4 Iterative Methods for Eigenvalue Problems

4.1 Methods for the partial eigenvalue problem

In this section, we discuss iterative methods for solving the *partial* eigenvalue problem of a general matrix $A \in \mathbb{K}^{n\times n}$.

4.1.1 The "Power Method"

Definition 4.1: *The "Power method" of v. Mises*[1] *generates, starting from some initial point* $z^0 \in \mathbb{C}^n$ *with* $\|z^0\| = 1$, *a sequence of iterates* $z^t \in \mathbb{C}^n$, $t = 1, 2, \ldots$, *by*

$$\tilde{z}^t = Az^{t-1}, \quad z^t := \|\tilde{z}^t\|^{-1}\tilde{z}^t. \tag{4.1.1}$$

The corresponding eigenvalue approximation is given by

$$\lambda^t := \frac{(Az^t)_r}{z_r^t}, \quad r \in \{1, \ldots, n\} : \ |z_r^t| = \max_{j=1,\ldots,n} |z_j^t|. \tag{4.1.2}$$

The normalization is commonly done using the norms $\|\cdot\| = \|\cdot\|_\infty$ or $\|\cdot\| = \|\cdot\|_2$. For the convergence anlysis of this method, we assume the matrix A to be diagonalizable, i. e., to be similar to a diagonal matrix, which is equivalent to the existence of a basis of eigenvectors $\{w^1, \ldots, w^n\}$ of A. These eigenvectors are associated to the eigenvalues ordered according to their modulus, $0 \le |\lambda_1| \le \ldots \le |\lambda_n|$, and are assumed to be normalized, $\|w^i\|_2 = 1$. Further, we assume that the initial vector z^0 has a nontrivial component with respect to the n-th eigenvector w^n,

$$z^0 = \sum_{i=1}^{n} \alpha_i w^i, \quad \alpha_n \neq 0. \tag{4.1.3}$$

In practice, this is not really a restrictive assumption since, due to round-off errors, it will be satisfied in general.

Theorem 4.1 (Power method): *Let the matrix* A *be diagonalizable and let the eigenvalue with largest modulus be separated from the other eigenvalues, i. e.,* $|\lambda_n| > |\lambda_i|$, $i = 1, \ldots, n-1$. *Further, let the starting vector* z^0 *have a nontrivial component with respect to the eigenvector* w^n. *Then, there are numbers* $\sigma_t \in \mathbb{C}$, $|\sigma_t| = 1$ *such that*

$$\|z^t - \sigma_t w^n\| \to 0 \quad (t \to \infty), \tag{4.1.4}$$

[1] *Richard von Mises (1883–1953): Austrian mathematician; Prof. of applied Mathematics in Straßburg (1909-1918), in Dresden and then founder of the new Institute of Applied Mathematics in Berlin (1919-1933), emigration to Turkey (Istambul) and eventually to the USA (1938); Prof. at Harward University; important contributions to Theoretical Fluid Mechanics (introduction of the "stress tensor"), Aerodynamics, Numerics, Statistics and Probability Theory.*

and the "maximum" eigenvalue $\lambda_{\max} = \lambda_n$ *is approximated with convergence speed*

$$\lambda^t = \lambda_{\max} + \mathcal{O}\Big(\Big|\frac{\lambda_{n-1}}{\lambda_{\max}}\Big|^t\Big) \quad (t \to \infty). \tag{4.1.5}$$

Proof. Let $z^0 = \sum_{i=1}^n \alpha_i w^i$ be the basis expansion of the starting vector. For the iterates z^t there holds

$$z^t = \frac{\tilde{z}^t}{\|\tilde{z}^t\|_2} = \frac{Az^{t-1}}{\|Az^{t-1}\|_2} = \frac{A\tilde{z}^{t-1}}{\|\tilde{z}^{t-1}\|_2}\,\frac{\|\tilde{z}^{t-1}\|_2}{\|A\tilde{z}^{t-1}\|_2} = \ldots = \frac{A^t z^0}{\|A^t z^0\|_2}.$$

Furthermore,

$$A^t z^0 = \sum_{i=1}^n \alpha_i \lambda_i^t w^i = \lambda_n^t \alpha_n \Big\{ w^n + \sum_{i=1}^{n-1} \frac{\alpha_i}{\alpha_n} \Big(\frac{\lambda_i}{\lambda_n}\Big)^t w^i \Big\}$$

and consequently, since $|\lambda_i/\lambda_n| < 1$, $i = 1, \ldots, n-1$,

$$A^t z^0 = \lambda_n^t \alpha_n \{ w^n + o(1) \} \quad (t \to \infty).$$

This implies

$$z^t = \frac{\lambda_n^t \alpha_n \{w^n + o(1)\}}{|\lambda_n^t \alpha_n|\,\|w^n + o(1)\|_2} = \underbrace{\frac{\lambda_n^t \alpha_n}{|\lambda_n^t \alpha_n|}}_{=:\,\sigma_t} w^n + o(1).$$

The iterates z^t converges to $\operatorname{span}\{w^n\}$. Further, since $\alpha_n \neq 0$, it follows that

$$\begin{aligned}\lambda^t &= \frac{(Az^t)_k}{z_k^t} = \frac{(A^{t+1}\,z^0)_k}{\|A^t z^0\|_2}\,\frac{\|A^t z^0\|_2}{(A^t z^0)_k} \\ &= \frac{\lambda_n^{t+1}\big\{\alpha_n w_k^n + \sum_{i=1}^{n-1} \alpha_i \big(\frac{\lambda_i}{\lambda_n}\big)^{t+1} w_k^i\big\}}{\lambda_n^t \big\{\alpha_n w_k^n + \sum_{i=1}^{n-1} \alpha_i \big(\frac{\lambda_i}{\lambda_n}\big)^t w_k^i\big\}} = \lambda_n + \mathcal{O}\Big(\Big|\frac{\lambda_{n-1}}{\lambda_n}\Big|^t\Big) \quad (t \to \infty).\end{aligned}$$

This completes the proof. Q.E.D.

For Hermitian matrices, one obtains improved eigenvalue approximations using the "Rayleigh quotient":

$$\lambda^t := (Az^t, z^t)_2\,, \quad \|z^t\|_2 = 1. \tag{4.1.6}$$

In this case $\{w_1, \ldots, w_n\}$ can be chosen as ONB of eigenvectors such that there holds

$$\begin{aligned}\lambda^t &= \frac{(A^{t+1}z^0,\, A^t z^0)}{\|A^t z^0\|^2} = \frac{\sum_{i=1}^n |\alpha_i|^2 \lambda_i^{2t+1}}{\sum_{i=1}^n |\alpha_i|^2 \lambda_i^{2t}} \\ &= \frac{\lambda_n^{2t+1}\,\big\{|\alpha_n|^2 + \sum_{i=1}^{n-1} |\alpha_i|^2 \big(\frac{\lambda_i}{\lambda_n}\big)^{2t+1}\big\}}{\lambda_n^{2t}\,\Big\{|\alpha_n|^2 + \sum_{i=1}^{n-1} |\alpha_i|^2 \big(\frac{\lambda_i}{\lambda_n}\big)^{2t}\Big\}} = \lambda_{\max} + \mathcal{O}\Big(\Big|\frac{\lambda_{n-1}}{\lambda_{\max}}\Big|^{2t}\Big).\end{aligned}$$

Here, the convergence of the eigenvalue approximations is twice as fast as in the non-Hermitian case.

Remark 4.1: The convergence of the power method is the better the more the modulus-wise largest eigenvalue λ_n is separated from the other eigenvalues. The proof of convergence can be extended to the case of diagonalizable matrices with multiple "maximum" eigenvalue for which $|\lambda_n| = |\lambda_i|$ necessarily implies $\lambda_n = \lambda_i$. For even more general, non-diagonalizable matrices convergence is not guaranteed. The proof of Theorem 4.1 suggests that the constant in the convergence estimate (4.1.5) depends on the dimension n and may therefore be very large for large matrices. The proof that this is actually not the case is posed as an exercise.

4.1.2 The "Inverse Iteration"

For practical computation the power method is of only limited value, as its convergence is very slow in general if $|\lambda_{n-1}/\lambda_n| \sim 1$. Further, it only delivers the "largest" eigenvalue. In most practical applications the "smallest" eigenvalue is wanted, i. e., that which is closest to zero. This is accomplished by the so-called "Inverse Iteration" of Wielandt[2]. Here, it is assumed that one already knows a good approximation $\tilde{\lambda}$ for an eigenvalue λ_k of the matrix A to be computed (obtained by other methods, e. g., Lemma of Gershgorin, etc.) such that

$$|\lambda_k - \tilde{\lambda}| \ll |\lambda_i - \tilde{\lambda}|, \quad i = 1, \ldots, n\,, \quad i \neq k. \tag{4.1.7}$$

In case $\tilde{\lambda} \neq \lambda_k$ the matrix $(A - \tilde{\lambda} I)^{-1}$ has the eigenvalues $\mu_i = (\lambda_i - \tilde{\lambda})^{-1}$, $i = 1, \ldots, n$, and there holds

$$|\mu_k| = \Big|\frac{1}{\lambda_k - \tilde{\lambda}}\Big| \gg \Big|\frac{1}{\lambda_i - \tilde{\lambda}}\Big| = |\mu_i|, \quad i = 1, \ldots, n\,, \quad i \neq k. \tag{4.1.8}$$

Definition 4.2: *The "Inverse Iteration" consists in the application of the power method to the matrix $(A - \tilde{\lambda} I)^{-1}$, where the so-called "shift" $\tilde{\lambda}$ is taken as an approximation to the desired eigenvalue λ_k. Starting from an initial point z^0 the method generates iterates z^t as solutions of the linear systems*

$$(A - \tilde{\lambda} I)\tilde{z}^t = z^{t-1}\,, \quad z^t = \|\tilde{z}^t\|^{-1}\tilde{z}^t, \quad t = 1, 2, \ldots. \tag{4.1.9}$$

The corresponding eigenvalue approximation is determined by

$$\mu^t := \frac{[(A - \tilde{\lambda} I)^{-1} z^t]_r}{z_r^t}, \quad r \in \{1, \ldots, n\} : \ |z_r^t| = \max_{j=1,\ldots,n} |z_j^t|, \tag{4.1.10}$$

or, in the Hermitian case, by the Rayleigh quotient

$$\mu^t := ((A - \tilde{\lambda} I)^{-1} z^t, z^t)_2. \tag{4.1.11}$$

[2]Helmut Wielandt (1910–2001): German mathematician; Prof. in Mainz (1946-1951) and Tübingen (1951-1977); contributions to Group Theory, Linear Algebra and Matrix Theory.

In the evaluation of the eigenvalue approximation in (4.1.10) and (4.1.11) the not yet known vector $\tilde{z}^{t+1} := (A - \tilde{\lambda}I)^{-1}z^t$ is needed. Its computation requires to carry the iteration, possibly unnecessarily, one step further by solving the corresponding linear system $(A - \tilde{\lambda}I)\tilde{z}^{t+1} = z^t$. This can be avoided by using the formulas

$$\lambda^t := \frac{(Az^t)_r}{z_r^t}, \quad \text{or in the symmetric case} \quad \lambda^t := (Az^t, z^t)_2, \tag{4.1.12}$$

instead. This is justified since z^t is supposed to be an approximation to an eigenvector of $(A - \tilde{\lambda}I)^{-1}$ corresponding to the eigenvalue μ_k, which is also an eigenvector of A corresponding to the desired eigenvalue λ_k.

In virtue of the above result for the simple power method, for any diagonalizable matrix A the "Inverse Iteration" delivers any eigenvalue, for which a sufficiently accurate approximation is known. There holds the error estimate

$$\mu^t = \mu_k + \mathcal{O}\Big(\Big|\frac{\mu_{k-1}}{\mu_k}\Big|^t\Big) \quad (t \to \infty), \tag{4.1.13}$$

where μ_{k-1} is the eigenvalue of $(A - \tilde{\lambda}I)^{-1}$ closest to the "maximum" eigenvalue μ_k. From this, we infer

$$\mu^t = \frac{1}{\lambda_k - \tilde{\lambda}} + \mathcal{O}\Big(\Big|\frac{\lambda_k - \tilde{\lambda}}{\lambda_{k-1} - \tilde{\lambda}}\Big|^t\Big) \quad (t \to \infty), \tag{4.1.14}$$

where $\lambda_{k-1} := 1/\mu_{k-1} + \tilde{\lambda}$, and eventually,

$$\lambda_k^t := \frac{1}{\mu^t} + \tilde{\lambda} = \lambda_k + \mathcal{O}\Big(\Big|\frac{\lambda_k - \tilde{\lambda}}{\lambda_{k-1} - \tilde{\lambda}}\Big|^t\Big) \quad (t \to \infty). \tag{4.1.15}$$

We collect the above results for the special case of the computation of the "smallest" eigenvalue $\lambda_{\min} = \lambda_1$ of a diagonalizable matrix A in the following theorem.

Theorem 4.2 (Inverse Iteration): *Let the matrix A be diagonalizable and suppose that the eigenvalue with smallest modulus is separated from the other eigenvalues, i. e., $|\lambda_1| < |\lambda_i|$, $i = 2, \ldots, n$. Further, let the starting vector z^0 have a nontrivial component with respect to the eigenvector w^1. Then, for the "Inverse Iteration" (with shift $\tilde{\lambda} := 0$) there are numbers $\sigma_t \in \mathbb{C}$, $|\sigma_t| = 1$ such that*

$$\|z^t - \sigma_t w^1\| \to 0 \quad (t \to \infty), \tag{4.1.16}$$

and the "smallest" eigenvalue $\lambda_{\min} = \lambda_1$ of A is approximated with convergence speed, in the general non-Hermitian case using (4.1.10),

$$\lambda^t = \lambda_{\min} + \mathcal{O}\Big(\Big|\frac{\lambda_{\min}}{\lambda_2}\Big|^t\Big) \quad (t \to \infty). \tag{4.1.17}$$

and with squared power $2t$ in the Hermitian case using (4.1.11).

Remark 4.2: The inverse iteration allows the approximation of any eigenvalue of A for which a sufficiently good approximation is known, where "sufficiently good" depends on the separation of the desired eigenvalue of A from the other ones. The price to be paid for this flexibility is that each iteration step requires the solution of the nearly singular system $(A - \tilde{\lambda} I)z^t = z^{t-1}$. This means that the better the approximation $\tilde{\lambda} \approx \lambda_k$, i. e., the faster the convergence of the Inverse Iteration is, the more expensive is each iteration step. This effect is further amplified if the Inverse Iteration is used with "dynamic shift" $\tilde{\lambda} := \lambda_k^t$, in order to speed up its convergence.

The solution of the nearly singular linear systems (4.1.9),

$$(A - \tilde{\lambda} I)\tilde{z}^t = z^{t-1},$$

can be accomplished, for moderately sized matrices, by using an a priori computed LR or Cholesky (in the Hermitian case) decomposition and, for large matrices, by the GMRES or the BiCGstab method and the CG method (in the Hermitian case). The matrix $A - \tilde{\lambda} I$ is very ill-conditioned with condition number

$$\mathrm{cond}_2(A - \tilde{\lambda} I) = \frac{|\lambda_{\max}(A - \tilde{\lambda} I)|}{|\lambda_{\min}(A - \tilde{\lambda} I)|} = \frac{\max_{j=1,\ldots,n} |\lambda_j - \tilde{\lambda}|}{|\lambda_k - \tilde{\lambda}|} \gg 1.$$

Therefore, preconditioning is mandatory. However, only the "direction" of the iterate $\tilde{z}^t$ is needed, which is a much better conditoned task almost independent of the quality of the approximation $\tilde{\lambda}$ to λ_k. In this case a good preconditioning is obtained by the *incomplete* LR (or the *incomplete* Cholesky) decomposition.

Example 4.1: We want to apply the considered methods to the eigenvalue problem of the model matrix from Section 3.4. The determination of vibration mode and frequency of a membrane over the square domain $\Omega = (0,1)^2$ (drum) leads to the eigenvalue problem of the Laplace operator

$$\begin{aligned} -\frac{\partial^2 w}{\partial x^2}(x,y) - \frac{\partial^2 w}{\partial y^2}(x,y) &= \mu w(x,y) \quad \text{for } (x,y) \in \Omega, \\ w(x,y) &= 0 \quad \text{for } (x,y) \in \partial\,\Omega. \end{aligned} \qquad (4.1.18)$$

This eigenvalue problem in function space shares several properties with that of a symmetric, positive definite matrix in $\mathbb{R}^n$. First, there are only countably many real, positive eigenvalues with finite (geometric) multiplicities. The corresponding eigenspaces span the whole space $L^2(\Omega)$. The smallest of theses eigenvalues, $\mu_{\min} > 0$, and the associated eigenfunction, $w_{\min}$, describe the fundamental tone and the fundamental oscillation mode of the drum. The discretization by the 5-point difference operator leads to the matrix eigenvalue problem

$$Az = \lambda z, \quad \lambda = h^2 \mu, \qquad (4.1.19)$$

with the same block-tridiagonal matrix A as occurring in the corresponding discretization

of the boundary value problem discussed in Section 3.4. Using the notation from above, the eigenvalues of A are explicitly given by

$$\lambda_{kl} = 4 - 2(\cos(kh\pi) + \cos(lh\pi)), \quad k, l = 1, \ldots, m.$$

We are interested in the smallest eigenvalue $\lambda_{\min}$ of A, which by $h^{-2}\lambda_{\min} \approx \mu_{\min}$ yields an approximation to the smallest eigenvalue of problem (4.1.18). For $\lambda_{\min}$ and the next eigenvalue $\lambda^* > \lambda_{\min}$ there holds

$$\begin{aligned} \lambda_{\min} &= 4 - 4\cos(h\pi) = 2\pi^2 h^2 + O(h^4), \\ \lambda^* &= 4 - 2(\cos(2h\pi) + \cos(h\pi)) = 5\pi^2 h^2 + O(h^4). \end{aligned}$$

For computing $\lambda_{\min}$, we may use the inverse iteration with shift $\lambda = 0$. This requires in each iteration the solution of a linear system like

$$Az^t = z^{t-1}. \tag{4.1.20}$$

For the corresponding eigenvalue approximation

$$\lambda^t = (z^{t+1}, z^t)_2, \tag{4.1.21}$$

there holds the convergence estimate

$$|\lambda^t - \lambda_{\min}| \approx \left(\frac{\lambda_{\min}}{\lambda^*}\right)^{2t} \approx \left(\frac{2}{5}\right)^{2t}, \tag{4.1.22}$$

i. e., the convergence is independent of the mesh size h or the dimension $n = m^2 \approx h^{-2}$ of A. However, in view of the relation $\mu_{\min} = h^{-2}\lambda_{\min}$ achieving a prescribed accuracy in the approximation of $\mu_{\min}$ requires the scaling of the tolerance in computing $\lambda_{\min}$ by a factor h^2, which introduces a logarithmic h-dependence in the work count of the algorithm,

$$t(\varepsilon) \approx \frac{\log(\varepsilon h^2)}{\log(2/5)} \approx \log(n). \tag{4.1.23}$$

This strategy for computing $\mu_{\min}$ is not very efficient if the solution of the subproblems (4.1.20) would be done by the PCG method. For reducing the work, one may use an iteration-dependent stopping criterion for the inner PCG iteration by which its accuracy is balanced against that of the outer inverse iteration.

Remark 4.3: Another type of iterative methods for computing single eigenvalues of symmetric or nonsymmetric large-scale matrices is the "Jacobi-Davidson method" (Davidson [30]), which is based on the concept of defect correction. This method will not be discussed in these lecture notes, we rather refer to the literature, e. g., Crouzeix et al. [29] and Sleijpen & Van der Vorst [48]

4.2 Methods for the full eigenvalue problem

In this section, we consider iterative methods for solving the *full* eigenvalue problem of an arbitrary matrix $A \in \mathbb{R}^{n\times n}$. Since these methods use successive factorizations of matrices, which for general full matrices have arithmetic complexity $\mathcal{O}(n^3)$, they are only applied to matrices with special sparsity pattern such as general Hessenberg or symmetric tridiagonal matrices. In the case of a general matrix therefore at firsr a reduction to such special structure has to be performed (e. g., by applying Householder transformations as discussed in Section 2.5.1). As application of such a method, we discuss the computation of the singular value decomposition of a general matrix. In order to avoid confusion between "indexing" and "exponentiation", in the following, we use the notation $A^{(t)}$ instead of the short version A^t for elements in a sequence of matrices.

4.2.1 The LR and QR method

I) The "LR method" of Rutishauser[3] (1958), starting from some initial guess $A^{(1)} := A$, generates a sequence of matrices $A^{(t)}$, $t \in \mathbb{N}$, by the prescription

$$A^{(t)} = L^{(t)}R^{(t)} \text{ (LR decomposition)}, \quad A^{(t+1)} := R^{(t)}L^{(t)}. \tag{4.2.24}$$

Since

$$A^{(t+1)} = R^{(t)}L^{(t)} = L^{(t)-1}L^{(t)}R^{(t)}L^{(t)} = (L^{(t)-1}A^{(t)}L^{(t)},$$

all iterates $A^{(t)}$ are similar to A and therefore have the same eigenvalues as A. Under certain conditions on A, one can show that, with the eigenvalues λ_i of A:

$$\lim_{t\to\infty} A^{(t)} = \lim_{t\to\infty} R^{(t)} = \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}, \quad \lim_{t\to\infty} L^{(t)} = I. \tag{4.2.25}$$

The LR method requires in each step the computation of an LR decomposition and is consequently by far too costly for general full matrices. For Hessenberg matrices the work is acceptable. The most severe disadvantage of the LR method is the necessary existence of the LR decompositions $A^{(t)} = L^{(t)}R^{(t)}$. If only a decomposition $P^{(t)}A^{(t)} = L^{(t)}R^{(t)}$ exists with a perturbation matrix $P^{(t)} \neq I$ the method may not converge. This problem is avoided by the so-called "QR method".

II) The "QR method" of Francis[4] (1961) is considered as the currently most efficient method for solving the full eigenvalue problem of Hessenberg matrices. Starting from

[3]Heinz Rutishauser (1918–1970): Swiss mathematician and computer scientist; since 1962 Prof. at ETH Zurich; contributions to Numerical Linear Algebra (LR method: "Solution of eigenvalue problems with the LR transformation", Appl. Math. Ser. nat. Bur. Stand. 49, 47-81(1958).) and Analysis as well as to the foundation of Computer Arithmetik.

[4]J. F. G. Francis: "The QR transformation. A unitary analogue to the LR transformation", Computer J. 4, 265-271 (1961/1962).

some initial guess $A^{(1)} = A$ a sequence of matrices $A^{(t)}$, $t \in \mathbb{N}$, is generated by the prescription

$$A^{(t)} = Q^{(t)}R^{(t)} \text{ (QR decomposition)}, \quad A^{(t+1)} := R^{(t)}Q^{(t)}, \tag{4.2.26}$$

where $Q^{(t)}$ is unitary and $R^{(t)}$ is an upper triangular matrix with *positive* diagonal elements (in order to ensure its uniqueness). The QR decomposition can be obtained, e. g., by employing Householder transformations. Because of the high costs of this method for a general full matrix the QR method is economical only for Hessenberg matrices or, in the symmetric case, only for tridiagonal matrices. Since

$$A^{(t+1)} = R^{(t)}Q^{(t)} = Q^{(t)T}Q^{(t)}R^{(t)}Q^{(t)} = Q^{(t)T}A^{(t)}Q^{(t)},$$

all iterates $A^{(t)}$ are similar to A and therefore have the same eigenvalues as A. The proof of convergence of the QR method will use the following auxiliary lemma.

Lemma 4.1: *Let $E^{(t)} \in \mathbb{R}^{n\times n}$, $t \in \mathbb{N}$, be regular matrices, which satisfy $\lim_{t\to\infty} E^{(t)} = I$ and possess the QR decompositions $E^{(t)} = Q^{(t)}R^{(t)}$ with $r_{ii} > 0$. Then, there holds*

$$\lim_{t\to\infty} Q^{(t)} = I = \lim_{t\to\infty} R^{(t)}. \tag{4.2.27}$$

Proof. Since

$$\|E^{(t)} - I\|_2 = \|Q^{(t)}R^{(t)} - Q^{(t)}{Q^{(t)}}^T\|_2 = \|Q^{(t)}(R^{(t)} - {Q^{(t)}}^T)\|_2 = \|R^{(t)} - {Q^{(t)}}^T\|_2 \to 0,$$

it follows that $q_{jk}^{(t)} \to 0\,(t \to \infty)$ for $j < k$. In view of

$$I = Q^{(t)}{Q^{(t)}}^T = \begin{bmatrix} \square & & & \to & 0 \\ & \square & & * & \\ & & \ddots & & \\ & * & & \square & \\ & & & & \square \end{bmatrix} \begin{bmatrix} \square & & & & \\ & \square & & * & \\ & & * & \ddots & \\ & & & & \square \\ \to & 0 & & & \square \end{bmatrix},$$

we conclude that

$$q_{jj}^{(t)} \to \pm 1\,, \quad q_{jk}^{(t)} \to 0 \quad (t \to \infty), \quad j > k.$$

Hence $Q^{(t)} \to \mathrm{diag}(\pm 1)$ (t $\to \infty$). Since

$$Q^{(t)}R^{(t)} = E^{(t)} \to I \quad (t \to \infty)\,, \quad r_{jj} > 0,$$

also $\lim_{t\to\infty} Q^{(t)} = I$. Then,

$$\lim_{t\to\infty} R^{(t)} = \lim_{t\to\infty} {Q^{(t)}}^T E^{(t)} = I,$$

what was to be shown. Q.E.D.

Theorem 4.3 (QR method): *Let the eigenvalues of the matrix $A \in \mathbb{R}^{n\times n}$ be separated with respect to their modulus, i. e., $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$. Then, the matrices $A^{(t)} = (a_{jk}^{(t)})_{j,k=1,\ldots,n}$ generated by the QR method converge like*

$$\{\lim_{t\to\infty} a_{jj}^{(t)} \mid j = 1,\ldots,n\} = \{\lambda_1,\ldots,\lambda_n\}. \tag{4.2.28}$$

Proof. The separation assumption implies that all eigenvalues of the matrix A are simple. There holds

$$\begin{aligned} A^{(t)} &= R^{(t-1)}Q^{(t-1)} = Q^{(t-1)T}Q^{(t-1)}R^{(t-1)}Q^{(t-1)} = Q^{(t-1)T}A^{(t-1)}Q^{(t-1)} \\ &= \ldots = [Q^{(1)}\ldots Q^{(t-1)}]^T A[Q^{(1)}\ldots Q^{(t-1)}] =: P^{(t-1)T}AP^{(t-1)}. \end{aligned} \tag{4.2.29}$$

The normalized eigenvectors w^i, $\|w^i\| = 1$, associated to the eigenvalues λ_i are linearly independent. Hence, the matrix $W = [w_1,\ldots,w_n]$ is regular and there holds the relation $AW = W\Lambda$ with the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_\mathrm{i})$. Consequently,

$$A = W\Lambda W^{-1}.$$

Let $QR = W$ be a QR decomposition of W and $LS = PW^{-1}$ an LR decomposition of PW^{-1} (P an appropriate permutation matrix). In the following, we consider the simple case that $P = I$. There holds

$$\begin{aligned} A^t &= [W\Lambda W^{-1}]^t = W\Lambda^t W^{-1} = [QR]\Lambda^t[LS] = QR[\Lambda^t L\Lambda^{-t}]\Lambda^t S \\ &= QR\begin{bmatrix} 1 & & 0 \\ & \ddots & \\ l_{jk}\left(\frac{\lambda_j}{\lambda_k}\right)^t & & 1 \end{bmatrix}\Lambda^t S \\ &= QR[I + N^{(t)}]\Lambda^t S = Q[R + RN^{(t)}]\Lambda^t S, \end{aligned}$$

and, consequently,

$$A^t = Q[I + RN^{(t)}R^{-1}]R\Lambda^t S. \tag{4.2.30}$$

By the assumption on the separation of the eigenvalues λ_i, we have $|\lambda_j/\lambda_k| < 1$, $j > k$, which yields

$$N^{(t)} \to 0\,, \quad RN^{(t)}R^{-1} \to 0 \quad (t\to\infty).$$

Then, for the (uniquely determined) QR decomposition $\tilde{Q}^{(t)}\tilde{R}^{(t)} = I + RN^{(t)}R^{-1}$ with $\tilde{r}_{ii}^{(t)} > 0$, Lemma 4.1 implies

$$\tilde{Q}^{(t)} \to I\,, \quad \tilde{R}^{(t)} \to I \quad (t\to\infty).$$

Further, recalling (4.2.30),

$$A^t = Q[I + RN^{(t)}R^{-1}]R\Lambda^t S = Q[\tilde{Q}^{(t)}\tilde{R}^{(t)}]R\Lambda^t S = [Q\tilde{Q}^{(t)}][\tilde{R}^{(t)}R\Lambda^t S]$$

is obviously a QR decomposition of A^t (but with not necessarily positive diagonal elements of R). By (4.2.29) and $Q^{(t)}R^{(t)} = A^{(t)}$ there holds

$$\underbrace{[Q^{(1)} \dots Q^{(t)}]}_{=P^{(t)}} \underbrace{[R^{(t)} \dots R^{(1)}]}_{=:S^{(t)}} = \underbrace{[Q^{(1)} \dots Q^{(t-1)}]}_{=P^{(t-1)}} A^{(t)} \underbrace{[R^{(t-1)} \dots R^{(1)}]}_{=:S^{(t-1)}}$$

$$= P^{(t-1)} [P^{(t-1)T} A P^{(t-1)}] S^{(t-1)} = A P^{(t-1)} S^{(t-1)},$$

and observing $P^{(1)}S^{(1)} = A$,

$$P^{(t)}S^{(t)} = AP^{(t-1)}S^{(t-1)} = \dots = A^{t-1}P^{(1)}S^{(1)} = A^t. \tag{4.2.31}$$

This yields another QR decomposition of A^t, i. e.,

$$[Q\tilde{Q}^{(t)}][\tilde{R}^{(t)}R\Lambda^t S] = A^t = P^{(t)}S^{(t)}.$$

Since the QR decomposition of a matrix is unique up to the scaling of the column vectors of the unitary matrix Q, there must hold

$$P^{(t)} = Q\tilde{Q}^{(t)}D^{(t)} =: QT^{(t)},$$

with certain diagonal matrices $D^{(t)} = \mathrm{diag}(\pm 1)$. Then, recalling again the realtion (4.2.29) and observing that

$$A = W\Lambda W^{-1} = QR\Lambda[QR]^{-1} = QR\Lambda R^{-1}Q^T,$$

we conclude that

$$\begin{aligned} A^{(t+1)} &= P^{(t)T}AP^{(t)} = [QT^{(t)}]^T AQT^{(t)} \\ &= T^{(t)T}Q^T[QR\Lambda R^{-1}Q^T]QT^{(t)} = T^{(t)T}R\Lambda R^{-1}T^{(t)} \\ &= T^{(t)T} \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} T^{(t)} = D^{(t)}\tilde{Q}^{(t)T} \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \tilde{Q}^{(t)}D^{(t)}. \end{aligned}$$

Since $\tilde{Q}^{(t)} \to I$ $(t \to \infty)$ and $D^{(t)}D^{(t)} = I$, we obtain

$$D^{(t)}A^{(t+1)}D^{(t)} \to \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} \quad (t \to \infty).$$

In case that W^{-1} does not possess an LR decomposition, then the eigenvalues λ_i do not appear ordered according to their modulus. Q.E.D.

Remark 4.4: The separation assumption $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$ means that all eigenvalues of A are simple, which implies that A is necessarily diagonalizable. For more general matrices the convergence of the QR method is not guaranteed. However, convergence in a suitable sense can be shown in case of multiple eigenvalues (such as in the model problem of Section 3.4). For a more detailed discussion, we refer to the literature, e. g., Deuflhard & Hohmann [33], Stoer & Bulirsch [50], Golub & Loan [36], and Parlett [44].

The speed of convergence of the QR method, i. e., the convergence of the off-diagonal elements in $A^{(t)}$ to zero, is determined by the size of the quotients

$$\left|\frac{\lambda_j}{\lambda_k}\right| < 1, \quad j > k,$$

The convergence is the faster the better the eigenvalues of A are modulus-wise separated. This suggests to use the QR algorithm with a "shift" σ for the matrix $A - \sigma I$, such that

$$\left|\frac{\lambda_j - \sigma}{\lambda_k - \sigma}\right| \ll \left|\frac{\lambda_j}{\lambda_k}\right| < 1,$$

for the most interesting eigenvalues. The QR method with (dynamic) shift starts form some initial guess $A^{(1)} = A$ and constructs a sequence of matrices $A^{(t)}$, $t \in \mathbb{N}$, by the prescription

$$A^{(t)} - \sigma_t I = Q^{(t)} R^{(t)} \ (QR \text{ decomposition}), \quad A^{(t+1)} := R^{(t)} Q^{(t)} + \sigma_t I, \tag{4.2.32}$$

This algorithm again produces a sequence of similar matrices:

$$\begin{aligned} A^{(t+1)} &= R^{(t)} Q^{(t)} + \sigma_t I \\ &= Q^{(t)T} Q^{(t)} R^{(t)} Q^{(t)} + \sigma_t I = Q^{(t)T} [A^{(t)} - \sigma_t I] Q^{(t)} + \sigma_t I \\ &= Q^{(t)T} A^{(t)} Q^{(t)}. \end{aligned} \tag{4.2.33}$$

For this algorithm a modified version of the proof of Theorem 4.3 yields a convergence estimate

$$|a_{jk}^{(t)}| \le c\Big(\Big|\frac{\lambda_j - \sigma_1}{\lambda_k - \sigma_1}\Big| \cdots \Big|\frac{\lambda_j - \sigma_t}{\lambda_k - \sigma_t}\Big|\Big), \quad j > k, \tag{4.2.34}$$

for the lower off-diagonal elements of the iterates $A^{(t)} = (a_{jk}^{(t)})_{j,k=1}^n$.

Remark 4.5: For positive definite matrices the QR method converges twice as fast as the corresponding LR method, but requires about twice as much work in each iteration. Under certain structural assumpotions on the matrix A, one can show that the QR method with varying shifts converges with quadratic order for Hermitian tridiagonal matrices and even with cubic order for *unitary* Hessenberg matrices (see Wang & Gragg [56]).

$$|\lambda^{(t)} - \lambda| \le c|\lambda^{(t-1)} - \lambda|^3,$$

As the LR method, for economy reasons, also the QR method is applied only to pre-reduced matrices for which the computation of the QR decomposition is of acceptable cost, e. g., Hessenber matrices, symmetric tridiagonal matrices or more general band matrices with bandwidth $2m+1 \ll n = m^2$ (e. g., the model matrix considered in Section 3.4). This is justified by the following observation.

Lemma 4.2: *If A is a Hessenberg matrix (or a symmetric $2m+1$-band matrix), then the same holds true for the matrices $A^{(t)}$ generated by the QR method.*

Proof. The proof is posed as exercise. Q.E.D.

4.2.2 Computation of the singular value decomposition

The numerically stable computation of the singular value decomposition (SVD) is rather costly. For more details, we refer to the literature, e. g., the book by Golub & van Loan [36]. The SVD of a matrix $A \in \mathbb{C}^{n\times k}$ is usually computed by a two-step procedure. In the first step, the matrix is reduced to a *bidiagonal* matrix. This requires $\mathcal{O}(kn^2)$ operations, assuming that $k \leq n$. The second step is to compute the SVD of the bidiagonal matrix. This step needs an iterative method since the problem to be solved is generically nonlinear. For fixe4 accuracy requirement (e. g., round-off error level) this takes $\mathcal{O}(n)$ iterations, each costing $\mathcal{O}(n)$ operations. Thus, the first step is more expensive and the overall cost is $\mathcal{O}(kn^2)$ operations (see Trefethen & Bau [54]). The first step can be done using Householder reflections for a cost of $\mathcal{O}(kn^2 + n^3)$ operations, assuming that only the singular values are needed and not the singular vectors.

The second step can then very efficiently be done by the QR algorithm. The LAPACK subroutine DBDSQR[9] implements this iterative method, with some modifications to cover the case where the singular values are very small. Together with a first step using Householder reflections and, if appropriate, QR decomposition, this forms the LAPACK DGESVD[10] routine for the computation of the singular value decomposition.

If the matrix A is very large, i. e., $n \geq 10^4 - 10^8$, the method described so far for computing the SVD is too expensive. In this situation, particularly if $A \in \mathbb{C}^{n\times n}$ is square and regular, the matrix is first reduced to smaller dimension,

$$A \to A^{(m)} = Q^{(m)T} A Q^{(m)} \in \mathbb{C}^{m\times m},$$

with $m \ll n$, by using, e. g., the Arnoldi process described below in Section 4.3.1, and then the above method is applied to this reduced matrix. For an appropriate choice of the orthonormal transformation matrix $Q^{(m)} \in \mathbb{C}^{n\times m}$ the singular values of $A^{(m)}$ are approximations of those of A, especially the "largest" ones (by modulus). If one is interested in the "smallest" singular values of A, what is typically the case in applications, the dimensional reduction process has to be applied to the inverse matrix A^{-1}.

4.3 Krylov space methods

"Krylov space methods" for solving eigenvalue problems follow essentially the same idea as in the case of the solution of linear systems. The original high-dimensional problem is reduced to smaller dimension by applying the Galerkin approximation in appropriate subspaces, e. g., so-called "Krylov space", which are sucessively constructed using the given matrix and sometimes also its transpose. The work per iteration should amount to about one matrix-vector multiplication. We will consider the two most popular variants of such methods, the "Arnoldi[5] method" for general, not necessarily Hermitian matrices, and its specialization for Hermitian matrices, the "Lanczos[6] method".

First, we introduce the general concept of such a "model reduction" by "Galerkin approximation". Consider a general eigenvalue problem

$$Az = \lambda z, \tag{4.3.35}$$

with a higher-dimensional matrix $A \in \mathbb{C}^{n\times n}$, $n \geq 10^4$, which may have resulted from the discretization of the eigenvalue problem of a partial differential operator. This eigenvalue problem can equivalently be written in variational from as

$$z \in \mathbb{C}^n,\ \lambda \in \mathbb{C}: \quad (Az, y)_2 = \lambda(z, y)_2 \quad \forall y \in \mathbb{C}^n. \tag{4.3.36}$$

Let $K_m = \operatorname{span}\{q^1, \ldots, q^m\}$ be an appropriately chosen subspace of $\mathbb{C}^n$ of smaller dimension $\dim K_m = m \ll n$. Then, the n-dimensional eigenvalue problem (4.3.36) is approximated by the m-dimensional "Galerkin eigenvalue problem"

$$z \in K_m,\ \lambda \in \mathbb{C}: \quad (Az, y)_2 = \lambda(z, y)_2 \quad \forall y \in K_m. \tag{4.3.37}$$

Expanding the eigenvector $z \in K_m$ with respect to the given basis, $z = \sum_{j=1}^m \alpha_j q^j$, the Galerkin system takes the form

$$\sum_{j=1}^m \alpha_j (Aq^j, q^i)_2 = \lambda \sum_{j=1}^m \alpha_j (q^j, q^i)_2, \quad i = 1, \ldots, m, \tag{4.3.38}$$

[5]Walter Edwin Arnoldi (1917–1995): US-American engineer; graduated in Mechanical Engineering at the Stevens Institute of Technology in 1937; worked at United Aircraft Corp. from 1939 to 1977; main research interests included modelling vibrations, Acoustics and Aerodynamics of aircraft propellers; mainly known for the "Arnoldi iteration"; the paper "The principle of minimized iterations in the solution of the eigenvalue problem", Quart. Appl. Math. 9, 17-29 (1951), is one of the most cited papers in Numerical Linear Algebra.

[6]Cornelius (Cornel) Lanczos (1893–1974): Hungarian mathematician and physicist; PhD in 1921 on Relativity Theory; assistant to Albert Einstein 19281929; contributions to exact solutions of the Einstein field equation; discovery of the fast Fourier transform (FFT) 1940; worked at the U.S. National Bureau of Standards after 1949; invented the "Lanczos algorithm" for finding eigenvalues of large symmetric matrices and the related conjugate gradient method; in 1952 he left the USA for the School of Theoretical Physics at the Dublin Institute for Advanced Studies in Ireland, where he succeeded Schrödinger and stayed until 1968; Lanczos was author of many classical text books.

Within the framework of Galerkin approximation this is usually written in compact form as a generalized eigenvalue problem

$$\mathcal{A}\alpha = \lambda \mathcal{M}\alpha, \tag{4.3.39}$$

for the vector $\alpha = (\alpha_j)_{j=1}^n$, involving the matrices $\mathcal{A} = \big((Aq^j, q^i)_2\big)_{i,j=1}^n$ and $\mathcal{M} = \big((q^j, q^i)_2\big)_{i,j=1}^n$.

In the following, we use another formulation. With the Cartesian representations of the basis vectors $q^i = (q_j^i)_{j=1}^n$ the Galerkin eigenvalue problem (4.3.37) is written in the form

$$\sum_{j=1}^m \alpha_j \sum_{k,l=1}^n a_{kl} q_k^j \bar{q}_l^i = \lambda \sum_{j=1}^m \alpha_j \sum_{k=1}^n q_k^j \bar{q}_k^i, \quad i = 1, \ldots, m. \tag{4.3.40}$$

Then, using the matrix $Q^{(m)} := [q^1, \ldots, q^m] \in \mathbb{C}^{n\times m}$ and the vector $\alpha = (\alpha_j)_{j=1}^m \in \mathbb{C}^m$ this can be written in compact form as

$$\bar{Q}^{(m)T} A Q^{(m)} \alpha = \lambda \bar{Q}^{(m)T} Q^{(m)} \alpha. \tag{4.3.41}$$

If $\{q^1, \ldots, q^m\}$ were an ONB of K_m this reduces to the normal eigenvalue problem

$$\bar{Q}^{(m)T} A Q^{(m)} \alpha = \lambda \alpha, \tag{4.3.42}$$

of the reduced matrix $H^{(m)} := \bar{Q}^{(m)T} A Q^{(m)} \in \mathbb{C}^{m\times m}$. If the reduced matrix $H^{(m)}$ has a particular structure, e. g., a Hessenberg matrix or a symmetric tridiagonal matrix, then, the lower-dimensional eigenvalue problem (4.3.42) can efficiently be solved by the QR method. Its eigenvalues may be considered as approximations to some of the dominant eigenvalues of the original matrix A and are called "Ritz[7] eigenvalues" of A. In view of this preliminary consideration the "Krylov methods" consist in the following steps:

1. Choose an appropriate subspace $K_m \subset \mathbb{C}^n$, $m \ll n$ (a "Krylov space"), using the matrix A and powers of it.
2. Construct an ONB $\{q^1, \ldots, q^m\}$ of K_m by the stabilized version of the Gram-Schmidt algorithm, and set $Q^{(m)} := [q^1, \ldots, q^m]$.
3. Form the matrix $H^{(m)} := \bar{Q}^{(m)T} A Q^{(m)}$, which then by construction is a Hessenberg matrix or, in the Hermitian case, a Hermitian tridiagonal matrix.
4. Solve the eigenvalue problem of the reduced matrix $H^{(m)} \in \mathbb{C}^{m\times m}$ by the QR method.
5. Take the eigenvalues of $H^{(m)}$ as approximations to the dominant (i. e., "largest") eigenvalues of A. If the "smallest" eigenvalues (i. e., those closest to the origin) are

[7] Walter Ritz (1878–1909): Swiss physicist; Prof. in Zürich and Göttingen; contributions to Spectral Theorie in Nuclear Physics and Electromagnetism.

to be determined the whole process has to be applied to the inverse matrix A^{-1}, which possibly makes the construction of the subspace K_m expensive.

Remark 4.6: In the above form the Krylov method for eigenvalue problems is analogous to its version for (real) linear systems described in Section 3.3.3. Starting from the variational form of the linear system

$$x \in \mathbb{R}^n : \quad (Ax, y)_2 = (b, y)_2 \quad \forall y \in \mathbb{R}^n,$$

we obtain the following reduced system for $x^m = \sum_{j=1}^m \alpha_j q^j$:

$$\sum_{j=1}^{m} \alpha_j \sum_{k,l=1}^{n} a_{kl} q_k^j q_l^i = \sum_{k=1}^{n} b_k q_k^i, \qquad i = 1, \ldots, m.$$

This is then equivalent to the m-dimensional algebraic system

$$Q^{(m)T} A Q^{(m)} \alpha = Q^{(m)T} b.$$

4.3.1 Lanczos and Arnoldi method

The "power method" for computing the largest eigenvalue of a matrix only uses the current iterate $A^m q$, $m \ll n$, for some normalized starting vector $q \in \mathbb{C}^n$, $\|q\|_2 = 1$, but ignores the information contained in the already obtained iterates $\{q, Aq, \ldots, A^{(m-1)}q\}$. This suggests to form the so-called "Krylov matrix"

$$K_m = [q, Aq, A^2 q, \ldots, A^{m-1} q], \quad 1 \le m \le n.$$

The columns of this matrix are not orthogonal. In fact, since $A^t q$ converges to the direction of the eigenvector corresponding to the largest (in modulus) eigenvalue of A, this matrix tends to be badly conditioned with increasing dimension m. Therefore, one constructs an orthogonal basis by the Gram-Schmidt algorithm. This basis is expected to yield good approximations of the eigenvectors corresponding to the m largest eigenvalues, for the same reason that $A^{m-1}q$ approximates the dominant eigenvector. However, in this simplistic form the method is unstable due to the instability of the standard Gram-Schmidt algorithm. Instead the "Arnoldi method" uses a stabilized version of the Gram-Schmidt process to produce a sequence of orthonormal vectors, $\{q^1, q^2, q^3, \ldots\}$ called the "Arnoldi vectors", such that for every m, the vectors $\{q^1, \ldots, q^m\}$ span the Krylov subspace K_m. For the following, we define the orthogonal projection operator

$$\mathrm{proj}_u(v) := \|u\|_2^{-2} (v, u)_2 u\,,$$

which projects the vector v onto $\mathrm{span}\{u\}$. With this notation the classical Gram-Schmidt orthonormalization process uses the recurrence formulas:

$$
\begin{aligned}
&q^1 = \|q\|_2^{-1} q, \quad t = 2, \ldots, m: \\
&\tilde{q}^t = A^{t-1} q - \sum_{j=1}^{t-1} \mathrm{proj}_{q^j}(A^{t-1} q), \quad q^t = \|\tilde{q}^t\|_2^{-1} \tilde{q}^t.
\end{aligned}
\tag{4.3.43}
$$

Here, the t-th step projects out the component of $A^{t-1}q$ in the directions of the already determined orthonormal vectors $\{q^1, \ldots, q^{t-1}\}$. This algorithm is numerically unstable due to round-off error accumulation. There is a simple modification, the so-called "modified Gram-Schmidt algorithm", where the t-th step projects out the component of Aq^t in the directions of $\{q^1, \ldots, q^{t-1}\}$:

$$
\begin{aligned}
&q^1 = \|q\|_2^{-1} q, \quad t = 2, \ldots, m: \\
&\tilde{q}^t = A q^{t-1} - \sum_{j=1}^{t-1} \mathrm{proj}_{q^j}(A q^{t-1}), \quad q^t = \|\tilde{q}^t\|_2^{-1} \tilde{q}^t.
\end{aligned}
\tag{4.3.44}
$$

Since q^t, $\tilde{q}^t$ are aligned and $\tilde{q}^t \perp K_t$, we have

$$
(q^t, \tilde{q}^t)_2 = \|\tilde{q}^t\|_2 = \big(\tilde{q}^t, A q^{t-1} - \sum\nolimits_{j=1}^{t-1} \mathrm{proj}_{q^j}(A q^{t-1})\big)_2 = (\tilde{q}^t, A q^{t-1})_2.
$$

Then, with the setting $h_{i,t-1} := (Aq^{t-1}, q^i)_2$, from (4.3.44), we infer that

$$
A q^{t-1} = \sum_{i=1}^{t} h_{i,t-1} q^i, \quad t = 2, \ldots, m+1. \tag{4.3.45}
$$

In practice the algorithm (4.3.44) is implemented in the following equivalent recursive form:

$$
\begin{aligned}
&q^1 = \|q\|_2^{-1} q, \quad t = 2, \ldots, m: \\
&j = 1, \ldots, t-1: \quad q^{t,1} = A q^{t-1}, \\
&\qquad\qquad q^{t,j+1} = q^{t,j} - \mathrm{proj}_{q^j}(q^{t,j}), \quad q^t = \|q^{t,t}\|_2^{-1} q^{t,t}.
\end{aligned}
\tag{4.3.46}
$$

This algorithm gives the same result as the original formula (4.3.43) but introduces smaller errors in finite-precision arithmetic. Its cost is asymptotically $2nm^2$ a. op.

Definition 4.3 (Arnoldi algorithm): *For a general matrix $A \in \mathbb{C}^{n\times n}$ the Arnoldi method determines a sequence of orthonormal vectors $q^t \in \mathbb{C}^n$, $1 \le t \le m \ll n$ ("Arnoldi basis"), by applying the modified Gram-Schmidt method (4.3.46) to the basis $\{q, Aq, \ldots, A^{m-1}q\}$ of the Krylov space K_m:*

$$
\begin{aligned}
&\textit{Starting vector:} && q^1 = \|q\|_2^{-1} q. \\
&\textit{Iterate for } 2 \le t \le m\textit{:} && q^{t,1} = A q^{t-1}, \\
&j = 1, \ldots, t-1: && h_{j,t} = (q^{t,j}, q^j)_2, \quad q^{t,j+1} = q^{t,j} - h_{j,t} q^j, \\
& && h_{t,t} = \|q^{t,t}\|_2, \quad q^t = h_{t,t}^{-1} q^{t,t}
\end{aligned}
$$

Let $Q^{(m)}$ denote the $n \times m$-matrix formed by the first m Arnoldi vectors $\{q^1, q^2, \ldots, q^m\}$, and let $H^{(m)}$ be the (upper Hessenberg) $m \times m$-matrix formed by the numbers h_{jk}:

$$Q^{(m)} := [q^1, q^2, \ldots, q^m], \qquad H^{(m)} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & \ldots & h_{1m} \\ h_{21} & h_{22} & h_{23} & \ldots & h_{2m} \\ 0 & h_{32} & h_{33} & \ldots & h_{3m} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & & 0 & h_{m,m-1} & h_{mm} \end{bmatrix}.$$

The matrices $Q^{(m)}$ are orthonormal and in view of (4.3.45) satisfy ("Arnoldi relation")

$$AQ^{(m)} = Q^{(m)} H^{(m)} + h_{m+1,m}[0, \ldots, 0, q^{m+1}]. \tag{4.3.47}$$

Multiplying by $\bar{Q}^{(m)T}$ from the left and observing $\bar{Q}^{(m)T} Q^{(m)} = I$ and $\bar{Q}^{(m)T} q^{m+1} = 0$, we infer that

$$H^{(m)} = \bar{Q}^{(m)T} A Q^{(m)}. \tag{4.3.48}$$

In the limit case $m = n$ the matrix $H^{(n)}$ is similar to A and, therefore, has the same eigenvalues. This suggests that even for $m \ll n$ the eigenvalues of the reduced matrix $H^{(m)}$ may be good approximations to some eigenvalues of A. When the algorithm stops (in exact arithmetic) for some $m < n$ by $h_{m+1,m} = 0$, then the Krylov space $\mathbb{K}_m$ is an invariant subspace of the matrix A and the reduced matrix $H^{(m)} = \bar{Q}^{(m)T} A Q^{(m)}$ has m eigenvalues in common with A (exercise), i. e.,

$$\sigma(H^{(m)}) \subset \sigma(A).$$

The following lemma provides an a posteriori bound for the accuracy in approximating eigenvalues of A by those of $H^{(m)}$.

Lemma 4.3: *Let $\{\mu, w\}$ be an eigenpair of the Hessenberg matrix $H^{(m)}$ and let $v = Q^{(m)} w$ so that $(\mu, v\}$ is an approximate eigenpair of A. Then, there holds*

$$\|Aw - \mu w\|_2 = |h_{m+1,m}|\, |w_m|, \tag{4.3.49}$$

where w_m is the last component of the eigenvector w.

Proof. Multiplying in (4.3.47) by w yields

$$\begin{aligned} Av = AQ^{(m)} w &= Q^{(m)} H^{(m)} w + h_{m+1,m}[0, \ldots, 0, q^{m+1}] w \\ &= \mu Q^{(m)} w + h_{m+1,m}[0, \ldots, 0, q^{m+1}] w = \mu v + h_{m+1,m}[0, \ldots, 0, q^{m+1}] w. \end{aligned}$$

Consequently, observing $\|q^{m+1}\|_2 = 1$,

$$\|Av - \mu v\|_2 = |h_{m+1,m}|\, |w_m|,$$

which is the asserted identity. Q.E.D.

The relation (4.3.49) does not provide a priori information about the convergence of the eigenvalues of $H^{(m)}$ against those of A for $m \to n$, but in view of $\sigma(H^{(n)}) = \sigma(A)$ this is not the question. Instead, it allows for an a posteriori check on the basis of the computed quantities $h_{m+q,m}$ and w_m whether the obtained pair $\{\mu, w\}$ is a reasonable approximation.

Remark 4.7: i) Typically, the Ritz eigenvalues converge to the extreme ("maximal") eigenvalues of A. If one is interested in the "smallest" eigenvalues, i. e., those which are closest to zero, the method has to be applied to the inverse matrix A^{-1}, similar to the approach used in the "Inverse Iteration". In this case the main work goes into the generation of the Krylov space $K_m = \text{span}\{q, A^{-1}q, \ldots, (A^{-1})^{m-1}q\}$, which requires the successive solution of linear systems,

$$v^0 := q, \quad Av^1 = v^0, \quad \ldots \quad Av^m = v^{m-1}.$$

ii) Due to practical storage consideration, common implementations of Arnoldi methods typically restart after some number of iterations. Theoretical results have shown that convergence improves with an increase in the Krylov subspace dimension m. However, an a priori value of m which would lead to optimal convergence is not known. Recently a dynamic switching strategy has been proposed, which fluctuates the dimension m before each restart and thus leads to acceleration of convergence.

Remark 4.8: The algorithm (4.3.46) can be used also for the stable orthonormalization of a general basis $\{v^1, \ldots, v^m\} \subset \mathbb{C}^n$:

$$\begin{aligned} &u^1 = \|v^1\|_2^{-1} v^1, \quad t = 2, \ldots, m: \\ &j = 1, \ldots, t-1: \quad u^{t,1} = v^t, \\ &\qquad\qquad u^{t,j+1} = u^{t,j} - \text{proj}_{u^j}(u^{t,j}), \quad u^t = \|u^{t,t}\|_2^{-1} u^{t,t}. \end{aligned} \tag{4.3.50}$$

This "modified" Gram-Schmidt algorithm (with exact arithmetic) gives the same result as its "classical" version (exercise)

$$\begin{aligned} &u^1 = \|v^1\|_2^{-1} v^1, \quad t = 2, \ldots, m: \\ &\tilde{u}^t = v^t - \sum_{j=1}^{t-1} \text{proj}_{u^j}(v^t), \quad u^t = \|\tilde{u}^t\|_2^{-1} \tilde{u}^t. \end{aligned} \tag{4.3.51}$$

Both algorithms have the same arithmetic complexity (exercise). In each step a vector is determined orthogonal to its preceding one and also orthogonal to any errors introduced in the computation, which enhances stability. This is supported by the following stability estimate for the resulting "orthonormal" matrix $U = [u^1, \ldots, u^m]$

$$\|U^T U - I\|_2 \leq \frac{c_1 \, \text{cond}_2(A)}{1 - c_2 \, \text{cond}_2(A)} \, \varepsilon. \tag{4.3.52}$$

The proof can be found in Björck & Paige [26].

Remark 4.9: Other orthogonalization algorithms use Householder transformations or Givens rotations. The algorithms using Householder transformations are more stable than the stabilized Gram-Schmidt process. On the other hand, the Gram-Schmidt process produces the t-th orthogonalized vector after the t-th iteration, while orthogonalization using Householder reflections produces all the vectors only at the end. This makes only the Gram-Schmidt process applicable for iterative methods like the Arnoldi iteration. However, in Quantum Mechanics there are several orthogonalization schemes with characteristics even better suited for applications than the Gram-Schmidt algorithm

As in the solution of linear systems by Krylov space methods, e. g., the GMRES method, the high storage needs for general matrices are avoided in the case of Hermitian matrices due to the availability of short recurrences in the orthonormalization process. This is exploited in the "Lanczos method". Suppose that the matrix A is Hermitian. Then, the recurrence formula of the Arnoldi method

$$\tilde{q}^t = Aq^{t-1} - \sum_{j=1}^{t-1} (Aq^{t-1}, q^j)_2 q^j, \quad t = 2, \ldots, m+1,$$

because of $(Aq^{t-1}, q^j)_2 = (q^{t-1}, Aq^j)_2 = 0$, $j = 1, \ldots, t-3$, simplifies to

$$\tilde{q}^t = Aq^{t-1} - \underbrace{(Aq^{t-1}, q^{t-1})_2}_{=:\,\alpha_{t-1}} q^{t-1} - \underbrace{(Aq^{t-1}, q^{t-2})_2}_{=:\,\beta_{t-2}} q^{t-2} = Aq^{t-1} - \alpha_{t-1} q^{t-1} - \beta_{t-2} q^{t-2}.$$

Clearly, $\alpha_{t-1} \in \mathbb{R}$ since A Hermitian. Further, multiplying this identity by q^t yields

$$\|\tilde{q}^t\|_2 = (q^t, \tilde{q}^t)_2 = (q^t, Aq^{t-1} - \alpha_{t-1} q^{t-1} - \beta_{t-2} q^{t-2})_2 = (q^t, Aq^{t-1})_2 = (Aq^t, q^{t-1})_2 = \beta_{t-1}.$$

This implies that also $\beta_{t-1} \in \mathbb{R}$ and $\beta_{t-1} q^t = \tilde{q}^t$. Collecting the foregoing relations, we obtain

$$Aq^{t-1} = \beta_{t-1} q^t + \alpha_{t-1} q^{t-1} + \beta_{t-2} q^{t-2}, \quad t = 2, \ldots, m+1. \tag{4.3.53}$$

These equations can be written in matrix form as follows:

$$AQ^{(m)} = Q^{(m)} \underbrace{\begin{bmatrix} \alpha_1 & \beta_2 & 0 & \ldots & \ldots & 0 \\ \beta_2 & \alpha_2 & \beta_3 & 0 & & \vdots \\ 0 & \beta_3 & \alpha_3 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \beta_{m-1} & 0 \\ \vdots & & 0 & \beta_{m-1} & \alpha_{m-1} & \beta_m \\ 0 & \ldots & \ldots & 0 & \beta_m & \alpha_m \end{bmatrix}}_{=:\,T^{(m)}} + \beta_m [0, \ldots, 0, q^{m+1}],$$

where the matrix $T^{(m)} \in \mathbb{R}^{m \times m}$ is real symmetric. From this so-called "Lanczos relation",

we finally obtain

$$\bar{Q}^{(m)T} A Q^{(m)} = T^{(m)}. \tag{4.3.54}$$

Definition 4.4 (Lanczos Algorithm): *For a Hermitian matrix $A \in \mathbb{C}^{n\times n}$ the Lanczos method determines a set of orthonormal vectors $\{q^1, \ldots, q^m\}$, $m \ll n$, by applying the modified Gram-Schmidt method to the basis $\{q, Aq, \ldots, A^{m-1}q\}$ of the Krylov space K_m:*

$$\begin{aligned}
&\textit{Starting values:} && q^1 = \|q\|_2^{-1} q, \quad q^0 = 0, \ \beta_1 = 0.\\
&\textit{Iterate for } 1 \le t \le m-1\textit{:} && r^t = Aq^t, \quad \alpha_t = (r^t, q^t)_2,\\
&&& s^t = r^t - \alpha_t q^t - \beta_t q^{t-1},\\
&&& \beta^{t+1} = \|s^t\|_2, \quad q^{t+1} = \beta_{t+1}^{-1} s^t,\\
&&& r^m = Aq^m, \quad \alpha_m = (r^m, q^m)_2.
\end{aligned}$$

After the matrix $T^{(m)}$ is calculated, one can compute its eigenvalues λ_i and their corresponding eigenvectors w^i, e. g., by the QR algorithm. The eigenvalues and eigenvectors of $T^{(m)}$ can be obtained in as little as $\mathcal{O}(m^2)$ work. It can be proven that the eigenvalues are approximate eigenvalues of the original matrix A. The Ritz eigenvectors v^i of A can then be calculated by $v^i = Q^{(m)} w^i$.

4.3.2 Computation of the pseudo-spectrum

As an application of the Krylov space methods described so far, we discuss the computation of the pseudo-spectrum of a matrix $A_h \in \mathbb{R}^{n\times n}$, which resulted from the discretization of a dynamical system governed by a differential operator in the context of linearized stability analysis. Hence, we are interested in the most "critical" eigenvalues, i. e., in those which are close to the origin or to the imaginary axis. This requires to consider the inverse of matrix, $T = A_h^{-1}$. Thereby, we follow ideas developed in Trefethen & Embree [22], Trefethen [21], and Gerecht et al. [35]. The following lemma collects some useful facts on the pseudo-spectra of matrices.

Lemma 4.4: *i) The ε-pseudo-spectrum of a matrix $T \in \mathbb{C}^{n\times n}$ can be equivalently defined in the following way:*

$$\sigma_\varepsilon(T) := \{z \in \mathbb{C} \,|\, \sigma_{\min}(zI - T) \le \varepsilon\}, \tag{4.3.55}$$

where $\sigma_{\min}(B)$ denotes the smallest singular value of the matrix B, i. e.,

$$\sigma_{\min}(B) := \min\{\lambda^{1/2} |\ \lambda \in \sigma(\bar{B}^T B)\},$$

with the (complex) adjoint $\bar{B}^T$ of B.

ii) The ε-pseudo-spectrum $\sigma_\varepsilon(T)$ of a matrix $T \in \mathbb{C}^{n\times n}$ is invariant under orthonormal transformations, i. e., for any unitary matrix $Q \in \mathbb{C}^{n\times n}$ there holds

$$\sigma_\varepsilon(\bar{Q}^T T Q) = \sigma_\varepsilon(T). \tag{4.3.56}$$

Proof. i) There holds

$$\begin{aligned}\|(zI-T)^{-1}\|_2 &= \max\{\mu^{1/2} |\, \mu \text{ singular value of } (zI-T)^{-1}\} \\ &= \min\{\mu^{1/2} |\, \mu \text{ singular value of } zI-T\}^{-1} = \sigma_{\min}(zI-T)^{-1},\end{aligned}$$

and, consequently,

$$\begin{aligned}\sigma_\varepsilon(T) &= \{z \in \mathbb{C} \,|\, \|(zI-T)^{-1}\|_2 \geq \varepsilon^{-1}\} \\ &= \{z \in \mathbb{C} \,|\, \sigma_{\min}(zI-T)^{-1} \geq \varepsilon^{-1}\} = \{z \in \mathbb{C} \,|\, \sigma_{\min}(zI-T) \leq \varepsilon\}.\end{aligned}$$

ii) The proof is posed as exercise. Q.E.D.

There are several different though equivalent definitions of the ε-pseudo-spectrum $\sigma_\varepsilon(T)$ of a matrix $T \in \mathbb{C}^{n\times n}$, which can be taken as starting point for the computation of pseudo-spectra (see Trefethen [21] and Trefethen & Embree [22]). Here, we use the definition contained in Lemma 4.4. Let $\sigma_\varepsilon(T)$ to be determined in a whole section $D \subset \mathbb{C}$. We choose a sequence of grid points $z_i \in D$, $i = 1, 2, 3, \ldots$, and in each z_i determine the smallest ε for which $z_i \in \sigma_\varepsilon(T)$. By interpolating the obtained values, we can then decide whether a point $z \in \mathbb{C}$ approximately belongs to $\sigma_\varepsilon(T)$.

Remark 4.10: The characterization

$$\sigma_\varepsilon(T) = \cup\{\sigma(T+E) \,|\, E \in \mathbb{C}^{n\times n},\ \|E\|_2 \leq \varepsilon\} \tag{4.3.57}$$

leads one to simply take a number of random matrices E of norm less than ε and to plot the union of the usual spectra $\sigma(T+E)$. The resulting pictures are called the "poor man's pseudo-spectra". This approach is rather expensive since in order to obtain precise information of the ε-pseudo-spectrum a really large number of random matrices are needed. It cannot be used for higher-dimensional matrices.

Remark 4.11: The determination of pseudo-spectra in hydrodynamic stability theory requires the solution of eigenvalue problems related to the linearized Navier-Stokes equations as described in Section 0.4.3:

$$\begin{aligned}&-\nu\Delta v + \hat{v}\cdot\nabla v + v\cdot\nabla\hat{v} + \nabla q = \lambda v, \quad \nabla\cdot v = 0, \quad \text{in } \Omega, \\ &v|_{\Gamma_{\text{rigid}}\cup\Gamma_{\text{in}}} = 0, \ \ \nu\partial_n v - qn|_{\Gamma_{\text{out}}} = 0,\end{aligned} \tag{4.3.58}$$

where $\hat{v}$ is the stationary "base flow" the stability of which is to be investigated. This eigenvalue problem is posed on the linear manifold described by the incompressibility constraint $\nabla\cdot v = 0$. Hence after discretization the resulting algebraic eigenvalue problems inherit the saddle-point structure of (4.3.58). We discuss this aspect in the context of a finite element Galerkin discretization with finite element spaces $H_h \subset H_0^1(\Omega)^d$ and $L_h \subset L^2(\Omega)$. Let $\{\varphi_h^i,\ i = 1, \ldots, n_v := \dim H_h\}$ and $\{\chi_h^j,\ j = 1, \ldots, n_p := \dim L_h\}$ be standard nodal bases of the finite element spaces H_h and L_h, respectively. The eigenvector $v_h \in H_h$ and the pressure $q_h \in L_h$ possess expansions $v_h = \sum_{i=1}^{n_v} v_h^i \varphi_h^i$, $q_h = \sum_{j=1}^{n_p} q_h^j \chi_h^j$, where the vectors of expansion coefficients are likewise denoted by $v_h = (v_h^i)_{i=1}^{n_v} \in \mathbb{C}^{n_v}$ and

$q_h = (q_h^j)_{j=1}^{n_p} \in \mathbb{C}^{n_p}$, respectively. With this notation the discretization of the eigenvalue problem (4.3.58) results in a generalized algebraic eigenvalue problem of the form

$$\begin{bmatrix} S_h & B_h \\ B_h^T & 0 \end{bmatrix} \begin{bmatrix} v_h \\ q_h \end{bmatrix} = \lambda_h \begin{bmatrix} M_h & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_h \\ q_h \end{bmatrix}, \tag{4.3.59}$$

with the so-called stiffness matrix S_h, gradient matrix B_h and mass matrix M_h defined by

$$S_h := \big(a'(\hat{v}_h; \varphi_h^j, \varphi_h^i)\big)_{i,j=1}^{n_v}, \quad B_h := \big((\chi_h^j, \nabla \cdot \varphi_h^i)_{L^2}\big)_{i,j=1}^{n_v,n_p}, \quad M_h := \big((\varphi_h^j, \varphi_h^i)_{L^2}\big)_{i,j=1}^{n_v}.$$

For simplicity, we suppress terms stemming from pressure and transport stabilization. The generalized eigenvalue problem (4.3.59) can equivalently be written in the form

$$\begin{bmatrix} M_h & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} S_h & B_h \\ B_h^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} M_h & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_h \\ q_h \end{bmatrix} = \mu_h \begin{bmatrix} M_h & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_h \\ q_h \end{bmatrix}, \tag{4.3.60}$$

where $\mu_h = \lambda_h^{-1}$. Since the pressure q_h only plays the role of a silent variable (4.3.60) reduces to the (standard) generalized eigenvalue problem

$$T_h v_h = \mu_h M_h v_h, \tag{4.3.61}$$

with the matrix $T_h \in \mathbb{R}^{n_v \times n_v}$ defined by

$$\begin{bmatrix} T_h & 0 \\ 0 & 0 \end{bmatrix} := \begin{bmatrix} M_h & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} S_h & B_h \\ B_h^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} M_h & 0 \\ 0 & 0 \end{bmatrix}.$$

The approach described below for computing eigenvalues of general matrices $T \in \mathbb{R}^{n \times n}$ can also be applied to this non-standard situation.

Computation of eigenvalues

For computing the eigenvalues of a (general) matrix $T \in \mathbb{R}^{n \times n}$, we use the *Arnoldi process*, which produces a lower-dimensional Hessenberg matrix the eigenvalues of which approximate those of T:

$$H^{(m)} = \bar{Q}^{(m)T} T Q^{(m)} = \begin{pmatrix} h_{1,1} & h_{1,2} & h_{1,3} & \cdots & h_{1,m} \\ h_{2,1} & h_{2,2} & h_{2,3} & \cdots & h_{2,m} \\ 0 & h_{3,2} & h_{3,3} & \cdots & h_{3,m} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_{m,m-1} & h_{m,m} \end{pmatrix},$$

where the matrix $Q^{(m)} = [q^1, \ldots, q^m]$ is formed with the orthonormal basis $\{q^1, \ldots, q^m\}$ of the Krylov space $K_m = \text{span}\{q, Tq, .., T^{m-1}q\}$. The corresponding eigenvalue problem is then efficiently solved by the QR method using only $\mathcal{O}(m^2)$ operaton. The obtained eigenvalues approximate those eigenvalues of T with largest modulus, which in turn are related to the desired eigenvalues of the differential operator with smallest real parts. Enlarging the dimension m of K_m improves the accuracy of this approximation as well as the number of the approximated "largest" eigenvalues. In fact, the pseudo-spectrum of $H^{(m)}$ approaches that of T for $m \to n$.

The construction of the Krylov space K_m is the most cost-intensive part of the whole process. It requires (m–1)-times the application of the matrix T, which, if T is the inverse of a given system matrix, amounts to the consecutive solution of m linear systems of dimension $n \gg m$. This may be achieved by a multigrid method implemented in available open source software (see Chapter 5). Since such software often does not support complex arithmetic the linear system $Sx = y$ needs to be rewritten in real arithmetic,

$$Sx = y \quad \Leftrightarrow \quad \begin{pmatrix} \text{Re}S & \text{Im}S \\ -\text{Im}S & \text{Re}S \end{pmatrix} \begin{pmatrix} \text{Re}x \\ -\text{Im}x \end{pmatrix} = \begin{pmatrix} \text{Re}y \\ -\text{Im}y \end{pmatrix}.$$

For the reliable approximation of the pseudo-spectrum of T in the subregion $D \subset \mathbb{C}$ it is necessary to choose the dimension m of the Krylov space sufficiently large, such that all eigenvalues of T and its perturbations located in D are well approximated by eigenvalues of $H^{(m)}$. Further, the QR method is to be used with maximum accuracy requiring the corresponding error tolerance TOL to be set in the range of the machine accuracy. An eigenvector w corresponding to an eigenvalue $\lambda \in \sigma(H^{(m)})$ is then obtained by solving the singular system

$$(H^{(m)} - \lambda I)w = 0. \tag{4.3.62}$$

By back-transformation of this eigenvector from the Krylov space K_m into the space $\mathbb{R}^n$, we obtain a corresponding approximate eigenvector of the full matrix T.

Practical computation of the pseudospectrum

We want to determine the "critical" part of the ε-pseudo-spectrum of the discrete operator A_h, which approximates the unbounded differential operator A. As discussed above, this requires the computation of the smallest singular value of the inverse matrix $T = A_h^{-1}$. Since the dimension n_h of T in practical applications is very high, $n_h \approx 10^4 - 10^8$, the direct computation of singular values of T or even a full singular value decomposition is prohibitively expensive. Therefore, the first step is the reduction of the problem to lower dimension by projection onto a Krylov space resulting in a (complex) Hessenberg matrix $H^{(m)} \in \mathbb{C}^{n\times n}$ the inverse of which, $H^{(m)-1}$, may then be viewed as a low-dimensional approximation to A_h capturing the critical "smallest" eigenvalues of A_h and likewise its pseudo-spectra. The pseudo-spectra of $H^{(m)}$ may then be computed using the approach described in Section 4.2.2. By Lemma 1.17 the pseudo-spectrum of $H^{(m)}$ is closely related

to that of $H^{(m)-1}$ but involving constants, which are difficult to control. Therefore, one tends to prefer to directly compute the pseudo-spectra of $H^{(m)-1}$ as an approximation to that of A_h. This, however, is expensive for larger m since the inversion of the matrix $H^{(m)}$ costs $\mathcal{O}(m^3)$ operations. Dealing directly with the Hessenberg matrix $H^{(m)}$ looks more attractive. Both procedures are discussed in the following. We choose a section $D \subset \mathbb{C}$ (around the origin), in which we want to determine the pseudo-spectrum. Let $D := \{z \in \mathbb{C} |\ \{\operatorname{Re} z, \operatorname{Im} z\} \in [a_r, b_r] \times [a_i, b_i]\}$ for certain values $a_r < b_r$ and $a_i < b_i$. To determine the pseudo-spectrum in the complete rectangle D, we cover D by a grid with spacing d_r and d_i, such that k points lie on each grid line. For each grid point, we compute the corresponding ε-pseudo-spectrum.

i) Computation of the pseudo-spectra $\sigma_\varepsilon(H^{(m)-1})$: For each $z \in D \setminus \sigma(H^{(m)-1})$ the quantity

$$\varepsilon(z, H^{(m)-1}) := \|(zI - H^{(m)-1})^{-1}\|_2^{-1} = \sigma_{\min}(zI - H^{(m)-1})$$

determines the smallest $\varepsilon > 0$, such that $z \in \sigma_\varepsilon(H^{(m)-1})$. Then, for any point $z \in D$, by computing $\sigma_{\min}(zI - H^{(m)-1})$, we obtain an approximation of the smallest ε, such that $z \in \sigma_\varepsilon(H^{(m)-1})$. For computing $\sigma_{\min} := \sigma_{\min}(zI - H^{(m)-1})$, we recall its definition as smallest (positive) eigenvalue of the Hermitian, positive definite matrix

$$S := (\overline{zI - H^{(m)-1}})^T (zI - H^{(m)-1})$$

and use the "inverse iteration", $z^0 \in \mathbb{C}^n$, $\|z^0\|_2 = 1$,

$$t \geq 1: \quad S\tilde{z}^t = z^{t-1}, \quad z^t = \|\tilde{z}^t\|_2^{-1} \tilde{z}^t, \quad \sigma_{\min}^t := (Sz^t, z^t)_2. \tag{4.3.63}$$

The linear systems in each iteration can be solved by pre-computing either directly an LR decomposition of S, or if this is too ill-conditioned, first a QR decomposition

$$zI - H^{(m)-1} = QR,$$

which then yields a Cholesky decomposition of S:

$$S = (\overline{QR})^T QR = \bar{R}^T \bar{Q}^T QR = \bar{R}^T R. \tag{4.3.64}$$

This preliminary step costs another $\mathcal{O}(m^3)$ operations.

ii) Computation of the pseudo-spectra $\sigma_\varepsilon(H^{(m)})$: Alternatively, one may compute a singular value decomposition of the Hessenberg matrix $zI - H^{(m)}$,

$$zI - H^{(m)} = U \Sigma \bar{V}^T,$$

where $U, V \in \mathbb{C}^{n \times n}$ are unitary matrices and $\Sigma = \operatorname{diag}\{\sigma_i,\ i = 1, \ldots, n\}$. Then,

$$\sigma_{\min}(zI - H^{(m)}) = \min\{\sigma_i,\ i = 1, \ldots, m\}.$$

For that, we use the LAPACK routine *dgesvd* within MATLAB. Since the operation count of the singular value decomposition growth like $\mathcal{O}(m^2)$, in our sample calculation, we limit the dimension of the Krylov space by $m \leq 200$.

Choice of parameters and accuracy issues

The described algorithm for computing the pseudo-spectrum of a differential operator at various stages requires the appropriate choice of parameters:

- The mesh size h in the finite element discretization on the domain $\Omega \subset \mathbb{R}^n$ for reducing the infinite dimensional problem to an matrix eigenvalue problem of dimension n_h.
- The dimension of the Krylov space $K_{m,h}$ in the Arnoldi method for the reduction of the n_h-dimensional (inverse) matrix T_h to the much smaller Hessenberg matrix $H_h^{(m)}$.
- The size of the subregion $D := [a_r, b_r] \times [a_i, b_i] \subset \mathbb{C}$ in which the pseudospectrum is to be determined and the mesh width k of interpolation points in $D \subset \mathbb{C}$.

Only for an appropriate choice of these parameters one obtains a reliable approximation to the pseudo-spectrum of the differential operator A. First, h is refined and m is increased until no significant change in the boundaries of the ε-pseudo-spectrum is observed anymore.

Example 1. Sturm-Liouville eigenvalue problem

As a prototypical example for the proposed algorithm, we consider the Sturm-Liouville boundary value problem (see Trefethen [21])

$$Au(x) = -u''(x) - q(x)u(x), \quad x \in \Omega = (-10, 10), \tag{4.3.65}$$

with the complex potential $q(x) := (3 + 3\mathbf{i})x^2 + \frac{1}{16}x^4$, and the boundary condition $u(-10) = 0 = u(10)$. Using the sesquilinear form

$$a(u, v) := (u', v') + (qu, v), \quad u, v \in H_0^1(\Omega),$$

the eigenvalue problem of the operator A reads in variational form

$$a(v, \varphi) = \lambda(v, \varphi) \quad \forall \varphi \in H_0^1(\Omega). \tag{4.3.66}$$

First, the interval $\Omega = (-10, 10)$ is discretized by eightfold uniform refinement resulting in the finest mesh size $h = 20 \cdot 2^{-8} \approx 0.078$ and $n_h = 256$. The Arnoldi algorithm for the corresponding discrete eigenvalue problem of the inverse matrix A_h^{-1} generates a Hessenberg matrix $H_h^{(m)}$ of dimension $m = 200$. The resulting reduced eigenvalue problem is solved by the QR method. For the determination of the corresponding pseudo-spectra, we export the Hessenberg matrix $H_h^{(m)}$ into a MATLAB file. For this, we use the routine *DGESVD* in *LAPACK* (singular value decomposition) on meshes with 10×10 and with 100×100 points. The ε-pseudo-spectra are computed for $\varepsilon = 10^{-1}, 10^{-2}, ..., 10^{-10}$ leading to the results shown in Fig. 4.1. We observe that all eigenvalues have negative real

part but also that the corresponding pseudo-spectra reach far into the positive half-plane of $\mathbb{C}$, i. e., small perturbations of the matrix may have strong effects on the location of the eigenvalues. Further, we see that already a grid with 10×10 points yields sufficiently good approximations of the pseudo-spectrum of the matrix $H_h^{(m)}$.

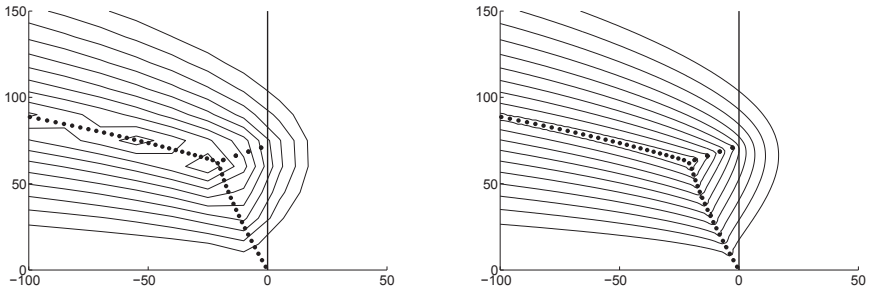

Figure 4.1: *Approximate eigenvalues and pseudo-spectra of the operator A computed from those of the inverse matrix A_h^{-1} on a 10×10 grid (left) and on a 100×100 grid (right): "dots" represent eigenvalues and the lines the boundaries of the ε-pseudo-spectra for $\varepsilon = 10^{-1}, ..., 10^{-10}$.*

Example 2. Stability eigenvalue problem of the Burgers operator

A PDE test example is the two-dimensional *Burgers equation*

$$-\nu\Delta v + v \cdot \nabla v = 0, \quad \text{in } \Omega. \tag{4.3.67}$$

This equation is sometimes considered as a simplified version of the Navier-Stokes equation since both equations contain the same nonlinearity. We use this example for investigating some questions related to the numerical techniques used, e. g., the required dimension of the Krylov spaces in the Arnoldi method.

For simplicity, we choose $\Omega := (0,2) \times (0,1) \subset \mathbb{R}^2$, and along the left-hand "inflow boundary" $\Gamma_{\text{in}} := \partial\Omega \cap \{x_1 = 0\}$ as well as along the upper and lower boundary parts $\Gamma_{\text{rigid}} := \partial\Omega \cap (\{x_2 = 0\} \cup \{x_2 = 1\})$ Dirichlet conditions and along the right-hand "outflow boundary" $\Gamma_{\text{out}} := \partial\Omega \cap \{x_1 = 2\}$ Neumann conditions are imposed, such that the exact solution has the form $\hat{v}(x) = (x_2, 0)$ of a Couette-like flow. We set $\Gamma_D := \Gamma_{\text{rigid}} \cup \Gamma_{\text{in}}$ and choose $\nu = 10^{-2}$. Linearization around this stationary solution yields the nonsymmetric stability eigenvalue problem for $v = (v_1, v_2)$:

$$\begin{aligned} -\nu\Delta v_1 + x_2\partial_1 v_1 + v_2 &= \lambda v_1, \\ -\nu\Delta v_2 + x_2\partial_1 v_2 \qquad &= \lambda v_2, \end{aligned} \tag{4.3.68}$$

in Ω with the boundary conditions $v|_{\Gamma_D} = 0$, $\partial_n v|_{\Gamma_{\text{out}}} = 0$. For discretizing this problem, we use the finite element method described above with conforming Q_1-elements combined

with transport stabilization by the SUPG (streamline upwind Petrov-Galerkin) method. We investigate the eigenvalues of the linearized (around Couette flow) Burgers operator with Dirichlet or Neumann inflow conditions. We use the Arnoldi method described above with Krylov spaces of dimension $m = 100$ or $m = 200$. For generating the contour lines of the ε-pseudospectra, we use a grid of 100×100 test points.

For testing the accuracy of the proposed method, we compare the quality of the pseudo-spectra computed on meshes of width $h = 2^{-7}$ ($n_h \approx 30,000$) and $h = 2^{-8}$, ($n_h \approx 130,000$) and using Krylov spaces of dimension $m = 100$ or $m = 200$. The results shown in Fig. 4.2 and Fig. 4.3 indicate that the choice $h = 2^{-7}$ and $m = 100$ is sufficient for the present example.

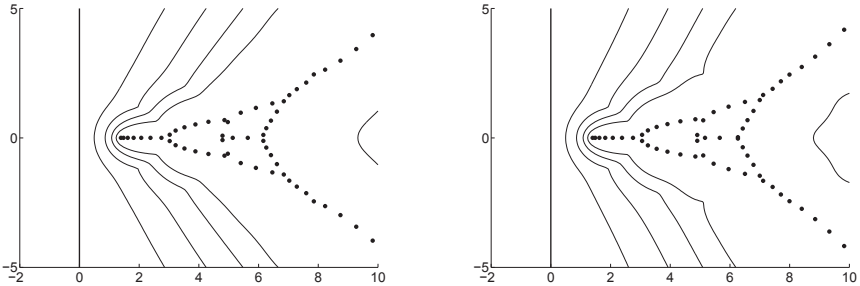

Figure 4.2: *Computed pseudo-spectra of the linearized Burgers operator with Dirichlet inflow condition for* $\nu = 0.01$ *and* $h = 2^{-7}$ *(left) and* $h = 2^{-8}$ *(right) computed by the Arnoldi method with* $m = 100$. *The "dots" represent eigenvalues and the lines the boundaries of the* ε*-pseudo-spectra for* $\varepsilon = 10^{-1}, ..., 10^{-4}$.

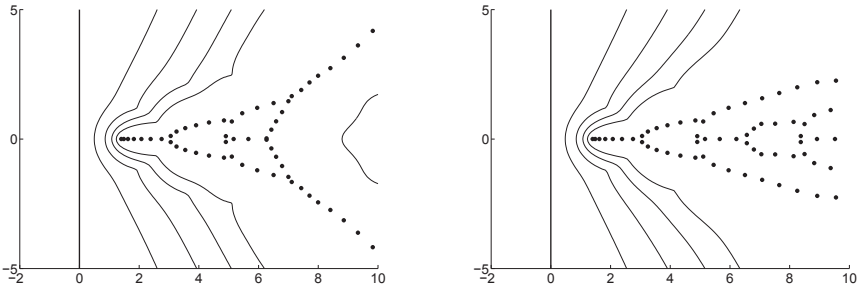

Figure 4.3: *Computed pseudo-spectra of the linearized Burgers operator with Dirichlet inflow condition for* $\nu = 0.01$ *and* $h = 2^{-8}$ *computed by the Arnoldi method with* $m = 100$ *(left) and* $m = 200$ *(right). The "dots" represent eigenvalues and the lines the boundaries of the* ε*-pseudo-spectra for* $\varepsilon = 10^{-1}, ..., 10^{-4}$.

Now, we turn to Neumann inflow conditions. In this particular case the first eigenval-

ues and eigenfunctions of the linearized Burgers operator can be determined analytically as $\lambda_k = \nu k^2\pi^2$, $v_k(x) = (\sin(k\pi x_2), 0)^T$, for $k \in \mathbb{Z}$. All these eigenvalues are degenerate. However, there exists another eigenvalue $\lambda_4 \approx 1.4039$ between the third and fourth one, which is not of this form, but also degenerate.

We use this situation for studying the dependence of the proposed method for computing pseudo-spectra on the size of the viscosity parameter, $0.001 \le \nu \le 0.01$. Again the discretization uses the mesh size $h = 2^{-7}$, Krylov spaces of dimension $m = 100$ and a grid of spacing $k = 100$. By varying these parameters, we find that only eigenvalues with $\mathrm{Re}\lambda \le 6$ and corresponding ε-pseudo-spectra with $\varepsilon \ge 10^{-4}$ are reliably computed. The results are shown in Fig. 4.4.

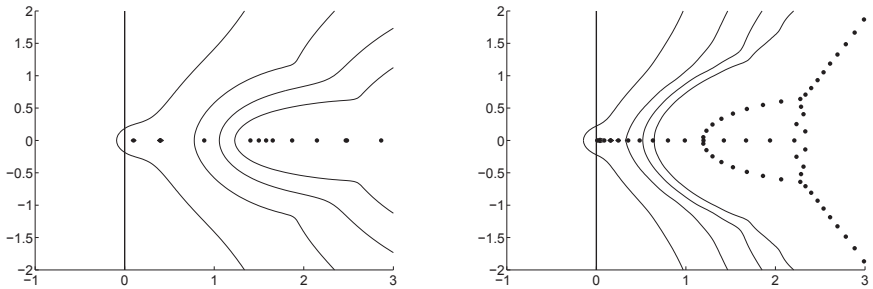

Figure 4.4: *Computed pseudos-pectra of the linearized (around Couette flow) Burger operator with Neumann inflow conditions for* $\nu = 0.01$ *(left) and* $\nu = 0.001$ *(right): The dots represent eigenvalues and the lines the boundaries of the* ε*-pseudo-spectra for* $\varepsilon = 10^{-1}, \ldots, 10^{-4}$.

For Neumann inflow conditions the most critical eigenvalue is significantly smaller than the corresponding most critical eigenvalue for Dirichlet inflow conditions, which suggests weaker stability properties in the "Neumann case". Indeed, in Fig. 4.4, we see that the 0.1-pseudo-spectrum reaches into the negative complex half-plane indicating instability for such perturbations. This effect is even more pronounced for $\nu = 0.001$ with $\lambda^N_{\mathrm{crit}} \approx 0.0098$.

Example 3. Stability eigenvalue problem of the Navier-Stokes operator

In this last example, we present some computational results for the 2d Navier-Stokes benchmark "channel flow around a cylinder" with the configuration shown in Section 0.4.3 (see Schäfer & Turek [65]). The geometry data are as follows: channel domain $\Omega := (0.00\mathrm{m}, 2.2\mathrm{m}) \times (0.00\mathrm{m}, 0.41\mathrm{m})$, diameter of circle $D := 0.10\mathrm{m}$, center of circle at $a := (0.20\mathrm{m}, 0.20\mathrm{m})$ (slightly nonsymmetric position). The Reynolds number is defined in terms of the diameter D and the maximum inflow velocity $\bar{U} = \max|v^{\mathrm{in}}| = 0.3\mathrm{m/s}$ (parabolic profile), $\mathrm{Re} = \bar{U}^2 D/\nu$. The boundary conditions are $v|_{\Gamma_{\mathrm{rigid}}} = 0$, $v|_{\Gamma_{\mathrm{in}}} = v^{\mathrm{in}}$, $\nu\partial_n v - np|_{\Gamma_{\mathrm{out}}} = 0$. The viscosity is chosen such that the Reynolds number is small enough, $20 \le \mathrm{Re} \le 40$, to guarantee stationarity of the base flow as shown in Fig. 4.5. Already for $\mathrm{Re} = 60$ the flow turns nonstationary (time periodic).

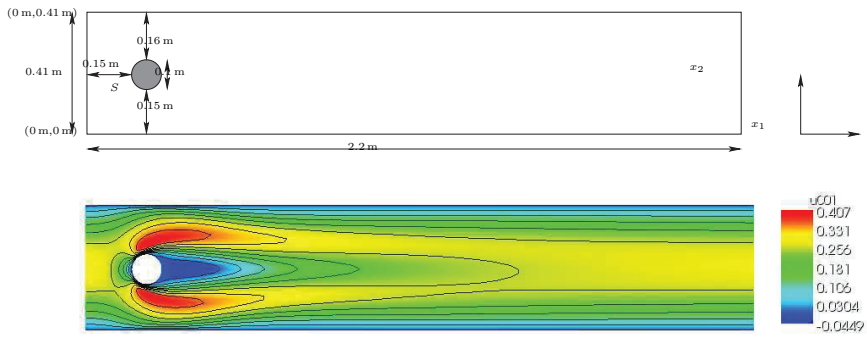

Figure 4.5: *Configuration of the "channel flow" benchmark and x_1-component of the velocity for* $\mathrm{Re} = 40$.

We want to investigate the stability of the computed base flow for several Reynolds numbers in the range $20 \leq \mathrm{Re} \leq 60$ and inflow conditions imposed on the admissible perturbations, Dirichlet or Neumann ("free"), by determining the corresponding critical eigenvalues and pseudo-spectra. This computation uses a "stationary code" employing the Newton method for linearization, which is known to potentially yield stationary solutions even at Reynolds numbers for which such solutions may not be stable.

Perturbations satisfying Dirichlet inflow conditions

We begin with the case of perturbations satisfying (homogeneous) Dirichlet inflow conditions. The pseudo-spectra of the critical eigenvalues for $\mathrm{Re} = 40$ and $\mathrm{Re} = 60$ are shown in Fig. 4.6.

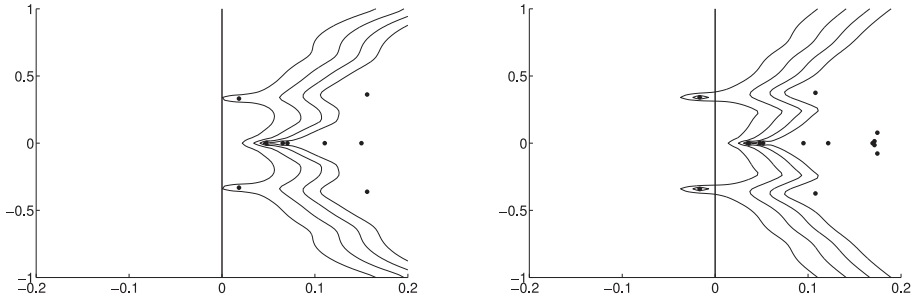

Figure 4.6: *Computed pseudo-spectra of the linearized Navier-Stokes operator ("channel flow" benchmark) for different Reynolds numbers,* $\mathrm{Re} = 40$ *(left) and* $\mathrm{Re} = 60$ *(right), with* **Dirichlet** *inflow condition: The "dots" represent eigenvalues and the lines the boundaries of ε-pseudospectra for* $\varepsilon = 10^{-2}, 10^{-2.5}, 10^{-3}, 10^{-3.5}$.

The computation has been done on meshes obtained by four to five uniform refinements of the (locally adapted) meshes used for computing the base flow. In the Arnoldi method, we use Krylov spaces of dimension $m = 100$. Computations with $m = 200$ give almost the same results. For $\mathrm{Re} = 40$ the relevant 10^{-2}-pseudo-spectrum does not reach into the negative complex half-plane indicating stability of the corresponding base solution in this case, as expected in view of the result of nonstationary computations. Obviously the transition from stationary to nonstationary (time periodic) solutions occurs in the range $40 \leq \mathrm{Re} \leq 60$. However, for this "instability" the sign of the real part of the critical eigenvalue seems to play the decisive role and not so much the size of the corresponding pseudo-spectrum.

Perturbations satisfying Neumann (free) inflow conditions

Next, we consider the case of perturbations satisfying (homogeneous) Neumann ("free") inflow conditions, i. e., the space of admissible perturbations is larger than in the preceding case. In view of the observations made before for Couette flow and Poiseuille flow, we expect weaker stability properties. The stationary base flow is again computed using Dirichlet inflow conditions but the associated eigenvalue problem of the linearized Navier-Stokes operator is considered with Neumann inflow conditions. In the case of perturbations satisfying Dirichlet inflow conditions the stationary base flow turned out to be stable up to $\mathrm{Re} = 45$. In the present case of perturbations satisfying Neumann inflow conditions at $\mathrm{Re} = 40$ the critical eigenvalue has positive but very small real part, $\mathrm{Re}\lambda_{\min} \approx 0.003$. Hence, the precise stability analysis requires the determination of the corresponding pseudo-spectrum. The results are shown in Fig. 4.3.2. Though, for $\mathrm{Re} = 40$ the real part of the most critical (positive) eigenvalue is rather small, the corresponding 10^{-2}-pseudo-spectrum reaches only a little into the negative complex half-plane.

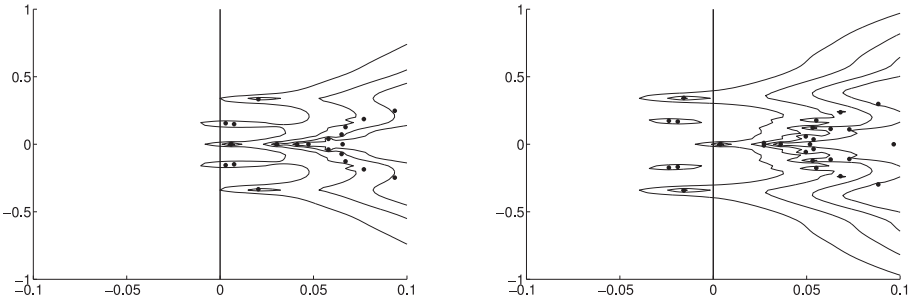

Figure 4.7: *Computed pseudo-spectra of the linearized Navier-Stokes operator ("channel flow") with* **Neumann** *inflow conditions for different Reynolds numbers,* $\mathrm{Re} = 40$ *(left) and* $\mathrm{Re} = 60$ *(right): The "dots" represent eigenvalues and the lines the boundaries of the* ε*-pseudospectra for* $\varepsilon = 10^{-2}, 10^{-2.5}, 10^{-3}, 10^{-3.5}$.

4.4 Exercises

Exercise 4.1: The proof of convergence of the "power method" applied to a symmetric, positive definite matrix $A \in \mathbb{R}^{n\times n}$ resulted in the identity

$$\lambda^t = (Az^t, z^t)_2 = \frac{(\lambda_n)^{2t+1}\left\{|\alpha_n|^2 + \sum_{i=1}^{n-1}|\alpha_i|^2\left(\frac{\lambda_i}{\lambda_n}\right)^{2t+1}\right\}}{(\lambda_n)^{2t}\left\{|\alpha_n|^2 + \sum_{i=1}^{n-1}|\alpha_i|^2\left(\frac{\lambda_i}{\lambda_n}\right)^{2t}\right\}} = \lambda_{\max} + \mathcal{O}\Big(\Big|\frac{\lambda_{n-1}}{\lambda_{\max}}\Big|^{2t}\Big),$$

where $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, n$, are the eigenvalues of A, $\{w^i,\, i = 1, \ldots, n\}$ a corresponding ONB of eigenvectors and α_i the coefficients in the expansion of the starting vector $z^0 = \sum_{i=1}^{n}\alpha_i w^i$. Show that, in case $\alpha_n \neq 0$, in the above identity the error term on the right-hand side is uniformly bounded with respect to the dimension n of A but depends linearly on $|\lambda_n|$.

Exercise 4.2: The "inverse iteration" may be accelerated by employing a dynamic "shift" taken from the presceding eigenvalue approximation ($\lambda_k^0 \approx \lambda_k$):

$$(A - \lambda_k^{t-1}I)\tilde{z}^t = z^{t-1}, \quad z^t = \frac{\tilde{z}^{t-1}}{\|\tilde{z}^{t-1}\|}, \quad \mu_k^t = ((A - \lambda_k^{t-1}I)^{-1}z^t, z^t)_2, \quad \lambda_k^t = \frac{1}{\mu_k^t} + \lambda_k^{t-1}.$$

Investigate the convergence of this method for the computation of the smallest eigenvalue $\lambda_1 = \lambda_{\min}$ of a symmetric, positive definite matrix $A \in \mathbb{R}^{n\times n}$. In detail, show the convergence estimate

$$|\lambda_1 - \lambda^t| \leq |\lambda^t - \lambda^{t-1}| \prod_{j=0}^{t-1}\Big|\frac{\lambda_1 - \lambda^j}{\lambda_2 - \lambda^j}\Big|^2 \frac{\|z^0\|_2^2}{|\alpha_1|^2}.$$

(Hint: Show that

$$\mu^t = \frac{\sum_{i=1}^{n}|\alpha_i|^2(\lambda_i - \lambda^{t-1})^{-1}\prod_{j=0}^{t-1}(\lambda_i - \lambda^j)^{-2}}{\sum_{i=1}^{n}|\alpha_i|^2\prod_{j=0}^{t-1}(\lambda_i - \lambda^j)^{-2}}$$

and proceed in a similar way as in the preceding exercise.)

Exercise 4.3: Let A be a Hessenberg matrix or a symmetric tridiagonal matrix. Show that in this case the same holds true for all iterates A^t generated by the QR method:

$$A^{(0)} := A,$$
$$A^{(t+1)} := R^{(t)}Q^{(t)}, \text{ with } A^{(t)} = Q^{(t)}R^{(t)}, \quad t \geq 0.$$

Exercise 4.4: Each matrix $A \in \mathbb{C}^{n\times n}$ possesses a QR decomposition $A = QR$, with a unitary matrix $Q = [q^1, \ldots, q^n]$ and an upper triangular matrix $R = (r_{ij})_{i,j=1}^n$. Clearly, this decomposition is not uniquely determined. Show that for regular A there exists a uniquely determined QR decomposition with the property $r_{ii} \in \mathbb{R}_+$, $i = 1, \ldots, n$.

(Hint: Use the fact that the QR decomposition of A yields a Cholesky decomposition of the matrix $\bar{A}^T A$.)

Exercise 4.5: For a matrix $A \in \mathbb{C}^{n\times n}$ and an arbitrary vector $q \in \mathbb{C}^n$, $q \neq 0$, form the Krylov spaces $K_m := \mathrm{span}\{q, Aq, \ldots, A^{m-1}q\}$. Suppose that for some $1 \leq m \leq n$ there holds $K_{m-1} \neq K_m = K_{m+1}$.

i) Show that then $K_m = K_{m+1} = \cdots = K_n = \mathbb{C}^n$ and $\dim K_m = m$.

ii) Let $\{q^1, \ldots, q^m\}$ be an ONB of K_m and set $Q^m := [q^1, \ldots, q^m]$. Show that there holds $\sigma(Q^{mT} A Q^m) \subset \sigma(A)$. In the case $m = n$ there holds $\sigma(Q^{nT} A Q^n) = \sigma(A)$.

Exercise 4.6: Recall the two versions of the Gram-Schmidt algorithm, the "classical" one and the "modified" one described in the text, for the successive orthogonalization of a general, linear independent set $\{v^1, \ldots, v^m\} \subset \mathbb{R}^n$.

i) Verify that both algorithms, used with exact arithmetic, yield the same result.

ii) Determine the computational complexity of these two algorithms, i. e., the number of arithmetic operations for computing the corresponding orthonormal set $\{u^1, \ldots, u^m\}$.

Exercise 4.7: Consider the nearly singular 3×3-matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ \varepsilon & \varepsilon & 0 \\ \varepsilon & 0 & \varepsilon \end{bmatrix} = [a^1, a^2, a^3],$$

where $\varepsilon > 0$ is small enough so that $1 + \varepsilon^2$ is rounded to 1 in the given floating-point arithmetic. Compute the QR decomposition of $A = [a^1, a^2, a^3]$ by orthonormalization of the set of its column vectors $\{a^1, a^2, a^3\}$ using (i) the *classical* Gram-Schmidt algorithm and (ii) its *modified* version. Compare the quality of the results by making the "Householder Test": $\|Q^T Q - I\|_\infty \approx 0$.

Exercise 4.8: Consider the model eigenvalue problem from the text, which originates from the 7-point discretization of the Poisson problem on the unit cube:

$$A = h^{-2} \underbrace{\begin{bmatrix} B & -I_{m^2} & \\ -I_{m^2} & B & \ddots \\ & \ddots & \ddots \end{bmatrix}}_{n=m^3} \qquad B = \underbrace{\begin{bmatrix} C & -I_m & \\ -I_m & C & \ddots \\ & \ddots & \ddots \end{bmatrix}}_{m^2} \qquad C = \underbrace{\begin{bmatrix} 6 & -1 & \\ -1 & 6 & \ddots \\ & \ddots & \ddots \end{bmatrix}}_{m}$$

where $h = 1/(m+1)$ is the mesh size. In this case the corresponding eigenvalues and eigenvectors are explicitly given by

$$\lambda_{ijk}^h = h^{-2}\big\{6 - 2\big(\cos[ih\pi] + \cos[jh\pi] + \cos[kh\pi]\big)\big\}, \quad i, j, k = 1, \ldots, m,$$
$$w_h^{ijk} = \big(\sin[pih\pi]\, \sin[qjh\pi]\, \sin[rkh\pi]\big)_{p,q,r=1}^m.$$

For this discretization, there holds the theoretical a priori error estimate

$$\frac{|\lambda_{ijk} - \lambda_{ijk}^h|}{|\lambda_{ijk}|} \leq \frac{1}{12}\lambda_{ijk} h^2,$$

where $\lambda_{ijk} = (i^2 + j^2 + k^2)\pi^2$ are the exact eigenvalues of the Laplace operator (and h sufficiently small).

i) Verify this error estimate using the given values for λ_{ijk} and λ_{ijk}^h.

ii) Derive an estimate for the number of eigenvalues (*not* counting multiplicities) of the Laplace operator that can be approximated reliably for a mesh size of $h = 2^{-7}$, if a uniform *relative* accuracy of $\mathrm{TOL} = 10^{-3}$ is required.

iii) How small has the mesh size h to be chosen if the first 1.000 eigenvalues (*counting multiplicities* for simplicity) of the Laplace operator have to be computed with relative accuracy $\mathrm{TOL} = 10^{-3}$? How large would the dimension n of the resulting system matrix A be in this case? (**Hint:** We are interested in an upper bound, so simplify accordingly.)

Exercise 4.9: Formulate the "inverse iteration" of Wielandt and the "Lanczos algorithm" (combined with the QR method) for computing the smallest eigenvalue of a large symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$. Suppose that matrix vector products as well as solving linear systems occurring in these processes can be accomplished with $\mathcal{O}(n)$ a. op.:

i) Compare the arithmetic work (# of a. op.) of these two approaches for performing 100 iterations.

ii) How do the two methods compare if not only the smallest but the 10 smallest eigenvalues are to be computed?

Exercise 4.10: The Krylov space method applied for general matrices $A \in \mathbb{C}^{n\times n}$ requires complex arithmetic, but many software packages provide only real arithmetic.

i) Verify that a (complex) linear system $Ax = b$ can equivalently be written in the following real "$(2n \times 2n)$-block form":

$$\begin{pmatrix} \mathrm{Re}\, A & \mathrm{Im}\, A \\ -\mathrm{Im}\, A & \mathrm{Re}\, A \end{pmatrix} \begin{pmatrix} \mathrm{Re}\, x \\ -\mathrm{Im}\, x \end{pmatrix} = \begin{pmatrix} \mathrm{Re}\, b \\ -\mathrm{Im}\, b \end{pmatrix}.$$

ii) Formulate (necessary and sufficient) conditions on A, which guarantee that this (real) coefficient block-matrix is a) regular, b) symmetric and c) positive definite?

Exercise 4.11: Show that the ε-pseudo-spectrum $\sigma_\varepsilon(T)$ of a matrix $T \in \mathbb{C}^{n\times n}$ is invariant under orthonormal transformations, i. e., for any unitary matrix $Q \in \mathbb{C}^{n\times n}$ there holds

$$\sigma_\varepsilon(T) = \sigma_\varepsilon(Q^{-1}TQ).$$

(Hint: Pick a suitable one of the many equivalent definitions of ε-pseudo-spectrum provided in the text.)