})P_{n-1}(X) - m_{n-1,n}P_{n-2}(X) - \cdots - m_{2n}P_1(X) - m_{1n}.$$

ii. Let $Q_1,\dots,Q_n \in k[X]$ be monic polynomials, with $\deg Q_k = k$. Show that there exists one and only one unit Hessenberg matrix M such that, for every $k = 1,\dots,n$, the characteristic polynomial of M_k equals Q_k. **Hint:** Argue by induction over n.

Note: The roots of the polynomials $P_1,\dots,P_n$ are called the *Ritz values* of M.