Course Material for the
Autumn School within PhD-Net

Hanoi 2010

Numerical Linear Algebra

Christian Mehl,
TU Berlin,

July 15, 2011

Literature

The material of this course is based on the following textbooks:

- G. Golub, C. Van Loan. *Matrix computations.* Baltimore, 1996.
- R. B. Lehoucq, D. C. Sorensen and C. Yang. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, Philadelphia, 1989.
- Y. Saad. *Numerical methods for large eigenvalue problems.* Manchester, 1992.
- L. Trefethen, D. Bau. *Numerical linear algebra.* Philadelphia, 1997.
- D. Watkins. *Fundamentals of matrix computations.* New York, 2002.

The following books are also useful to complement the material of these notes:

- J. Demmel. *Applied numerical linear algebra.* Philadelphia, 1997.
- N.J. Higham. *Accuracy and stability of numerical algorithms.* Philadelphia, 2002.
- G.W. Stewart. *Matrix algorithms.* Philadelphia, 1998-2001, 2 Volumes.
- G.W. Stewart, J.G. Sun. *Matrix perturbation theory.* Boston, 1990.

Chapter 0

Introduction

The main topics of *Numerical Linear Algebra* are the solution of different classes of *eigenvalue problems* and *linear systems.*
For the eigenvalue problem we discuss different classes.

(a) The *standard eigenvalue problem*: For a real or complex matrix $A \in \mathbb{C}^{n,n}$, determine $x \in \mathbb{C}^n, \lambda \in \mathbb{C}$, such that

$$Ax = \lambda x.$$

In many applications the coefficient matrices have extra properties such as being real and symmetric or complex Hermitian.
For linear systems:

$$Ax = b, x \in \mathbb{C}^n, b \in \mathbb{C}^m$$

with $A \in \mathbb{C}^{m,n}$ we again may extra properties for the coefficient matrices.
We will concentrate in this course on the numerical solution of standard and generalized eigenvalue problems and the solution of linear systems.
Applications: Eigenvalue problems arise in

- the vibrational analysis of structures and vehicles (classical mechanics);
- the analysis of the spectra and energy levels of atoms and molecules (quantum mechanics);
- model reduction techniques, where a large scale model is reduced to a small scale model by leaving out weakly important parts;
- many other applications.

Linear systems arise in almost any area of science and engineering such as

(a) frequency response analysis for excited structures and vehicles;

(b) finite element methods or finite difference methods for ordinary and partial differential equations;

(c) data mining, information retrieval;

(d) and many others.

We will distinguish between small and medium class problems where the full matrices fit into main memory, these are of today sizes $n = 10^2 - 10^5$ and large sparse problems, where the coefficient matrices are stored in sparse formats, and have sizes $n = 10^6$ and larger. We will mainly discuss the case of complex matrices. Many results hold equally well in the real case, but often the presentation becomes more clumsy. We will point out when the real case is substantially different.
We will discuss the following algorithms.

	A small	A large
EVP	QR-Algorithm	Lanczos, Arnoldi
LS		CG, GMRES

Chapter 1

Matrix theory

1.1 Basics

1.1.1 Eigenvalues and Eigenvectors

Let $A \in \mathbb{C}^{n,n}$, then $v \in \mathbb{C}^n \setminus \{0\}$ and $\lambda \in \mathbb{C}$ that satisfy

$$Av = \lambda v$$

are called *eigenvector and eigenvalue* of A.
The set

$$\sigma(A) \quad := \quad \{\lambda \in \mathbb{C} |\ \lambda \text{ eigenvalue of } A\}$$

is called *spectrum of* A.

1.1.2 Matrix norms

Let $A \in \mathbb{C}^{m,n}$, then

$$||A||_p := \sup_{x \neq 0} \frac{||Ax||_p}{||x||_p}$$

is the *matrix p-norm*, $p \in \mathbb{N} \cup \{\infty\}$ and for invertible matrices A

$$\kappa_p(A) := ||A||_p \cdot ||A^{-1}||_p$$

is called the *p-norm condition number of* A.

Special cases:

(a) $p = 1 \rightsquigarrow$ the column-sum norm:

$$||A||_1 = \max_j \sum_{i=1}^{m} |a_{ij}|$$

(b) $p = \infty \rightsquigarrow$ the row-sum norm:

$$||A||_\infty = ||A^T||_1$$

(c) $p = 2 \rightsquigarrow$ the spectral norm

$$||A||_2 = \text{square root of the largest eigenvalue of } A^*A$$

with $A^* = \overline{A}^T$.

Convention:

$$||A|| = ||A||_2, \quad \kappa(A) = \kappa_2(A)$$

1.1.3 Isometric and unitary matrices

Definition 1 *Let $U \in \mathbb{C}^{m,n}, m \geq n$.*

(a) U is called isometric *if $U^*U = I_k$;*

(b) U is called unitary *if U is isometric and $n = k$.*

Theorem 2 *Let $U \in \mathbb{C}^{n \times k}, k \leq n$. Then the following are equivalent.*

(a) U is isometric;

(b) the columns of U are orthonormal;

(c) $\langle Ux, Uy \rangle = \langle x, y \rangle$ for all $x, y \in \mathbb{C}^k$ ($\langle \cdot, \cdot \rangle$: standard real or complex scalar product);

(d) $||Ux|| = ||x||$ for all $x \in \mathbb{C}^k$;

For $k = n$, (a)-(d) are equivalent to

(e) $UU^* = I_n$;

(f) $U^{-1} = U^*$;

(g) the rows of U are orthonormal.

In this case, furthermore,

$$||U|| = 1 = ||U^{-1}|| = \kappa(U).$$

1.1.4 Subspaces

Definition 3 *A space $\mathcal{U} \subset \mathbb{C}^n$ is called subspace, if for all $x, y \in \mathcal{U}, \alpha \in \mathbb{C}$ we have*

$$x + y \in \mathcal{U},\ \alpha x \in \mathcal{U}.$$

Theorem 4 *Let $\mathcal{U} \subset \mathbb{C}^n$ be a subspace with basis $(x_1, \ldots, x_m)$ and $X = [x_1, \ldots, x_m]$, i.e. $Rank(X) = m$.*

(a) Then $\mathcal{U} = \mathcal{R}(X) := \{Xy |\ y \in \mathbb{C}^m\}$ (Range or column space of X).

(b) Let $Y \in \mathbb{C}^{n,m}$ with $Rank(Y) = m$, then

$$\mathcal{R}(X) = \mathcal{R}(Y) \Leftrightarrow X = YB,\ B \in \mathbb{C}^{m,m}.$$

In particular then B is invertible and

$$XB^{-1} = Y.$$

(c) The Gram-Schmidt method for $(x_1, \ldots, x_m)$ delivers an orthonormal basis $(q_1, \ldots, q_m)$ of $\mathcal{U}$ with

$$Span\{q_1, \ldots, q_j\} = Span\{x_1, \ldots, x_j\}$$

for $j = 1, \ldots, m$. This condition is equivalent to:

There exists an upper triangular matrix $R \in \mathbb{C}^{m,m}$ with $X = QR$ where $Q = [q_1, \ldots, q_m]$. (QR-decomposition)

1.1.5 Invariant subspaces

Definition 5 *Let $A \in \mathbb{C}^{n,n}$ and $\mathcal{U} \subset \mathbb{C}^n$. Then $\mathcal{U}$ is called A-*invariant*, if*

$$x \in \mathcal{U} \Rightarrow Ax \in \mathcal{U} \quad \text{for all } x \in \mathbb{C}^n.$$

Theorem 6 *Let $A \in \mathbb{C}^{n,n}, X \in \mathbb{C}^{n,n}$ and $\mathcal{U} = \mathcal{R}(X)$. Then the following are equivalent:*

(a) $\mathcal{U}$ is A-invariant;

(b) There exists $B \in \mathbb{C}^{k,k}$, such that:

$$AX = XB.$$

Furthermore, in this case for $\lambda \in \mathbb{C}$ and $v \in \mathbb{C}^k$:

$$Bv = \lambda v \Rightarrow AXv = \lambda Xv,$$

i.e., every eigenvalue von B is also an eigenvalue von A.

Remark 7 *If A, X, B satisfy $AX = XB$ and if X has only one column x, then B is a scalar λ and we obtain the eigenvalue equation*

$$Ax = x\lambda,$$

i.e., X can be viewed as a generalization of the concept of eigenvector.

1.2 Matrix decompositions

1.2.1 Schur decomposition

Theorem 8 (Schur, 1909)
Let $A \in \mathbb{C}^{n,n}$. Then there exists $U \in \mathbb{C}^{n,n}$ unitary such that

$$T := U^{-1}AU$$

is upper triangular.

Proof: By induction: $n = 1$ is trivial.
"$n-1 \Rightarrow n$": Let $v \in \mathbb{C}^n$ be an eigenvector of A to the eigenvalue $\lambda \in \mathbb{C}$. Let $q_1 := \frac{v}{||v||}$ and complete q_1 to an orthonormal basis $(q_1, \ldots, q_n)$ of $\mathbb{C}^n$. Then $Q = [q_1, \ldots, q_m]$ is unitary and

$$Q^{-1}AQ = \left[\begin{array}{c|c} \lambda & A_{12} \\ \hline 0 & A_{22} \end{array}\right]$$

By the inductive assumption there exists U_{22} unitary, such that $T_{22} := U_{22}^* A_{22} U_{22}$ is upper triangular. Setting

$$U = Q \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & U_{22} \end{array}\right],$$

then $T = U^*AU$ is upper triangular. □

Remark 9 *In the Schur decomposition U can be chosen such that the eigenvalues of A appear in arbitrary order on the diagonal.*

1.2.2 The singular value decomposition (SVD)

Theorem 10 (Singular value decomposition, SVD)
Let $A \in \mathbb{C}^{m,n}$ with Rank $(A) = r$. *Then there exist unitary matrices $U \in \mathbb{C}^{m,m}$ and $V \in \mathbb{C}^{n,n}$ such that*

$$A = U\Sigma V^*, \ \Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right] \in \mathbb{C}^{m,n}.$$

Furthermore, $\sigma_1 = ||A||_2$ and $\sigma_1, \ldots, \sigma_r$ are uniquely determined.

Proof: See Golub/Van Loan, Matrix Computations. □

Definition 11 *Let $A, U = [u_1, \ldots, u_m], V = [v_1, \ldots, v_n], \Sigma$ be as in the SVD and $\sigma_k := 0$ for $k = r+1, \ldots, \min\{m,n\}$. Then*

(a) $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$ *are called* singular values *of A.*

(b) $u_1, \ldots, u_m$ *are called* left singular vectors *of A.*

(c) $v_1, \ldots, v_n$ *are called* right singular vectors *of* A.

Remark 12 *(a) From the SVD one obtains*

$$A^*A = V\Sigma^*U^*U\Sigma V^* = V\Sigma^*\Sigma V^* = V \left[\begin{array}{ccc|c} \sigma_1^2 & & & \\ & \ddots & & 0 \\ & & \sigma_r^2 & \\ \hline & 0 & & 0 \end{array}\right] V^*$$

and

$$AA^* = U\Sigma V^*V\Sigma^*U^* = U\Sigma\Sigma^*U^* = U \left[\begin{array}{ccc|c} \sigma_1^2 & & & \\ & \ddots & & 0 \\ & & \sigma_r^2 & \\ \hline & 0 & & 0 \end{array}\right] U^*,$$

i.e. $\sigma_1^2, \ldots, \sigma_r^2$ *are the nonzero eigenvalues of* AA^* *and* A^*A*, respectively.*

(b) Since $AV = U\Sigma$ *one has* Kernel $(A) = Span\{v_{r+1}, \ldots, v_n\}$ *and* Image $(A) = \mathcal{R}(A) = Span\{u_1, \ldots, u_r\}$.

(c) The SVD allows optimal low-rank approximation of A*, since*

$$\begin{aligned} A &= U\Sigma V^* \\ &= U\left(\left[\begin{array}{ccc} \sigma_1 & & \\ & 0 & \\ & & \ddots \\ & & 0 \end{array}\right] + \ldots + \left[\begin{array}{ccc} 0 & & \\ & \ddots & \\ & & 0 \\ & & \sigma_n \end{array}\right]\right) V^* \\ &= \sum_{j=1}^{r} \sigma_j u_j v_j^*. \end{aligned}$$

Here $u_j v_j^*$ *is a rank one matrix of size* $m \times n$*. For* $0 \leq \nu \leq r$ *the matrix*

$$A_\nu := \sum_{i=1}^{\nu} \sigma_i u_j v_i^*$$

is the best rank ν *approximation to* A *in the sense that*

$$||A - A_\nu|| = \inf_{\substack{B \in \mathbb{C}^{m,n} \\ \text{Rank}\,(B) \leq \nu}} ||A - B|| = \sigma_{\nu+1},$$

where $\sigma_{r+1} := 0$.

(d) If A *is real, then also* U *and* V *can be chosen real.*

1.3 Perturbation theory

In the analysis of numerical methods, we will have to study the eigenvalues, eigenvectors, and invariant subspaces under small perturbations. For example, if we compute an invariant subspace numerically, then we introduce roundoff errors and the computed subspace will only be an approximation to the invariant subspace. How good is this approximation?

1.3.1 Canonical angles and vectors

Question: let $\mathcal{U}, \mathcal{V} \subset \mathbb{C}^n$ be subspaces of dimension k. How 'near' are $\mathcal{U}$ and $\mathcal{V}$?
Strategy: Compute successively the angles between $\mathcal{U}$ and $\mathcal{V}$ beginning with the smallest. Choose normalized vectors $x \in \mathcal{U}$ and $y \in \mathcal{V}$, such that

$$|\langle x, y\rangle| \stackrel{!}{=} \max.$$

Without loss of generality we can choose x and y such that their scalar product is real and nonnegative. Otherwise we can take $z \in \mathbb{C}$ with $|z| = 1$, so that $\langle x, zy\rangle = z\langle x, y\rangle$ is real and nonnegative. Then $|\langle x, y\rangle| = |\langle x, zy\rangle|$.

(a) Choose $x_1 \in \mathcal{U}$ and $y_1 \in \mathcal{V}$ with $||x_1|| = ||y_1|| = 1$ such that

$$\langle x_1, y_1\rangle = \max\{\mathrm{Re}\langle x, y\rangle \mid x \in \mathcal{U}, y \in \mathcal{V}, ||x|| = ||y|| = 1\}$$

Then $\langle x_1, y_1\rangle$ is real, $\vartheta_1 = \arccos\langle x_1, y_2\rangle$ is called first *canonical angle* and x_1, y_1 are called first *canonical vectors.*

(b) Suppose that we have determined $j-1$ canonical angles and vectors, i.e.,

$$x_1, \ldots, x_{j-1} \in \mathcal{U},\ y_1, \ldots, y_{j-1} \in \mathcal{V}$$

are determined with $(x_1, \ldots, x_{j-1})$ and $(y_1, \ldots, y_{j-1})$ orthonormal.

Choose $x_j \in \mathcal{U}$ and $y_j \in \mathcal{V}$ with $x_j \perp x_1, \ldots, x_{j-1}$ and $y_j \perp, y_1, \ldots, y_{j-1}$ and $||x_j|| = ||y_i|| = 1$, so that

$$\langle x_j, y_j\rangle$$

has maximal real part. Then $\langle x_j, y_j\rangle$ is real,

$$\vartheta_j := \arccos\langle x_j, y_j\rangle$$

is the j-th canonical angle, and x_j, y_j are j-th canonical vectors. Proceeding inductively we obtain k canonical angles $0 \le \vartheta_1 \le \ldots \le \vartheta_k \le \frac{\pi}{2}$ and orthonormal bases $(x_1, \ldots, x_k)$ and $(y_1, \ldots, y_k)$ of $\mathcal{U}$, $\mathcal{V}$, respectively.

Lemma 13 *For $i, j = 1, \ldots, k$ and $i \neq j$ the canonical vectors satisfy $\langle x_i, y_j\rangle = 0$.*

Proof: Exercise. □

Corollary 14 *Let* $X = [x_1, \ldots, x_k]$ *and* $Y = [y_1, \ldots, y_k]$. *Then*

$$X^*Y = (\langle x_i, y_j \rangle) = \begin{bmatrix} \cos\vartheta_1 & & 0 \\ & \ddots & \\ 0 & & \cos\vartheta_k \end{bmatrix}$$

with $\cos\vartheta_1 \geq \ldots \geq \cos\vartheta_k \geq 0$ *and this is a SVD.*

Practical computation of canonical angles and vectors

(a) Determine orthonormal bases of $\mathcal{U}$ and $\mathcal{V}$, i.e., isometric matrices $P, Q \in \mathbb{C}^{n,k}$ with

$$\mathcal{R}(P) = \mathcal{U},\ \mathcal{R}(Q) = \mathcal{V}.$$

(b) Compute the SVD of P^*Q

$$P^*Q = U\Sigma V^*$$

with the diagonal matrix

$$\Sigma = \underbrace{U^*P^*}_{X^*}\underbrace{QV}_{Y}$$

(c) Set $U = [u_1, \ldots, u_k]$ and $V = [v_1, \ldots, v_k]$. Then

(a) $\vartheta_j = \arccos\sigma_j, j = 1, \ldots, k$ are the canonical angles and

(b) $Pu_j, Qv_j, j = 1, \ldots, k$ are the canonical vectors.

1.3.2 Distance between subspaces

Definition 15 *Let* $\mathcal{U}, \mathcal{V} \in \mathbb{C}^n$ *be subspaces of dimension* k.

(a) For $x \in \mathcal{U}$ *we call*

$$d(x, \mathcal{V}) := \min_{y \in \mathcal{V}} ||x - y||$$

the distance from x to $\mathcal{V}$ *and*

(b)

$$d(\mathcal{U}, \mathcal{V}) := \max_{\substack{x \in \mathcal{U} \\ ||x||=1}} d(x, \mathcal{V})$$

the distance of $\mathcal{U}$ and $\mathcal{V}$.

Theorem 16 *Let* $\mathcal{U}, \mathcal{V} \subset \mathbb{C}^n$ *be subspaces of dimension* k *with canonical angles* $\vartheta_1 \leq \ldots \leq \vartheta_k$, *then*

$$d(\mathcal{U}, \mathcal{V}) = \sin\vartheta_k.$$

Proof: See Stewart/Sun. *Matrix perturbation theory.* Boston, 1990. □

Chapter 2

Eigenvalue problems with dense matrices

Situation: $A \in \mathbb{C}^{n,n}$, where n is small enough so that the matrix A can be fully stored and that we can manipulate the whole matrix by similarity transformations.

2.1 The power method

Idea: Take an arbitrary $q \in \mathbb{C}^n \setminus \{0\}$ and form the sequence $q, Aq, A^2q, \ldots$. What will happen?

Assumption: A is diagonalizable. Let $\lambda_1, \ldots, \lambda_n$ with $|\lambda_1| \geq \ldots \geq |\lambda_n|$ be the eigenvalues of A and let $(v_1, \ldots, v_n)$ be a basis of eigenvectors. Then there exist $c_1, \ldots, c_n$ with

$$q = c_1 v_1 + \ldots + c_n v_n.$$

Further assumption: $c_1 \neq 0$ (this happens with probability 1 if q is random). Then,

$$\begin{aligned} Aq &= c_1\lambda_1 v_1 + \ldots + c_n\lambda_n v_n, \\ A^k q &= c_1\lambda_1^k v_1 + \ldots + c_n\lambda_n^k v_n. \end{aligned}$$

For $|\lambda_1| > 1$ the powers $|\lambda_1^k|$ will grow, so we scale as

$$\frac{1}{\lambda_1^k} A^k q = c_1 v_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \ldots + c_n \left(\frac{\lambda_n}{\lambda_1}\right)^k v_n.$$

Third assumption: $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$. Then,

$$\begin{aligned} \left\| \frac{1}{\lambda_1^k} A^k q - c_1 v_1 \right\| &\leq |c_2| \left|\frac{\lambda_2}{\lambda_1}\right|^k ||v_2|| + \ldots + |c_n| \left|\frac{\lambda_n}{\lambda_1}\right|^k ||v_n|| \\ &\leq \left(|c_2|||v_2|| + \ldots + |c_n|||v_n||\right) \left|\frac{\lambda_2}{\lambda_1}\right|^k \overset{k\to\infty}{\longrightarrow} 0, \end{aligned}$$

and hence $\lim_{k\to\infty} \frac{1}{\lambda_1^k} A^k q = c_1 v_1$ and the convergence is linear with convergence rate $r \leq |\frac{\lambda_2}{\lambda_1}|$.

Definition 17 *A sequence* (x_k) converges linearly *to* x*, if there exists* r *with* $0 < r < 1$ *such that*

$$\lim_{k\to\infty} \frac{||x_{k+1} - x||}{||x_k - x||} = r.$$

Then r *is called the* convergence rate *of the sequence.*
We say that the convergence $(x_k) \to x$ *is of order* $m \geq 2$ *if*

$$\lim_{k\to\infty} \frac{||x_{k+1} - x||}{||x_k - x||^m} = c \neq 0.$$

If $m = 2$ *then we speak of quadratic convergence and if* $m = 3$ *of cubic convergence.*

In practice we do not know $1/\lambda_1^k$, thus we normalize differently and divide by the largest (in modulus) component of $A^k q$.

Algorithm: (Power method)
Computes the dominant eigenvalue λ_1 and the associated eigenvector v_1.

(a) Choose $q_0 \in \mathbb{C}^n \setminus \{0\}$

(b) Iterate, for $k = 1, 2, \ldots$ to convergence

$$q_k := \frac{1}{\alpha_k} A q_{k-1},$$

where α_k is the largest (in modulus) component of Aq_{k-1}.

The power method can also be used for large scale problem where only matrix vector multiplication is available. By the presented analysis we have proved the following theorem.

Theorem 18 *Suppose that* $A \in \mathbb{C}^{n,n}$ *has the eigenvalues* $\lambda_1, \ldots, \lambda_n$ *with* $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$*. If* $q_0 \in \mathbb{C}^n \setminus \{0\}$ *has a component in the invariant subspace associated to* λ_1*, (i.e., "*$c_1 \neq 0$*"), then the sequence of subspaces span*(q_k)*, where* q_k *is defined in the power method converges to the invariant subspace associated with* λ_1*. The convergence is linear with rate* $r \leq |\frac{\lambda_2}{\lambda_1}|$.

Remark 19 (a) This theorem also holds for non-diagonalizable matrices.

(b) Forming the full products Aq_k costs $2n^2$ flops and the scaling $O(n)$ flops. Hence m iterations will cost $2n^2 m$ flops.

(c) If (as is very common) $|\frac{\lambda_2}{\lambda_1}| \approx 1$ then the convergence is very slow.

2.2 Shift-and-Invert and Rayleigh-Quotient-Iteration

Observations: Let $A \in \mathbb{C}^{n,n}$ and $(\lambda, v) \in \mathbb{C} \times \mathbb{C}^n$ with $Av = \lambda v$. Then

(a) $A^{-1}v = \lambda^{-1}v$ for A invertible, and

(b) $(A - \varrho I)v = (\lambda - \varrho)v$ for all $\varrho \in \mathbb{C}$.

If $\lambda_1, \ldots, \lambda_n$ are again the eigenvalues of A with $|\lambda_1| \geq \ldots \geq |\lambda_n|$, then we can perform the following iterations.

Inverse Iteration. This is the power method applied to A^{-1}. If $|\lambda_n| < |\lambda_{n-1}|$, then the inverse iteration converges to an eigenvector to λ_n with convergence rate $|\frac{\lambda_n}{\lambda_{n-1}}|$ (which is small if $|\lambda_n| \ll |\lambda_{n-1}|$).

Shift and Invert Power Method. This is the power method applied to $(A - \varrho I)^{-1}$. Let λ_j, λ_k be the eigenvalues that are closest to ϱ, and suppose that $|\lambda_j - \varrho| < |\lambda_k - \varrho|$. Then the power method for $(A - \varrho I)^{-1}$ converges to an eigenvector associated with λ_j with rate

$$\left| \frac{\lambda_j - \varrho}{\lambda_k - \varrho} \right|.$$

This is small if $|\lambda_j - \varrho| \ll |\lambda_k - \varrho|$ and $\lambda_j \approx \varrho$ would be optimal.

Where do we get good shifts ϱ for the Shift and Invert Power Method? To answer this question we need some results on residuals and backward errors.

Definition 20 *Let $A \in \mathbb{C}^{n,n}$ and $(\mu, w) \in \mathbb{C} \times \mathbb{C}^n$. Then $Aw - \mu w$ is called the* residual *of (μ, w) with respect to A.*

Theorem 21 *Let $\mu \in \mathbb{C}, \varepsilon > 0$, $A \in \mathbb{C}^{n \times n}$, and $w \in \mathbb{C}^n$ with $||w|| = 1$. If $||Aw - \mu w|| = \varepsilon$, then there exists a matrix (*the backward error matrix*) $E \in \mathbb{C}^{n,n}$ with $||E|| \leq \varepsilon$ such that*

$$(A + E)w = \mu w.$$

Proof: Let $r := Aw - \mu w$ and $E = -rw^*$. Then

$$(A + E)w = Aw - r\underbrace{w^*w}_{=1} = \mu w$$

$$\text{and} \quad ||E|| = ||rw^*|| \leq ||r||||w^*|| = ||r|| = \varepsilon.$$

□

The idea to determine a good eigenvalue approximation (shift) from a given eigenvector approximation is to minimize the residual $||Aw - \mu w||$. Consider the over-determined linear system

$$w\mu = Aw$$

with the $n \times 1$-Matrix w, the unknown vector μ and the right hand side Aw. We can use the normal equations to solve $||Aw - \mu w|| = \min!$, i.e., we use

$$w^*w\mu = w^*Aw \qquad \text{respect.} \qquad \mu = \frac{w^*Aw}{w^*w}.$$

Definition 22 *Let $A \in \mathbb{C}^{n,n}$ and $w \in \mathbb{C}^n \setminus \{0\}$. Then*

$$r(w) := \frac{w^*Aw}{w^*w}$$

is called the Rayleigh-quotient *of w with respect to A.*

The following theorem gives an estimate for the distance of the Rayleigh-quotient from an eigenvalue.

Theorem 23 *Let $A \in \mathbb{C}^{n,n}$ and $(\lambda, v) \in \mathbb{C} \times \mathbb{C}^n$ be an eigenvalue/eigenvector pair of A with $\|v\| = 1$. Then for $w \in \mathbb{C}^n$ with $\|w\| = 1$ the following estimate holds:*

$$|\lambda - r(w)| \leq 2\|A\| \cdot \|v - w\|.$$

This gives an the idea for an iteration to iterate computing an approximate eigenvector and from this a Rayleigh-quotient, i.e., the following algorithm:

Algorithm: Rayleigh-Quotient-Iteration (RQI)
This algorithm computes an eigenvalue/eigenvector pair $(\lambda, v) \in \mathbb{C} \times \mathbb{C}^n$ of the matrix $A \in \mathbb{C}^{n,n}$.

(a) Start: Choose $q_0 \in \mathbb{C}^n$ with $||q_0|| = 1$ and set $\lambda_0 := q_0^* A q_0$.

(b) Iterate for $k = 1, 2, \ldots$ until convergence

(a) Solve the linear system $(A - \lambda_{k-1} I)x = q_{k-1}$ for x.

(b) $q_k := \dfrac{x}{||x||}$

(c) $\lambda_k := q_k^* A q_k$

Remark 24 (a) It is difficult to analyze the convergence of this algorithm but one observes practically that it almost always converges. The convergence rate is typically quadratic. For Hermitian matrices $A = A^*$ there is more analysis and one can even show cubic convergence.

(b) **Costs:** $O(n^3)$ flops per step if the linear system is solved with full Gaussian elimination. The costs are $O(n^2)$ for Hessenberg matrices (see Chapter 2.4.2) and they can be even smaller for banded or other sparse matrices.

2.3 Simultaneous subspace iteration

To compute several eigenvalues and the associated invariant subspace, we can generalize the power method to the subspace iteration. Consider $A \in \mathbb{C}^{n,n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, where $|\lambda_1| \geq \ldots \geq |\lambda_n|$.

Idea: Instead of $q_0 \in \mathbb{C}^n$, consider a set of linearly independent vectors $\{w_1, \ldots, w_m\} \subset \mathbb{C}^n$. Set

$$W_0 := [w_1, \ldots, w_m] \in \mathbb{C}^{n \times m},$$

and form the sequence $W_0, AW_0, A^2 W_0, \ldots$ via

$$W_k := A^k W_0 = \left[A^k w_1, \ldots, A^k w_m\right], \quad k \geq 1.$$

In general, we expect $\mathcal{R}(W_k)$ to converge to the invariant subspace $\mathcal{U}$ associated with the m eigenvalues $\lambda_1, \ldots, \lambda_m$. This iteration is called *simultaneous subspace iteration.*

Theorem 25 *Let $A \in \mathbb{C}^{n,n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ satisfy*

$$|\lambda_1| \geq \ldots \geq |\lambda_m| > |\lambda_{m+1}| \geq \ldots \geq |\lambda_n|.$$

Let $\mathcal{U}, \mathcal{V}$ be the invariant subspaces associated with $\lambda_1, \ldots, \lambda_m$, and $\lambda_{m+1}, \ldots, \lambda_n$ respectively. Furthermore, let $W \in \mathbb{C}^{n \times m}$ with $\operatorname{Rank}(W) = m$ and $\mathcal{R}(W) \cap \mathcal{V} = \{0\}$. Then for the iteration $W_0 := W$, $W_{k+1} = AW_k$ for $k \geq 0$ and for every ϱ with $\left|\frac{\lambda_{m+1}}{\lambda_m}\right| < \varrho < 1$, there exists a constant c such that

$$d(\mathcal{R}(W_k), \mathcal{U}) \leq c \cdot \varrho^k, \quad k \geq 1.$$

For the proof of the theorem, we need the following lemma:

Lemma 26 *Let*

$$\mathcal{U} = \mathcal{R}\left(\begin{bmatrix} I_m \\ 0 \end{bmatrix}\right) \quad \text{and} \quad \hat{\mathcal{U}} = \mathcal{R}\left(\begin{bmatrix} I_m \\ X \end{bmatrix}\right),$$

with $X \in \mathbb{C}^{(n-m)\times m}$, $m < n$, be m-dimensional subspaces of $\mathbb{C}^n$ and let $\theta_1, \ldots, \theta_m$ be the canonical angles between $\mathcal{U}$ and $\hat{\mathcal{U}}$. Then $||X|| = \tan\theta_m$.

Proof: See Stewart/Sun. *Matrix perturbation theory.* Boston, 1990. □

Proof: (of the theorem) We prove the theorem for the case that A is diagonalizable. We perform a similarity transformation

$$A_{\text{new}} = S^{-1}A_{\text{old}}S = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix},$$

with $A_1 = \text{diag}\,(\lambda_1, \ldots, \lambda_m)$ and $A_2 = \text{diag}\,(\lambda_{m+1}, \ldots, \lambda_n)$. Then A_1 is nonsingular, since $|\lambda_1| \geq \ldots \geq |\lambda_m| > 0$. Set

$$\mathcal{U}_{\text{new}} = S^{-1}\mathcal{U}_{\text{old}} = \mathcal{R}\left(\begin{bmatrix} I_m \\ 0 \end{bmatrix}\right)$$

and

$$\mathcal{V}_{\text{new}} = S^{-1}\mathcal{V}_{\text{old}} = \mathcal{R}\left(\begin{bmatrix} 0 \\ I_{n-m} \end{bmatrix}\right).$$

Furthermore, let

$$W_{\text{new}} = S^{-1}W_{\text{old}} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$$

for some $Z_1 \in \mathbb{C}^{m,m}$ and $Z_2 \in \mathbb{C}^{n-m,m}$. Then

(a) $d(\mathcal{R}(W_{\text{new}}),\mathcal{U}_{\text{new}}) \leq \kappa(S)d(\mathcal{R}(W_{\text{old}}),\mathcal{U}_{\text{old}})$ (Exercise)
(Here $\kappa(S) = \|S\|\|S^{-1}\|$ is the condition number of S with respect to inversion.)

(b) $\mathcal{R}(W_{\text{new}}) \cap \mathcal{V}_{\text{new}} = \{0\} \Leftrightarrow Z_1$ is nonsingular (Exercise)

In the following we drop the index 'new'. Then

$$W = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} I_m \\ Z_2Z_1^{-1} \end{bmatrix} Z_1 = \begin{bmatrix} I \\ X_0 \end{bmatrix} Z_1$$

with $X_0 = Z_2Z_1^{-1}$, and hence

$$\mathcal{R}(W) = \mathcal{R}\left(\begin{bmatrix} I_m \\ X_0 \end{bmatrix}\right)$$

as well as

$$\mathcal{R}(W_k) = \mathcal{R}(A^kW) = \mathcal{R}(A^k \begin{bmatrix} I \\ X_0 \end{bmatrix}).$$

Then it follows that

$$A^k \begin{bmatrix} I \\ X_0 \end{bmatrix} = \begin{bmatrix} A_1^k & 0 \\ 0 & A_2^k \end{bmatrix} \begin{bmatrix} I \\ X_0 \end{bmatrix} = \begin{bmatrix} A_1^k \\ A_2^k X_0 \end{bmatrix} = \begin{bmatrix} I_m \\ \underbrace{A_2^k X_0 A_1^{-k}}_{=:X_k} \end{bmatrix} A_1^k,$$

and thus,

$$\mathcal{R}(W_k) = \mathcal{R}\left(\begin{bmatrix} I_m \\ X_k \end{bmatrix} \right).$$

It remains to show that

$$d(\mathcal{R}(W_k), \mathcal{U}) \to 0.$$

Let $\Theta_m^{(k)}$ be the largest canonical angle between $\mathcal{R}(W_k)$ and $\mathcal{U}$. Then

$$\begin{aligned} d(\mathcal{R}(W_k), \mathcal{U}) &= \sin \Theta_m^{(k)} \leq \tan \Theta_m^{(k)} = ||X_k|| \leq ||A_2^k|| ||X_0|| ||A_1^{-k}|| \\ &= |\lambda_{m+1}^k| ||X_0|| |\lambda_m^{-k}|, \end{aligned}$$

which implies that

$$d(R(W_k), \mathcal{U}) \leq \tilde{c} \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^k.$$

Undoing the similarity transformation we obtain the desired result. For the diagonalizable case we do not need the bound ϱ. This will be only needed in the non-diagonalizable case. □

Remark 27 For $W_0 = [w_1, \ldots, w_m]$ we have

$$A^k W_0 = \left[A^k w_1, \ldots, A^k w_m \right],$$

i.e., we perform the iteration not only for W_0 but simultaneously also for all $W_0^{(j)} = [w_1, \ldots, w_j]$, since

$$A^k W_0^{(j)} = \left[A^k w_1, \ldots, A^k w_j \right].$$

Under appropriate assumptions, we then have convergence of

$$\text{Span} \left\{ A^k w_1, \ldots, A^k w_j \right\}$$

to the invariant subspace associated with $\lambda_1, \ldots, \lambda_j$ for all $j = 1, \ldots, m$. For this reason one speaks of 'simultaneous subspace iteration'.

Problems with subspace iteration in finite precision arithmetic:

Theory: $A^k W_0 = \left[A^k w_1, \ldots, A^k w_m \right]$ in general has Rank m (for generic starting values).

Practice: Unfortunately, in finite precision arithmetic rounding errors lead to linear dependence in $\mathcal{R}(W_k)$ already after few iterations.

The basic idea to cope with this problem is to orthonormalize the columns in every step.

Step 1: Factor $W_0 = [w_1, \ldots, w_m] = Q_0 R_0$ with $Q_0 \in \mathbb{C}^{n,m}$ isometric and $R_0 \in \mathbb{C}^{m,m}$ upper triangular. Then $\mathcal{R}(W_0) = \mathcal{R}(Q_0)$ and furthermore,

$$\text{Span}\{w_1, \ldots, w_j\} = \text{Span}\left\{q_1^{(0)}, \ldots, q_j^{(0)}\right\}, \; j = 1, \ldots, m,$$

where $Q_0 = \left[q_1^{(0)}, \ldots, q_m^{(0)}\right]$. (This follows form the triangular form of R_0.)

Situation after step $k-1$: $\mathcal{R}(W_{k-1}) = \mathcal{R}(Q_{k-1})$ with

$$Q_{k-1} = \left[q_1^{(k-1)}, \ldots, q_m^{(k-1)}\right] \in \mathbb{C}^{n,m}$$

isometric and

$$\text{Span}\left\{A^{k-1}w_1, \ldots, A^{k-1}w_j\right\} = \text{Span}\left\{q_1^{(k-1)}, \ldots, q_j^{(k-1)}\right\}, \quad j = 1, \ldots, m.$$

Step k: Let $AQ_{k-1} = Q_k R_k$ be a QR decomposition with $Q_k \in \mathbb{C}^{n,m}$ isometric and $R_k \in \mathbb{C}^{m,m}$ upper triangular. Then

$$\mathcal{R}(W_k) = \mathcal{R}(AW_{k-1}) = \mathcal{R}(AQ_{k-1}) = \mathcal{R}(Q_k),$$

and moreover

$$\text{Span}\left\{A^k w_1, \ldots, A^k w_j\right\} = \text{Span}\left\{q_1^{(k)}, \ldots, q_j^{(k)}\right\}, \quad j = 1, \ldots, m.$$

Algorithm: Unitary Subspace Iteration

(a) Start: Choose $Q_0 \in \mathbb{C}^{n,m}$ isometric.

(b) Iterate. For $k = 1, 2, \ldots$ to convergence:

  (a) Compute $Z_k = AQ_{k-1}$

  (b) Compute QR-decomposition $Z_k = Q_k R_k$.

Remark 28 Theoretically the convergence behavior of the unitary subspace iteration is as for the subspace iteration but, fortunately, the described problems in finite precision arithmetic do not arise.

2.4 The Francis QR Algorithm

2.4.1 Simple QR Algorithm Without Shifts

Let $A \in \mathbb{C}^{n,n}$ have the eigenvalues $\lambda_1, \ldots, \lambda_n$ where $|\lambda_1| \geq \ldots \geq |\lambda_n|$.

Idea: Use the n-dimensional unitary subspace iteration, i.e., choose $m = n$ and $Q_0 = I_n$. If $Q_k = \left[q_1^{(k)}, \ldots, q_n^{(k)}\right]$, then for every $1 \leq m \leq n$

$$\text{Span}\left\{q_1^{(k)}, \ldots, q_m^{(k)}\right\}$$

converges to the invariant subspace associated with $\lambda_1, \ldots, \lambda_m$ with a rate $\left|\frac{\lambda_{m+1}}{\lambda_m}\right|$ provided that $|\lambda_{m+1}| < |\lambda_m|$ and one does not run into an exceptional situation.
To observe the convergence, we form $A_k = Q_k^{-1} A Q_k$. If $\text{Span}\{q_1^{(k)}, \ldots, q_m^{(k)}\}$ converges to an invariant subspace, then we expect that in the matrix

$$A_k = \begin{matrix} \\ m \\ n-m \end{matrix} \begin{matrix} \begin{matrix} m & n-m \end{matrix} \\ \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \end{matrix}$$

the block A_{21} converges to 0 for $k \to \infty$. Since this happens for all m simultaneously, it follows that A_k converges to block-upper triangular matrix.
Another question is whether we can directly move from A_{k-1} to A_k? To see this, observe that

$$\begin{aligned} A_{k-1} &= Q_{k-1}^{-1} A Q_{k-1} \\ A_k &= Q_k^{-1} A Q_k \end{aligned}$$

and hence

$$A_k = Q_k^{-1} Q_{k-1} A_{k-1} \underbrace{Q_{k-1}^{-1} Q_k}_{=:U_k} = U_k^{-1} A_{k-1} U_k.$$

Thus we can reformulate the k-th step of the unitary subspace iteration

$$A Q_{k-1} = Q_k R_k$$

as

$$A_{k-1} = Q_{k-1}^{-1} A Q_{k-1} = Q_{k-1}^{-1} Q_k R_k = U_k R_k.$$

This is a QR decomposition of A_{k-1} and we have

$$A_k = U_k^{-1} A_{k-1} U_k = U_k^{-1} U_k R_k U_k = R_k U_k.$$

Algorithm: (QR Algorithm)(Francis and Kublanovskaya 1961)
For a given matrix $A \in \mathbb{C}^{n,n}$ this algorithm constructs a sequence (A_k) of similar matrices that converges to block upper-triangular form.

(a) Start with $A_0 = A$

(b) Iterate for $k = 1, 2, \ldots$ until convergence.

(a) Compute a QR-decomposition of A_{k-1}: $A_{k-1} = U_k R_k$

(b) Compute A_k via $A_k = R_k U_k$.

Theorem 29 (Convergence of the QR algorithm)
Let $A \in \mathbb{C}^{n,n}$ have eigenvalues $\lambda_1, \ldots, \lambda_n$, where $|\lambda_1| \geq \ldots \geq |\lambda_m| > |\lambda_{m+1}| \geq \ldots \geq |\lambda_n|$. Let $\mathcal{V} \subset \mathbb{C}^n$ be the invariant subspace associated with $\lambda_{m+1}, \ldots, \lambda_n$, and let (A_k) be the sequence generated by the QR Algorithm. If

$$Span\{e_1, \ldots, e_m\} \cap \mathcal{V} = \{0\}$$

and

$$A_k = \begin{array}{c} \\ m \\ n-m \end{array} \begin{array}{c} \begin{array}{cc} m & n-m \end{array} \\ \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & A_{22}^{(k)} \end{bmatrix} \end{array},$$

then for every ϱ with $\left|\frac{\lambda_{m+1}}{\lambda_m}\right| < \varrho < 1$ there exists a constant $\tilde{c}$ such that

$$||A_{21}^{(k)}|| \leq \tilde{c}\varrho^k.$$

Proof: (Sketch) Let $\mathcal{U}$ be the invariant subspace associated with $\lambda_1, \ldots, \lambda_m$ and

$$\mathcal{U}_k = \text{Span}\left\{q_1^{(k)}, \ldots, q_m^{(k)}\right\},$$

where

$$Q_k = \left[q_1^{(k)}, \ldots, q_n^{(k)}\right]$$

is the unitary matrix with $Q_k^{-1} A Q_k = A_k$ from the unitary subspace iteration. One first shows that

$$||A_{21}^{(k)}|| \leq 2\sqrt{2}||A||d(\mathcal{U}, \mathcal{U}_k)$$

Then using the convergence results for the subspace iteration there exists a constant $c > 0$ with

$$d(\mathcal{U}, \mathcal{U}_k) \leq c\varrho^k.$$

Then choose $\tilde{c} := 2\sqrt{2}||A||c$. □

In the special case that A is Hermitian, the sequence A_k converges to a diagonal matrix.

Remark 30 In the presented form the algorithm has two major disadvantages:

(a) It is expensive, since it costs $O(n^3)$ flops per iteration step.

(b) The convergence is slow (only linear).

A way to address the two problems is the Hessenberg reduction and the use of shifts.

2.4.2 Hessenberg reduction

Definition 31 *A matrix $A = [a_{ij}]$ is called* Hessenberg matrix *or* in Hessenberg form, *if $a_{ij} = 0$ for $i > j + 1$. A Hessenberg matrix A is called* unreduced *if $a_{i+1,i} \neq 0$ for all $i = 1, \ldots, n-1$.*

Householder transformations
The QR decomposition and other unitary transformations can be realized via *Householder transformations*

$$P = I - \frac{2}{v^* v} v v^*$$

for $v \in \mathbb{C}^n \setminus \{0\}$. Householder Transformations are Hermitian and unitary. (Exercise) Multiplication with Householder transformations is geometrically a reflection of a vector $x \in \mathbb{C}^n$ at the hyperplane $\mathrm{Span}\{v\}^\perp$. (Exercise.)
A typical task: Reflect $x \in \mathbb{C}^n \setminus \{0\}$ to a multiple of the first unit vector, i.e., determine v and from this P such that

$$Px = \pm ||x|| e_1.$$

To determine such a v we make an ansatz $v = x + \alpha e_1$ and obtain

$$v = x \pm ||x|| e_1 \qquad \text{and} \qquad Px = \mp ||x|| e_1. \qquad \text{(Exercise)}$$

For numerical reasons we take

$$v = \begin{cases} x + ||x|| e_1, & x_1 \geq 0, \\ x - ||x|| e_1, & x_1 < 0. \end{cases}$$

Advantage: The multiplication with Householder transformations is cheap. For $B \in \mathbb{C}^{n,m}$ the computation of PB only needs $\sim 4mn$ flops (instead of $\sim 2n^2 m$ flops for classical matrix/matrix multiplication).
Givens rotations
Another tool for unitary operations are *Givens rotations*

$$G_{i,j}(c,s) = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & & & s & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & -\bar{s} & & & \bar{c} & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix},$$

where $|c|^2 + |s|^2 = 1$ and where the matrix differs from an identity only in positions $(i,i), (i,j), (j,i), (j,j)$. Multiplication of a matrix with a Givens rotation allows to zero an element in any position.
E.g. choose

$$\hat{G}_{1,2}(c,s) = \begin{bmatrix} c & s \\ -\bar{s} & \bar{c} \end{bmatrix}$$

with $|c|^2 + |s|^2 = 1$ such $-\bar{s}a_{11} + \bar{c}a_{21} = 0$, then we have

$$\hat{G}_{1,2}(c,s) \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} * & * \\ 0 & * \end{bmatrix}.$$

(a) For a Hessenberg matrix $H \in \mathbb{C}^{n,n}$ the QR decomposition can be performed in $O(n^2)$ flops using Givens rotations.

(b) Hessenberg matrices are invariant under QR iterations.

2.4.3 The Francis QR Algorithm with Shifts

Deflation: Let $H \in \mathbb{C}^{n,n}$ be in Hessenberg form. If H is not unreduced, i.e., if $h_{m+1,m} = 0$ for some m, then

$$H = \begin{matrix} \\ m \\ n-m \end{matrix} \begin{matrix} \begin{matrix} m & n-m \end{matrix} \\ \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix} \end{matrix}$$

i.e., we can split our problem into two subproblems H_{11}, H_{22}.

Algorithm (QR Algorithm with Hessenberg reduction and shifts)
Given: $A \in \mathbb{C}^{n,n}$:

(a) Compute U_0 unitary such that

$$H_0 := U_0^* A U_0$$

is in Hessenberg form. We may assume that H_0 is unreduced, otherwise we can deflate right away.

(b) Iterate for $k = 1, 2, \ldots$ until deflation happens, i.e.,

$$h_{m+1,m}^{(k)} = O(\texttt{eps})(|\texttt{h}_{\texttt{m,m}}| + |\texttt{h}_{\texttt{m+1,m+1}}|)$$

for some m and the machine precision eps.

(i) Choose shift $\mu_k \in \mathbb{C}$.

(ii) Compute a QR decomposition $H_{k-1} - \mu_k I = Q_k R_k$ of $H_{k-1} - \mu_k I$.

(iii) Form $H_k = R_k Q_k + \mu_k I$.

Remark 32 *(a) Steps (ii) and (iii) of this algorithm correspond to a QR iteration step for $H_0 - \mu_k I$.*

(b) The sub-diagonal entry $h_{m+1,m}^{(k)}$ in H_k converges with rate $\left|\frac{\lambda_{m+1} - \mu_k}{\lambda_m - \mu_k}\right|$ to 0.

(c) If $h_{m+1,m}^{(k)} = 0$ or $h_{m+1,m}^{(k)} = O(eps)$, then we have deflation and we can continue with smaller problems.

(d) If μ_k is an eigenvalue then deflation happens immediately after one step.

Shift strategies:

(a) **Rayleigh-quotient shift:** For the special case that A is Hermitian, the sequence A_k converges to a diagonal matrix. Then $q_n^{(k)}$ is a good approximation to an eigenvector and a good approximation to the eigenvalue is the Rayleigh-quotient

$$r(q_n^{(k)}) = (q_n^{(k)})^* A q_n^{(k)}$$

which is just the n-th diagonal entry $a_{n,n}^{(k)}$ of $Q_k^* A Q_k$.

Heuristic: We expect in general $a_{n,n}^{(k)}$ to be a good approximation to an eigenvalue and therefore may choose

$$\mu_k = a_{n,n}^{(k)}$$

With this choice $h_{n,n-1}^{(k)}$ typically converges quadratically to 0.

(b) **Wilkinson-shift:** Problems with the Rayleigh-quotient shift arise when the matrix is real and has nonreal eigenvalues, e.g., for

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

A QR iteration for $A_0 = A$ yields $Q_0 = A_0, R_0 = I$ and hence,

$$R_0 Q_0 = I A_0 = A_0,$$

i.e., the algorithm stagnates. To avoid such situations, for $A \in \mathbb{C}^{n,n}$ in the k-th step, one considers the submatrix B in

$$A_k = \begin{array}{c} \\ n-2 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} n-2 & 2 \end{array} \\ \begin{bmatrix} * & * \\ * & B \end{bmatrix} \end{array}$$

and chooses the μ_k as shift that is nearest to $a_{nn}^{(k)}$.

(c) Another strategy, the double-shift will be discussed below.

(d) On the average 2-3 iterations are needed until a 1×1 or 2×2 block deflates.

Remark 33 (a) The empirical costs for the computation of all eigenvalues of A are approx. $10n^3$ flops, If also the transformation matrix Q is needed then this leads to approximately $25n^3$ flops.

(b) The convergence analysis is difficult, no global convergence proof is known.

Theorem 34 (Implicit Q Theorem) *Let $A \in \mathbb{C}^{n,n}$ and let $Q = [q_1, \ldots, q_n]$, $U = [u_1, \ldots, u_n]$ be unitary matrices such that*

$$H = Q^{-1} A Q = (h_{ij}) \qquad \text{and} \qquad G = U^{-1} A U = [g_{ij}]$$

are Hessenberg matrices. If $q_1 = u_1$ and H is unreduced, then

$$q_i = c_i u_i$$

for $c_i \in \mathbb{C}$ with $|c_i| = 1$ and $|h_{i,i-1}| = |g_{i,i-1}|$ for $i = 2, \ldots, n$, i.e., Q is determined already essentially uniquely by q_1.

Proof: Exercise. □

2.4.4 Implicit Shifts and 'Bulge-Chasing'

Let $H \in \mathbb{C}^{n,n}$ be a Hessenberg matrix and $\mu_1, \ldots, \mu_l \in \mathbb{C}$. Carry out l steps of the QR Algorithm with shifts $\mu_1, \ldots, \mu_l$.

$$\begin{aligned} H - \mu_1 I &= Q_1 R_1 \\ H_1 &= R_1 Q_1 + \mu_1 I \\ &\vdots \\ H_{l-1} - \mu_l I &= Q_l R_l \\ H_l &= R_l Q_l + \mu_l I. \end{aligned}$$

Then

$$H_l = Q_l^H Q_l R_l Q_l + \mu_l Q_l^H Q_l = Q_l^H (Q_l R_l + \mu_l I) Q_l = Q_l^H H_{l-1} Q_l$$

and thus per induction

$$H_l = Q_l^H \ldots Q_1^H H \underbrace{Q_1 \ldots Q_l}_{=:Q} = Q^H H Q.$$

This opens the question whether we can compute Q directly without carrying out l QR-iterations.

Lemma 35 $M := (H - \mu_l I) \ldots (H - \mu_1 I) = Q_1 \ldots Q_l \underbrace{R_l \ldots R_1}_{=:R} = QR.$

Proof: By induction we show that

$$(H - \mu_j I) \ldots (H - \mu_1 I) = Q_1 \ldots Q_j R_j \ldots R_1, \quad j = 1, \ldots, l.$$

$j = 1$**:** This is just the first step of the QR algorithm.

$j - 1 \to j$**:**

$$\begin{aligned} & Q_1 \ldots Q_j R_j \ldots R_1 \\ = \; & Q_1 \ldots Q_{j-1} (H_{j-1} - \mu_j I) R_{j-1} \ldots R_1 \\ = \; & Q_1 \ldots Q_{j-1} \Big(Q_{j-1}^H \ldots Q_1^H H Q_1 \ldots Q_{j-1} - \mu_j I \Big) R_{j-1} \ldots R_1 \\ = \; & (H - \mu_j I) Q_1 \ldots Q_{j-1} R_{j-1} \ldots R_1 \\ \overset{I.A.}{=} \; & (H - \mu_j I)(H - \mu_{j-1} I) \ldots (H - \mu_1 I). \end{aligned}$$

□

This leads to the idea to compute M and then the Householder QR decomposition of M, i.e., $M = QR$, and to set

$$\tilde{H} = Q^H R Q = H_l.$$

This means that one just needs one QR decomposition instead of l QR decompositions in each QR step. On the other hand we would have to compute M, i.e., $l - 1$ matrix-matrix multiplications. But this can be avoided by computing $\tilde{H}$ directly from H using the implicit Q Theorem.

Implicit shift-strategy:

(a) Compute

$$Me_1 = (H - \mu_l I) \dots (H - \mu_1 I) e_1,$$

the first column of M. Then the first $l+1$ entries are in general nonzero. If l is not too large, then this costs only $O(1)$ flops.

(b) Determine a Householder matrix P_0 such that $P_0(Me_1)$ is a multiple of e_1. Transform H with P_0 as

$$\begin{aligned} & & & l+1 & n-l-1 \\ P_0 &= & l+1 & \begin{bmatrix} * & 0 \end{bmatrix} \\ & & n-l-1 & \begin{bmatrix} 0 & I \end{bmatrix} \\ & & & l+2 & n-l-2 \\ P_0 H P_0 &= & l+2 & \begin{bmatrix} * & * \end{bmatrix} \\ & & n-l-2 & \begin{bmatrix} 0 & \hat{H} \end{bmatrix} \end{aligned}$$

P_0 changes only rows and columns $1, \dots, l+1$ of H. This gives a Hessenberg matrix with a bulge.

(c) Determine Householder matrices $P_1, \dots, P_{n-2}$ to restore the Hessenberg form. This is called *bulge chasing*, since we chase the bulge the down the diagonal. This yields

$$\tilde{H} := P_{n-2} \dots P_1 P_0 H P_0 \dots P_{n-2}$$

that is again in Hessenberg form and $P_k e_1 = e_1$ for $k = 1, \dots, n-2$.

(d) P_0 has the same first column as Q. As in the first step for P_0 we have $P_k e_1 = e_1$, then also

$$P_0 P_1 \dots P_{n-2}$$

has the same first column as P_0 and Q, respectively. With the implicit Q Theorem then also Q and $P_0 \dots, P_{n-2}$ and therefore also $\tilde{H}$ and $Q^H H Q$ are essentially equal, thus we have computed $\tilde{H}$ and H_l directly from H.

Algorithm (Francis QR algorithm with implicit double-shift strategy): (Francis 1961)
Given $A \in \mathbb{C}^{n,n}$:

(a) Determine U_0 unitary so that $H_0 := U_0^H A U_0$ is in Hessenberg form.

(b) Iterate for $k = 1, 2, \dots$ to convergence (deflation):

(a) Compute the eigenvalues μ_1, μ_2 of the lower right 2×2 submatrix of H_{k-1}.

(b) Compute (with the implicit shift strategy) for $l = 2$ the matrix $\tilde{Q}_k$ that one obtains with 2 steps of the QR Algorithm with shifts μ_1, μ_2.

(c)

$$H_k := \tilde{Q}_k^H H_{k-1} \tilde{Q}_k$$